

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/360192059>

# A Survey on Feature Selection and Classification Techniques for EEG Signal Processing

Chapter · January 2022

DOI: 10.1007/978-981-16-5652-1\_13

---

CITATIONS

3

READS

1,983

3 authors, including:



Saranya .K

Sri Ramakrishna Institute of Technology

26 PUBLICATIONS 112 CITATIONS

SEE PROFILE

Gunasekaran Manogaran  
A. Shanthini  
G. Vadivu *Editors*

# Proceedings of International Conference on Deep Learning, Computing and Intelligence

ICDCI 2021

# **Advances in Intelligent Systems and Computing**

**Volume 1396**

## **Series Editor**

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

## **Advisory Editors**

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Indexed by DBLP, INSPEC, WTI Frankfurt eG, zbMATH, Japanese Science and Technology Agency (JST).

All books published in the series are submitted for consideration in Web of Science.

For proposals from Asia please contact Aninda Bose ([aninda.bose@springer.com](mailto:aninda.bose@springer.com)).

More information about this series at <https://link.springer.com/bookseries/11156>

Gunasekaran Manogaran · A. Shanthini · G. Vadivu  
Editors

# Proceedings of International Conference on Deep Learning, Computing and Intelligence

ICDCI 2021



Springer

*Editors*

Gunasekaran Manogaran  
University of California  
Davis, CA, USA

A. Shanthini  
Department of Information Technology  
SRM Institute of Science and Technology  
Kattankulathur, India

G. Vadivu  
Department of Information Technology  
SRM Institute of Science and Technology  
Kattankulathur, India

ISSN 2194-5357

ISSN 2194-5365 (electronic)

Advances in Intelligent Systems and Computing

ISBN 978-981-16-5651-4

ISBN 978-981-16-5652-1 (eBook)

<https://doi.org/10.1007/978-981-16-5652-1>

© 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

# **Organization**

## **Conference Chairs**

Dr. A. Shanthini, Associate Professor, SRMIST, India  
Dr. Gunasekaran Manogaran, UC Davis, USA  
Dr. G. Vadivu, Professor, SRMIST, India

## **Keynote Speakers**

Dr. Ching-Hsien-Hsu, Asia University, Taiwan  
Dr. Gunasekaran Manoharan, Beirut Arab University, Beirut, Lebanon  
Dr. Oscar Sanjuan Martinez, Universidad Internacional de la Rioja, Logroño, Spain  
Dr. Ashish Kr. Luhach, PNG University of Technology, Papua, New Guinea

## **Chief Patrons**

Dr. T. R. Paarivendhar, Founder Chancellor, SRMIST  
Mr. Ravi Pachamoothoo, Pro Chancellor (Administration), SRMIST  
Dr. P. Sathyaranayanan, Pro Chancellor (Academics), SRMIST  
Dr. R. Shiva Kumar, Vice President, SRMIST

## **Advisory Committee**

Dr. Sandeep Sancheti, Vice Chancellor, SRMIST  
Dr. T. P. Ganesan, Pro-VC (P&D), SRMIST

- Dr. N. Sethuraman, Registrar, SRMIST  
Dr. C. Muthamizhchelvan, Pro-VC (E&T), SRMIST  
Dr. T. V. Gopal, Dean (CET), SRMIST  
Dr. Revathi Venkataraman, Chairperson (School of Computing)  
Dr. K. Ramasamy, Director (Research), SRMIST  
Dr. S. R. S. Prabhakaran, Joint Director (Research), SRMIST  
Dr. B. Neppolian, Dean Research, SRMIST

## Conference Coordinators

- Mr. Ravi Pachamoothoo, Assistant Professor  
Dr. P. Sathyaranarayanan, Assistant Professor

## International Technical Advisory Committee

- Dr. Ali Kashif Bashir, Manchester Metropolitan University, UK  
Dr. Nirmala Shenoy, Rochester Institute of Technology, USA  
Dr. Ultu Kose, SDU, Turkey  
Dr. Xavier Fernando, RU, Canada  
Dr. Dominic P. D. D., UTP, Malaysia  
Dr. Sheeba Backia Mary, Huawei Tech, Sweden  
Dr. A. S. M. Kayes, Latrobe University, Australia  
Dr. Priyan Malarvizhi Kumar, Kyung Hee University, South Korea  
Dr. Sujatha Krishnamoorthy, Wenzhou Kean University, China  
Dr. Oscar Sanjuan Martinez, Universidad Internacional de la Rioja, Logroño, Spain  
Dr. Carlos Enrique Montenegro Marin, Universidad Distrital Francisco José de Caldas, Colombia  
Dr. Vijayalakshmi Saravanan, Ryerson University, Canada  
Dr. Fadi Al-Turjman, Antalya Bilim University, Turkey  
Dr. Vicente García Díaz, University of Oviedo, Spain  
Dr. Syed Hassan Ahmed, JMA Wireless, USA  
Dr. Mamoun Alazab, Charles Darwin University, Australia  
Dr. Ashish Kr. Luhach, The PNG University of Technology, Papua New Guinea  
Dr. Hassan Fouad Mohamed-El-Sayed, Helwan University, Egypt  
Dr. Sanjeevi Pandiyan, Jiangnan University, China  
Dr. P. Mohamed Shakeel, Universiti Teknikal Malaysia Melaka, Malaysia  
Dr. Adhiyaman Manickam, Shenzhen University, China  
Dr. Sathishkumar V. E., Sunchon National University, Republic of Korea  
Dr. Ivan Sanz-Prieto, Universidad Internacional de La Rioja, Spain  
Dr. Gabriel A. Ogunmola, Sharda University, Uzbekistan  
Dr. Dinh-Thuan Do, Sunchon Ton Duc Thang University, Vietnam

Dr. F. H. A. Shibly, South Eastern University of Sri Lanka, Sri Lanka  
Dr. Asma Sbeih, Palestine Ahliya University, Palestine  
Dr. Osamah Ibrahim Khalaf, Al-Nahrain University, Iraq  
Mr. Awais Khan Jumani, ILMA University, Pakistan  
Mr. Melvin Johnson, Google Inc, USA  
Ms. Shruthi Prabhu, Twitter Inc, USA

## National Technical Advisory Committee

Dr. Gunasekaran Raja, Anna University  
Dr. Thangavelu, University of Madras  
Dr. Jayashree P., Anna University  
Dr. Kishore Kumar, NIT Warangal  
Dr. Aravamudhan, PEC  
Dr. Janakiraman, Pondicherry University  
Dr. Vinayak Shukla, CMC  
Dr. Ramanujam R., IMSc  
Dr. Purusothaman, GCT, Coimbatore  
Dr. Preethi, Anna University, Coimbatore  
Dr. Sanjay Kumar Singh, IIT-BHU  
Dr. Harisekar, SanInfotech  
Dr. Sudhakar T., Amrita University  
Dr. Murugan K., VIT  
Dr. Nandhakumar R., VIT  
Dr. Rajesh, Anna University  
Dr. B. B. Gupta, NIT Kurukshetra  
Dr. Balamurugan S., iRCS  
Dr. Karthik S., SNSCT  
Dr. Karthikeyan N., SNSCE  
Dr. Sivaparthipan C. B., Infinite Bullseye  
Dr. Bala Anand Muthu, Thirsty Crowz  
Dr. Vinayak Shukla, CMC  
Dr. Anbarasan M., Sairam Institution of Technology  
Dr. B. Santhosh Kumar, GMR Institute of Technology  
Dr. R. Sabitha, Karunya Institute of Technology and Sciences  
Dr. J. Jayashree, VIT University  
Dr. J. Vijayashree, VIT University  
Dr. S. Velliangiri, CMR Institute of Technology  
Dr. Ebin Deni Raj, IIIT Kottayam  
Dr. Rama Subbareddy, VNRRJIET  
Dr. I. Kala, PSG Institute of Technology and Applied Research  
Dr. V. Praveena, Sri Sakthi Institute of Engineering and Technology

Dr. C. Chandru Vignesh, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology  
Dr. R. Kamalraj, Jain Deemed to be University  
Dr. Renjith V. Ravi, M.E.A. Engineering College  
Mr. Thanjai Vadivel, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology  
Mr. Yamin Khan, Microsoft, India

## **Finance Chairs**

Ms. K. Sornalakshmi, Assistant Professor  
Ms. C. Fancy, Assistant Professor  
Ms. P. Nithyakani, Assistant Professor

## **Publicity and Publication Chairs**

Dr. K. Kottilingam, Associate Professor  
Dr. P. Selvaraj, Assistant Professor  
Dr. Godwin Ponsam, Assistant Professor

## **Organizing Committee**

Dr. Suresh, Professor  
Dr. M. B. Mukesh Krishnan, Associate Professor  
Dr. G. Maragatham, Associate Professor  
Dr. M. Saravanan, Associate Professor  
Dr. N. Arivazhagan, Assistant Professor  
Dr. L. N. B. Srinivas, Assistant Professor  
Mr. A. Arokiaraj Jovith, Assistant Professor  
Dr. D. Hemavathi, Assistant Professor  
Mr. P. Rajasekar, Assistant Professor  
Ms. G. Sujatha, Assistant Professor  
Dr. K. Venkatesh, Assistant Professor  
Mr. M. Anand, Assistant Professor  
Mr. K. Senthil Kumar, Assistant Professor  
Dr. S. Metilda Florence, Assistant Professor  
Mr. S. Murugaanandam, Assistant Professor  
Ms. D. Saveetha, Assistant Professor  
Ms. M. Safa, Assistant Professor

Mr. V. Joseph Raymond, Assistant Professor  
Ms. M. Rajalakshmi, Assistant Professor  
Dr. T. Karthick, Assistant Professor  
Ms. S. Dhivya, Assistant Professor  
Ms. S. Srividhya, Assistant Professor  
Ms. S. Sindhu, Assistant Professor  
Ms. S. Nithiya, Assistant Professor  
Ms. S. Deepanjali, Assistant Professor  
Ms. S. Sivasankari, Assistant Professor  
Ms. D. Sai Santhiya, Assistant Professor  
Ms. V. Vijayalakshmi, Assistant Professor  
Mr. J. Prabaharan, Assistant Professor  
Ms. A. Helen Victoria, Assistant Professor  
Mr. V. Nallarasan, Assistant Professor  
Ms. M. Sangeetha, Assistant Professor

## **Registration Committee**

Dr. D. Hemavathi, Assistant Professor  
Dr. M. B. Mukesh Krishnan, Assistant Professor  
Dr. G. Maragatham, Assistant Professor  
Dr. M. Saravanan, Assistant Professor  
Dr. N. Arivazhagan, Assistant Professor  
Dr. L. N. B. Srinivas, Assistant Professor

# Preface

Artificial intelligence, a sub-field of computer science, has made remarkable transformations in many applications. It aims at creating smart intelligent systems resembling human-based thinking and decision making. Artificial intelligence works better with the huge amount of data that is named as big data where data can either be monitored, optimized, and analyzed for achieving useful predictive inferences. Hence, both AI and big data go hand in hand. Health care is one such field that has continued ongoing challenges to be dealt, with the help of artificial intelligence. Health care aims at adopting new technologies with the massive number of medical reports based on data which are predominantly images. Deep learning gives breakthrough results for computer vision-based applications. Over a relatively short time, the sophistication of deep learning has exploded, leaving researchers a stack of technologies and algorithms to choose from.

The main purpose of this book is to give useful insights through surveys, explore deep learning-based solutions, the latest trends and developments for medical imaging systems. The book covers a detailed theoretical introduction of deep learning for health information mining and medical imaging. This is followed by several surveys that foster the need for deep learning algorithms for health care, in terms of image acquisition, detection, diagnostic analysis, quantitative measurements, and reconstruction of images.

This book focuses on four major areas. The first area is artificial intelligence. It has covered varied types of neural network models for image detection, identification, object detection, classification, and medical image processing techniques for two-dimensional and three-dimensional images. It has covered different types of deep learning approaches.

Second, this book has highlighted the prevailing challenges and areas of big data in the field of medical imaging. This book will serve as a handbook to different types of applications. We also included the algorithms used for advanced data analytics. This is the driving factor for the next-generation intelligent systems.

The third area covered in this book is machine learning concepts. This discusses the significance of ML in various fields. It provides a collective report on the usages

of predictive and prescriptive analysis. Various other contributions along with the existing techniques will help to improvise on ML.

The last area presents the contributions of artificial intelligence in networks and security. The evolution of smart and intelligent networks is the future of network industry. We also discuss at the advanced projections toward security in this field. It also examines the complex network administrative service requirements. This provides solutions for the improvisation on device for network optimization.

This book not only describes the role of AI and its usage, but also gives insights about the future research. This book will provide the detailed introductory to beginners and provides new directions to researchers. The primary audience will be researchers and graduate students, researching in artificial intelligence and health care. The secondary audience includes medical experts, pathologists, and clinical scientists who would leverage the capabilities of artificial intelligence to yield better outcomes to address the challenges of healthcare industry.

Davis, USA  
Kattankulathur, India  
Kattankulathur, India  
May 2021

Dr. Gunasekaran Manogaran  
Dr. A. Shanthini  
Dr. G. Vadivu

# Contents

<b>Classification of Blast Cells in Leukemia Using Digital Image Processing and Machine Learning .....</b>	1
T. Karthick, M. Ramprasath, and M. Sangeetha	
<b>Secured Cloud Storage System Using Auto Passkey Generation .....</b>	19
Sudha Ayatti, K. Gouri, Pavan Kunchur, and Sadhana P. Bangarashetti	
<b>A Novel Multi-Objective Memetic Algorithm for Mining Classifiers .....</b>	33
K. R. Ananthapadmanaban, S. Muruganandam, and Sujatha Srinivasan	
<b>An Intelligent Road Transportation System .....</b>	43
S. Muruganandam, K. R. Ananthapadmanaban, and Sujatha Srinivasan	
<b>Securing Data from Active Attacks in IoT: An Extensive Study .....</b>	51
C. Silpa, G. Niranjana, and K. Ramani	
<b>Building an Enterprise Data Lake for Educational Organizations for Prediction Analytics Using Deep Learning .....</b>	65
Palanivel Kuppusamy and K. Suresh Joseph	
<b>Events of Interest Extraction from Forensic Timeline Using Natural Language Processing (NLP) .....</b>	83
Palash Dusane and G. Sujatha	
<b>Survey on Voice-Recognized Home Automation System Using IOT .....</b>	95
Shreya Mittal, Tarun Bharadwaj, and T. Y. J. Naga Malleswari	
<b>Fingerprinting of Image Files Based on Metadata and Statistical Analysis .....</b>	105
J. Harish Kumar and T. Kirthiga Devi	
<b>Comparative Study of Risk Assessment of COVID-19 Patients with Comorbidities .....</b>	119
Satwika Kesana, Meghana Avadhanam, and T. Y. J. Naga Malleswari	

<b>Neural Network for 3-D Prediction of Breast Cancer Using Lossless Compression of Medical Images .....</b>	133
P. Renukadevi and M. Syed Mohamed	
<b>An Investigation of Paralysis Attack Using Machine Learning Approach .....</b>	143
S. Surya and S. Ramamoorthy	
<b>A Survey on Feature Selection and Classification Techniques for EEG Signal Processing .....</b>	155
K. Saranya, M. Paulraj, and M. Brindha	
<b>Deployment of Sentiment Analysis of Tweets Using Various Classifiers .....</b>	167
Shatakshi Brijpuriya and M. Rajalakshmi	
<b>Predictive Analysis on HRM Data: Determining Employee Promotion Factors Using Random Forest and XGBoost .....</b>	179
D. Vishal Balaji and J. Arunmehru	
<b>A Review on Deaf and Dumb Communication System Based on Various Recognitions Aspect .....</b>	191
G. Arun Prasath and K. Annapurani	
<b>Refined Search and Attribute-Based Encryption Over the Cloud Data .....</b>	205
M. S. Iswariya and K. Annapurani	
<b>Facial Recognition Techniques and Their Applicability to Student Concentration Assessment: A Survey .....</b>	213
Mukul Lata Roy, D. Malathi, and J. D. Dorathi Jayaseeli	
<b>Malware Detection Using API Function Calls .....</b>	227
Bashar Hayani and E. Poovammal	
<b>Implementing Blockchain Technology for Health-Related Early Response Service in Emergency Situations .....</b>	237
Nemanja Zdravković, Milena Bogdanović, Miroslav Trajanović, and Vijayakumar Ponnusamy	
<b>Interpreting Chest X-rays for COVID-19 Applying AI and Deep Learning: A Technical Review .....</b>	245
A. Veronica Nithila Sugirtham and C. Malathy	
<b>Technical Survey on Covid Precautions Monitoring System Using Machine Learning .....</b>	257
K. Lekha, Nisha Yadav, and T. Y. J. Naga Malleswari	
<b>Trust-Based Node Selection Architecture for Software-Defined Networks .....</b>	271
Aditya Kumar and C. Fancy	

Contents	xv
<b>A Review on Video Tampering Analysis and Digital Forensic .....</b>	287
Pavithra Yallamandhala and J. Godwin	
<b>Efficient Sensitive File Encryption Strategy with Access Control and Integrity Auditing .....</b>	295
D. Udhaya Mugil and S. Metilda Florence	
<b>Zone Based Crop Forecast at Tamil Nadu Using Artificial Neural Network .....</b>	305
S. Nithiya, S. Srividhya, and G. Parimala	
<b>Secure Cloud Data Storage and Retrieval System Using Regenerating Code .....</b>	313
S. Yuvaraman and D. Saveetha	
<b>Visual Question Answering System for Relational Networks .....</b>	323
M. Fathima Najiya and N. Arivazhagan	
<b>Crop Yield Prediction on Soybean Crop Applying Multi-layer Stacked Ensemble Learning Technique .....</b>	335
S. Iniyi and R. Jebakumar	
<b>A Comparative Analysis to Obtain Unique Device Fingerprinting .....</b>	349
T. Sabhanayagam	
<b>Adware and Spyware Detection Using Classification and Association .....</b>	355
Kalyan Anumula and Joseph Raymond	
<b>An Effective Approach to Explore Vulnerabilities in Android Application and Perform Social Engineering Attack .....</b>	363
Joseph Raymond and P. Selvaraj	
<b>History of Deception Detection Techniques .....</b>	373
D. Viji, Nikita Gupta, and Kunal H. Parekh	
<b>Climate Change Misinformation Detection System .....</b>	389
Sagar Saxena and K. Nimala	
<b>Phishing Website Detection and Classification .....</b>	401
D. Viji, Vaibhav Dixit, and Vishal Jha	
<b>Generations of Wireless Mobile Networks: An Overview .....</b>	413
Burla Sai Teja and Vivia Mary John	
<b>Quantum Networking—Design Challenges .....</b>	419
S. Mohammed Rifas and Vivia Mary John	
<b>A Comparative Analysis of Classical Machine Learning and Deep Learning Approaches for Diabetic Peripheral Neuropathy Prediction .....</b>	427
R. Usharani and A. Shanthini	

<b>Zone-Based Multi-clustering in Non-rechargeable Wireless Sensor Network Using Optimization Technique .....</b>	437
S. Srividhya, M. Rajalakshmi, L. Paul Jasmine Rani, and V. Rajaram	
<b>A Review and Design of Depression and Suicide Detection Model Through Social Media Analytics .....</b>	443
Michelle Catherina Prince and L. N. B. Srinivas	
<b>Artificial Intelligence Used in Accident Detection Using YOLO .....</b>	457
S. V. Gautham and D. Hemavathi	
<b>An Ensemble Learning Method on Mammogram Data for Breast Cancer Prediction—Comparative Study .....</b>	469
T. Sreehari and S. Sindhu	
<b>Media Bias Detection Using Sentimental Analysis and Clustering Algorithms .....</b>	485
Sachin Rawat and G. Vadivu	
<b>A Review on Video Summarization .....</b>	495
R. V. Krishna Vamsi and Dhivya Subburaman	
<b>Prediction and Grading Approach for Cataract Diagnosis Using Deep Convolutional Neural Network .....</b>	505
P. Nithyakani, R. Kheerthana, A. Shrikrishna, S. Selva Ganesan, and Anurag Wadhwa	
<b>A Technical Review and Framework Design for Influence Extraction from Social Networks .....</b>	515
Akash Saini and K. Sornalakshmi	
<b>Sentiment Diagnosis in Text Using Convolutional Neural Network .....</b>	525
C. Sindhu, Shilpi Adak, and Soumya Celina Tigga	
<b>Enhanced Movie Recommender System Using Hybrid Approach .....</b>	539
R. Lavanya, V. S. Bharat Raam, and Nikil Pillaithambi	
<b>Optimization of CNN in Capsule Networks for Alzheimer's Disease Prediction Using CT Images .....</b>	551
P. R. Ananya, Vedika Pachisia, and S. Ushasukhanya	
<b>Red Lesion Detection in Color Fundus Images for Diabetic Retinopathy Detection .....</b>	561
P. Saranya, K. M. Umamaheswari, Satish Chandra Patnaik, and Jayvardhan Singh Patyal	
<b>Video Based Human Gait Activity Recognition Using Fusion of Deep Learning Architectures .....</b>	571
P. Nithyakani and M. Ferni Ukrat	

Contents	xvii
<b>A Deep Learning Based Palmar Vein Recognition: Transfer Learning and Feature Learning Approaches .....</b>	581
M. Rajalakshmi and K. Annapurani	
<b>Survey of Popular Linear Dimensionality Reduction Techniques .....</b>	593
Anne Lourdu Grace and M. Thenmozhi	
<b>Healthcare Monitoring System Using Medical Smart Card .....</b>	605
G. Sujatha, D. Hemavathi, K. Sornalakshmi, and S. Sindhu	
<b>Enhanced Energy Distributed Unequal Clustering Protocol for Wireless Ad Hoc Sensor Networks .....</b>	621
G. Parimala, A. Razia Sulthana, and S. Nithiya	
<b>Classification and Prediction of Leaf Diseases in Grape Using Optimized CNN .....</b>	631
J. Sujithra and M. Ferni Ukrat	
<b>Diabetic Retinopathy Image Segmentation Using Region-Based Convolutional Neural Network .....</b>	637
D. Vanusha and B. Amutha	
<b>A Secure and Reliable IoT-Edge Framework with Blockchains for Smart MicroGrids .....</b>	651
Rijo Jackson Tom, Vivia Mary John, and G. Renukadevi	
<b>Sentiment Analysis on the Performance of Engineering Students in Continuous Assessment Evaluation System: A Non-parametric Approach Using Kruskal-Wallis Method .....</b>	671
A. Vijay Bharath, A. Shanthini, and A. Subbarayan	
<b>Centralized and Decentralized Data Backup Approaches .....</b>	687
Rahul Kumar and K. Venkatesh	
<b>Author Index .....</b>	699

# Editors and Contributors

## About the Editors

**Dr. Gunasekaran Manogaran** is currently working as Big Data Scientist at University of California, Davis, USA. He is also Adjunct Assistant Professor, Department of Computer Science and Information Engineering, Asia University, Taiwan, and Adjunct Faculty, in School of Computing, SRM Institute of Science and Technology, Kattankulathur, India. He is Visiting Researcher/Scientist at the University of La Frontera, Colombia, and the International University of La Rioja, Spain. He received his Ph.D. from the Vellore Institute of Technology University, India. He received his Bachelor of Engineering and Master of Technology from Anna University, India, and Vellore Institute of Technology University, India, respectively. He is author/co-author of more than 100 papers in conferences, book chapters, and journals, including *IEEE Transactions on Industrial Informatics*, *IEEE Transactions on Computational Social Systems*, *IEEE Internet of Things*, *IEEE Intelligent System*, *IEEE Access*, *ACM Transactions on Multimedia Computing, Communications, and Applications*.

**Dr. A. Shanthini** is currently working as Associate Professor in the Department of Information and Technology, SRM Institute of Science and Technology, Kattankulathur Campus, India. She received her Bachelor of Engineering, Master of Engineering and Ph.D. from Annamalai University, India. Her current research interests include IoT, machine learning, and deep learning in health care. She published two patents in the Patent Office Journal and one in Australian patent in the year 2018 and 2020, respectively. Currently, she is Principal Investigator of the project titled “Prog-nosis of Microaneurysm, and early diagnosis system for non-proliferative Diabetic Retinopathy using Deep Convolutional neural network” sponsored by SPARC-IIITK, MHRD, Government of India, which is associated with University of California, Davis Campus, USA, and SRM IST, India, for 67 Lakhs in March 2019. She is author/co-author in 12 research articles in international journals and conferences, including SCI and Scopus indexed papers. She is Active Member of IEEE, ACM, and ISC.

**Dr. G. Vadivu** is working in the teaching profession for more than two decades. Currently, she is designated Professor and Head in the Department of Information Technology at SRM Institute of Science and Technology, Kattankulathur Campus, India. Her research areas include big data analytics, semantic web, data mining, and database systems. She published more than 30 research articles in a reputed journal listed in SCIE and Scopus. She has organized UGC sponsored workshop on .NET Technologies during 2006 and 2009. She received the Best Teaching Faculty Award and Certificate of Appreciation for the journal published in the year 2012. Also, she has completed Oracle Certification, Certification in Database Administration-Microsoft Technology Association, High-Impact Teaching Skills Certified by Dale Carnegie, and IBM-DB2 certification.

## Contributors

**Shilpi Adak** Department of CSE, SRM Institute of Science and Technology, Kattankulathur, India

**B. Amutha** Department of CSE, SRM Institute of Science and Technology, Chennai, India

**K. R. Ananthapadmanaban** Faculty of Science and Humanities, SRM Institute of Science and Technology, Chennai, Tamilnadu, India

**P. R. Ananya** Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, India

**K. Annapurani** Department of Networking and Communications, School of Computing, SRM Institute of Science and Technology, Kattankulathur, Kanchipuram, Tamil Nadu, India

**Kalyan Anumula** Department of Information Technology, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, India

**N. Arivazhagan** Department of Information Technology, SRM University Chennai, Chennai, India

**G. Arun Prasath** Department of Networking and Communications, SRM Institute of Science and Technology, Kattankulathur, India

**J. Arunnehru** Department of Computer Science and Engineering, SRM Institute of Science and Technology, Vadapalani Campus, Chennai, India

**Meghana Avadhanam** Department of CSE, SRMIST, Chennai, India

**Sudha Ayatti** Department of Computer Science and Engineering, KLS Gogte Institute of Technology, Belagavi, India

**Sadhana P. Bangarashetti** Department of Information Science and Engineering, BEC, Bagalkot, India

**Tarun Bharadwaj** SRM Institute of Science and Technology, Chennai, India

**Milena Bogdanović** Faculty of Information Technology, Belgrade Metropolitan University, Belgrade, Serbia

**Shatakshi Brijpuriya** SRM Institute of Science and Technology, Chennai, India

**M. Brindha** Department of Computer Science and Engineering, Chennai Institute of Technology, Chennai, Tamil Nadu, India

**Vaibhav Dixit** Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai, India

**J. D. Dorathi Jayaseeli** Department of Computer Science and Engineering, SRM University, Chennai, India

**Palash Dusane** Department of Information Technology, SRM Institute of Science and Technology, Chennai, India

**C. Fancy** SRM Institute of Science and Technology, Chennai, India

**M. Fathima Najiya** Department of Information Technology, SRM University, Chennai, India

**M. Ferni Ukrat** School of Computing, Department of Information Technology, SRM Institute of Science and Technology, Chennai, India

**S. Selva Ganesan** SRM Institute of Science and Technology, Chennai, India

**S. V. Gautham** School of Computing, SRM Institute of Science and Technology, Chennai, India

**J. Godwin** Department of Information Technology, SRM Institute of Science and Technology, Kattankulathur, India

**K. Gouri** Department of Computer Science and Engineering, KLS Gogte Institute of Technology, Belagavi, India

**Anne Lourdu Grace** Department of Computer Science, SRM University, Chennai, India

**Nikita Gupta** Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, India

**J. Harish Kumar** SRM Institute of Science and Technology, Chennai, India

**Bashar Hayani** Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, India

**D. Hemavathi** School of Computing, SRM Institute of Science and Technology, Chennai, India

**S. Iniyar** SRM Institute of Science and Technology, Chennai, India

**M. S. Iswariya** Department of Networking and Communications, SRM Institute of Science and Technology, Kattankulathur, Kanchipuram, Tamil Nadu, India

**R. Jebakumar** SRM Institute of Science and Technology, Chennai, India

**Vishal Jha** Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai, India

**Vivian Mary John** Department of Computer Science and Engineering, CMR Institute of Technology, Bengaluru, India

**T. Karthick** Department of Information Technology, SRM Institute of Science and Technology, Chennai, India

**Satwika Kesana** Department of CSE, SRMIST, Chennai, India

**R. Kheerthan** SRM Institute of Science and Technology, Chennai, India

**T. Kirthiga Devi** SRM Institute of Science and Technology, Chennai, India

**Aditya Kumar** SRM Institute of Science and Technology, Chennai, India

**Rahul Kumar** School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai, India

**Pavan Kunchur** Department of Computer Science and Engineering, KLS Gogte Institute of Technology, Belagavi, India

**Palanivel Kuppusamy** Pondicherry University, Puducherry, India

**R. Lavanya** Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai, India

**K. Lekha** Department of CSE, SRMIST, Kattangulathur, India

**D. Malathi** Department of Computer Science and Engineering, SRM University, Chennai, India

**C. Malathy** Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India

**S. Metilda Florence** Department of Information Technology, SRM Institute of Science and Technology, Kattankulathur, India

**Shreya Mittal** SRM Institute of Science and Technology, Chennai, India

**S. Muruganandam** Department of Computer Science, SRM Institute for Training and Development, Chennai, Tamilnadu, India

**T. Y. J. Naga Malleswari** Department of CSE, SRM Institute of Science and Technology, Kattangulathur, Chennai, India

**K. Nimala** SRM Institute of Science and Technology, Kattankulathur, India

**G. Niranjana** Department of CSE, SRMIST, Chennai, India

**S. Nithiya** SRM Institute of Science and Technology, Kattankulathur, Chennai, India

**P. Nithyakani** School of Computing, Department of Information Technology, SRM Institute of Science and Technology, Chennai, India

**Vedika Pachisia** Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, India

**Kunal H. Parekh** Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, India

**G. Parimala** SRM Institute of Science and Technology, Kattankulathur, Chennai, India

**Satish Chandra Patnaik** Department of Computer Science and Engineering, SRM Institute of Science and Technology, Tamilnadu, India

**Jayvardhan Singh Patyal** Department of Computer Science and Engineering, SRM Institute of Science and Technology, Tamilnadu, India

**M. Paulraj** Department of Computer Science and Engineering, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India

**Nikil Pillaithambi** Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai, India

**Vijayakumar Ponnusamy** Department of ECE, SRM Institute of Science and Technology, Kattankulathur, India

**E. Poovammal** Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, India

**Michelle Catherine Prince** Department of Information Technology, College of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India

**V. S. Bharat Raam** Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai, India

**M. Rajalakshmi** School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai, India

**V. Rajaram** Department of Information Technology, Sri Venkateswara College of Engineering, Sriperumbuthur, India

**S. Ramamoorthy** Associate professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India

**K. Ramani** Department of IT, SVEC, Tirupati, India

**M. Ramprasath** Department of Information Technology, SRM Institute of Science and Technology, Chennai, India

**L. Paul Jasmine Rani** Department of Computer Science and Engineering, Rajalakshmi Institute of Science and Technology, Chennai, Tamil Nadu, India

**Sachin Rawat** IT Department, SRMIST Kattankulathur, Chennai, India

**Joseph Raymond** Department of Information Technology, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, India

**A. Razia Sulthana** Birla Institute of Science and Technology-Pilani, Dubai Campus, United Arab Emirates

**P. Renukadevi** Department of Information Technology, Sri Ram Nallamani Yadava College of Arts and Science (Affiliated to Manonmaniam Sundaranar University, Tirunelveli), Tenkasi, India

**S. Mohammed Rifas** CMR Institute of Technology, Bengaluru, Karnataka, India

**Mukul Lata Roy** Department of Computer Science and Engineering, SRM University, Chennai, India

**T. Sabhanayagam** Department of Computing Technologies, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India

**Akash Saini** Department of Data Science and Business Systems, School of Computing, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Chennai, TN, India

**M. Sangeetha** Department of Information Technology, SRM Institute of Science and Technology, Chennai, India

**K. Saranya** Department of Computer Science and Engineering, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India

**P. Saranya** Department of Computer Science and Engineering, SRM Institute of Science and Technology, Tamilnadu, India

**D. Saveetha** Department of Information Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India

**Sagar Saxena** SRM Institute of Science and Technology, Kattankulathur, India

**P. Selvaraj** SRM Institute of Science and Technology, Kattankulathur, India

**A. Shanthini** Department of Data Science and Business Systems, School of Computing, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India

**A. Shrikrishna** SRM Institute of Science and Technology, Chennai, India

**C. Silpa** Department of CSE, SRMIST, Chennai, India

**C. Sindhu** Department of CSE, SRM Institute of Science and Technology, Kattankulathur, Chennai, India

**S. Sindhu** Department of Information Technology, SRM University, Chennai, India

**K. Sornalakshmi** Department of Data Science and Business Systems, School of Computing, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Chennai, TN, India

**T. Sreehari** Department of Information Technology, SRM University, Chennai, India

**Sujatha Srinivasan** Department of Computer Science, SRM Institute for Training and Development, Chennai, Tamilnadu, India

**S. Srividhya** School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai, India

**L. N. B. Srinivas** Department of Information Technology, College of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India

**A. Subbarayan** Directorate of Research, SRM Institute of Science and Technology, Tamilnadu, Chennai, India

**Dhivya Subburaman** Department of Information Technology, SRM Institute of Science and Technology, Kattankulathur, India

**A. Veronica Nithila Sugirtham** Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu, India

**G. Sujatha** Department of Information Technology, SRM Institute of Science and Technology, Chennai, India

**J. Sujithra** School of Computing, SRM Institute of Science and Technology, Chennai, India

**K. Suresh Joseph** Pondicherry University, Puducherry, India

**S. Surya** Research Scholar, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Chennai, India

**M. Syed Mohamed** Department of Information Technology, Sri Ram Nallamani Yadava College of Arts and Science (Affiliated to Manonmaniam Sundaranar University,Tirunelveli), Tenkasi, India

**Burla Sai Teja** CMR Institute of Technology, Bengaluru, Karnataka, India

**M. Thenmozhi** Department of Information Technology, SRM University, Chennai, India

**Soumya Celina Tingga** Department of CSE, SRM Institute of Science and Technology, Kattankulathur, India

**Rijo Jackson Tom** Department of Computer Science and Engineering, CMR Institute of Technology, Bengaluru, India

**Miroslav Trajanović** Faculty of Mechanical Engineering, University of Niš, Niš, Serbia

**D. Udhaya Mugil** Department of Information Technology, SRM Institute of Science and Technology, Kattankulathur, India

**K. M. Umamaheswari** Department of Computer Science and Engineering, SRM Institute of Science and Technology, Tamilnadu, India

**R. Usharani** Department of Data Science and Business Systems, School of Computing, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India

**S. Ushasukhanya** Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, India

**G. Vadivu** IT Department, SRMIST Kattankulathur, Chennai, India

**R. V. Krishna Vamsi** M. Tech (Big Data Analytics), Department of Information Technology, SRM Institute of Science and Technology, Kattankulathur, India

**D. Vanusha** Department of CSE, SRM Institute of Science and Technology, Chennai, India

**K. Venkatesh** School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai, India

**A. Vijay Bharath** Department of Data Science and Business Systems, SRM Institute of Science and Technology, Tamilnadu, Chennai, India

**D. Viji** Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai, India

**D. Vishal Balaji** Department of Computer Science and Engineering, SRM Institute of Science and Technology, Vadapalani Campus, Chennai, India

**Anurag Wadhwa** SRM Institute of Science and Technology, Chennai, India

**Nisha Yadav** Department of CSE, SRMIST, Kattangulathur, India

**Pavithra Yallamandhala** Department of Information Technology, SRM Institute of Science and Technology, Kattankulathur, India

**S. Yuvaraman** Department of Information Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamil Nadu, India

**Nemanja Zdravković** Faculty of Information Technology, Belgrade Metropolitan University, Belgrade, Serbia

# Classification of Blast Cells in Leukemia Using Digital Image Processing and Machine Learning



T. Karthick, M. Ramprasath, and M. Sangeetha

**Abstract** Leukemia is a fast-developing type of blood cancer that gets worse quickly in the children and adults and needs prompt treatment. Thus, this work displays an attempt that has been made to design a fast and cost-effective computer-aided system for classification of blast cells in leukemia using digital image processing and machine learning. The white blood cells in the microscopic images of blood smears are initially extracted by the Otsu's method, and a cell separation algorithm is applied to break up the overlapped cells. Subsequently, several features are extracted from the whole cell, nucleus, and cytoplasm. The resulted overall accuracy of 90.32% for identification of leukemia affected cells is achieved. Therefore, the current system yielded very promising results in terms of classification accuracy of leukemia-affected cells that can be distinguished from the unaffected ones. However, future research is still needed to develop the diagnostic accuracy.

**Keywords** Leukemia · Machine learning · Diagnostic accuracy · Computer-aided system

## 1 Introduction

The detection of cancer is always a major issue for the pathologists and medical practitioners for diagnosis and treatment. The manual identification of leukemia from microscopic biopsy images may vary from expert to expert depending on their expertise and errors may include lack of specific and accurate quantitative measures

---

T. Karthick (✉) · M. Ramprasath · M. Sangeetha

Department of Information Technology, SRM Institute of Science and Technology,  
Kattankulathur, Chennai, India  
e-mail: [karthict@srmist.edu.in](mailto:karthict@srmist.edu.in)

M. Ramprasath  
e-mail: [ramprasm@srmist.edu.in](mailto:ramprasm@srmist.edu.in)

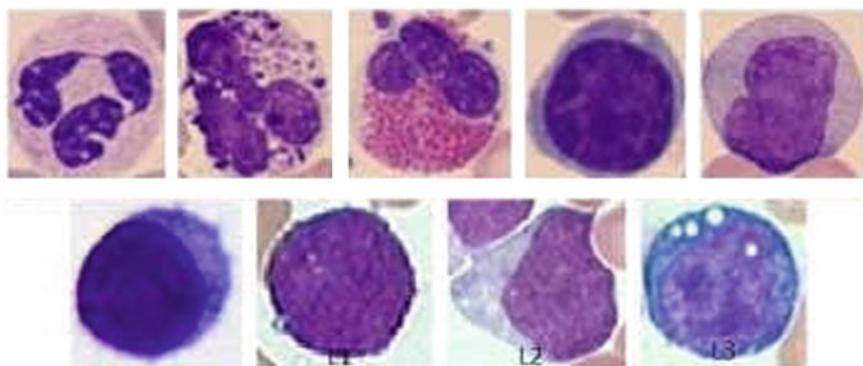
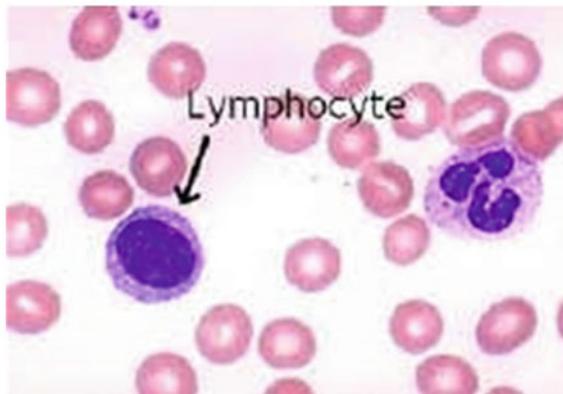
M. Sangeetha  
e-mail: [sangeetk@srmist.edu.in](mailto:sangeetk@srmist.edu.in)

to classify the biopsy images as healthy or leukemic one [1]. The probability of getting affected by leukemia may be due to tobacco addiction, diet habits. Curing the disease is increased due to recent combined advancement in medicine and engineering. The treatment selection depends on its level of malignancy. Thus, the microscopic tissue structure of patient is examined for accurate detection. The malignant level of the disease is based on the amount of blast cells. Hence, chronic leukemia is the difficult one to treat. The analysis of white blood cells (WBCs) allows for the detection of acute lymphoblastic leukemia (ALL), a blood cancer that can be fatal if left untreated (Fig. 1).

The round, uniform nucleus and small amount of cytoplasm surrounding it are the best identifying characteristics for this cell. (Bottom) A comparison between lymphocytes subring from ALL: a healthy lymphocyte, followed by lymphoblasts classified as  $L_1$ ,  $L_2$ , and  $L_3$ , respectively (Fig. 2).

Lymphocytes are regularly shaped and have a compact nucleus with regular and continuous edges, whereas lymphoblasts are irregularly shaped and contain small

**Fig. 1** Lymphocyte



**Fig. 2** Comparison between healthy and affected Leukocytes

cavities in the cytoplasm, termed vacuoles, and spherical particles within the nucleus, termed nucleoli.

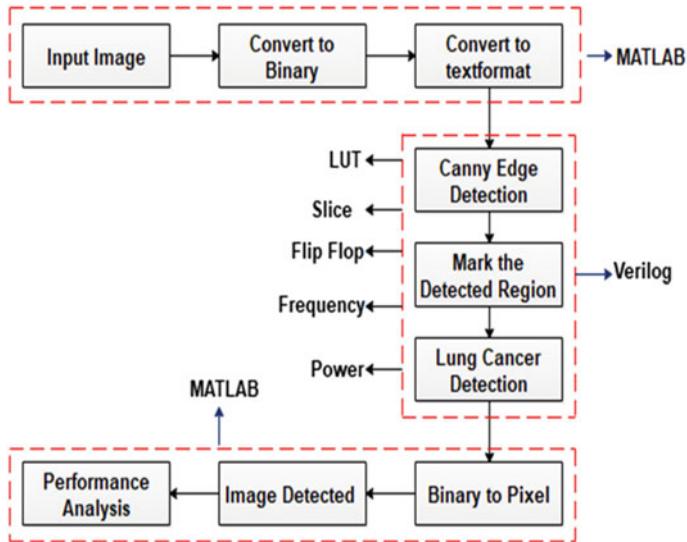
For the monitoring of the tumor micro-environment at a cell level, cell identification is an integral part of automatic photographs anal pipelines [1]. It is a difficult issue, since the cells in the tumor terrestrial region differ in their scale, form, and morphology. Cell identification is also superior to segmentation so pathologists can accurately capture the simple reality (punctures rather than free-hand drawings). Owing to their positive outcomes when working with large cohort, deep learning approaches have become a method of choice [2]. Cires et al. [3] published one of the early methods of identifying mitosis in photographs of breast cancer. They trained a CNN in order to reduce the chance of mitosis or non-mitosis for any pixel. The spatially restricted CNN (SCCNN) was augmented by two layers to be added to totally linked layer in Sirinukunwattana et al. [1]. The probability of a pixel being the nucleus core is determined by these layers.

This structure was expanded by Kashif et al. [4] with the addition of hand-crafting elements which slightly enhanced the  $F_1$  scoring and echoed the accuracy. Chen et al. [5] suggested a depth regression network that acquires its probability map parameters provided by a mitotic-cell segmentation mask. Xie et al. [6] reversed the map to follow the maximum detection of local [6]. Xue et al. [7] indicated that an embedded vector could be regressed to retrieve sparse cell positions combined with identification. Xie et al. [8] suggested systematic regression to acquire maps of higher values close to the centers of the cell. For the improvement of probability maps, Tofighi et al. [9] recently used form priors by manual labeling on core boundaries. In probability maps, local maximum is then determined the location of cells.

In order to identify the 17 myelogenous blood cell forms, Markiewicz et al. [10] use different features related to form, geometry and histograms in tandem with the support vector machine (SVM) classifier. In their study, Aimi et al. [11] demonstrated the good performance of the multilayer perceptron network (MLP) trained by the Bayesian Control algorithm. Mohapatra et al. [12] suggested a new system in which malignant and harmless blood smear cells were graded.

It has also shown that ensemble classifiers are superior to other classification strategies such as NB, KNN, MLP, radial base feature network (RBFN), and SVM for the classification of cells in blood smears. In the field of medical photographs processing, local binary patterns (LBPs) [13] made important strides in the use of texture descriptors. Vanika et al. [14] carried out a comparative analysis for classification of dysfunctional lymphocytes in cell images between form and LBP characteristics, demonstrating that LBP functions better than form for classification.

Madhloom et al. [15] experimented with pattern recognition to distinguish normal lymphocytes and acute lymphoblastic leukemia. The technique reveals that pattern identification of normal lymphocytes and acute lymphoblastic leukemia is implemented. Cho et al. [16] suggest the comparative study of the features and machine learning classifiers to predict cancer. The correlation coefficient, the distance to Euclid, the cosine coefficient, the data collection, reciprocal knowledge, and the noise ratio of a signal to derive the features are used. They have been used. The



**Fig. 3** DRAM-Optimal Adder-CED methodology

description is rendered using the nerve history, the network, the network and adaptive function diagram, the SVM, the decision tree, and the KNN. Their study shows that the best finding was neuronal gradient propagation with Pearson's correlation. Discrimination against the regular B-lymphoblast cells is particularly difficult since all cell groups are exactly the same morphologically.

The DRAM-Optimal Adder-CED technique is shown in Fig. 3. The first step is to read input images from the MATLAB tool and to prepare a binary format from this data. Next, it is written in text format for the binary value. The output of the text format provides the Verilog input. CED technology, which uses Gaussian filtering to smooth out the inside picture and erase artificial boundaries, detects lung cancer cells [17–20].

It is critical that the edge of the cells affected by cancer in the lung is established for the diagnosis of lung cancer. Canny edge detection is derived from the characteristics of the lung picture. The CED technology spread achieves good performance. The CED can measure low and high thresholds. The following five steps of the CED technique are: 1. The vertical and horizontal differential is measured. 2. Gradient path. 3. Deletion of non-maximum regression (NMS) regression. 4. Low and high threshold values calculation. 5. Threshold for hysteresis [21].

At the outset, a median filtering was built to filter the noise in the image. The gradient measurement is carried out with a mask called finite impulse response (FIR). In the third stage, each pixel suppresses the pixel with a low gradient relative to the size and orientation of the neighboring pixel. First, high and low thresholds are measured with the aid of an entire image gradient histogram. DRAM is used to react without a clock in the horizontal and vertical module to handle and control point

inputs. The interface of the DRAM system consists of two ports, with the following signals: address, control and data.

## 2 Proposed Methodology

### 2.1 WBC Identification

WBC identification consists of several phases: 1. Conversion from RGB to CMYK color model. 2. Histogram equalization or contrast stretching operations. 3. Segmentation by threshold using Zack algorithm. 4. Background removal operation.

#### 2.1.1 Conversion from RGB to CMYK Color Model

The initial image is in the RGB scale. We first convert the image from RGB to CMYK scale because the leukocytes have no yellow component and appear completely black in the yellow component.

#### 2.1.2 Histogram Equalization or Contrast Stretching Operations

Contrast stretching is done to increase the contrast so that the leukocytes can be easily thresholded.

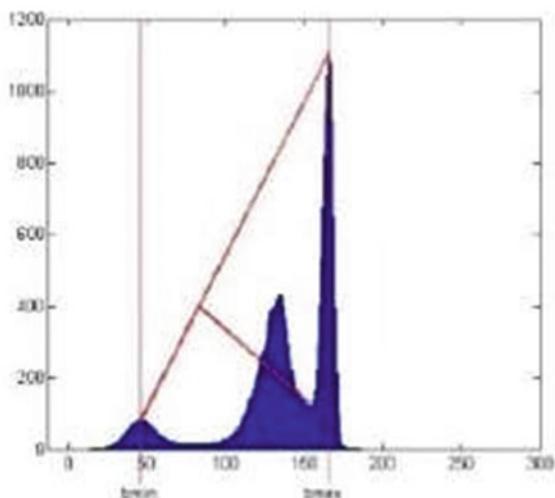
#### 2.1.3 Segmentation by Threshold Using Zack Algorithm

The triangle method is applied to the image histogram, resulting in a straight line that connects the highest histogram value ( $h[b_{\max}]$ ) and the lowest histogram value ( $h[b_{\min}]$ ), where  $b_{\max}$  and  $b_{\min}$  indicate the values of the gray levels where the histogram reaches its maximum and minimum, respectively. Then, the distance between the marked line and the histogram values between  $b_{\min}$  and  $b_{\max}$  is calculated. The intensity value where the distance  $d$  reaches its maximum denotes the threshold value (Fig. 4).

#### 2.1.4 Background Removal Operation

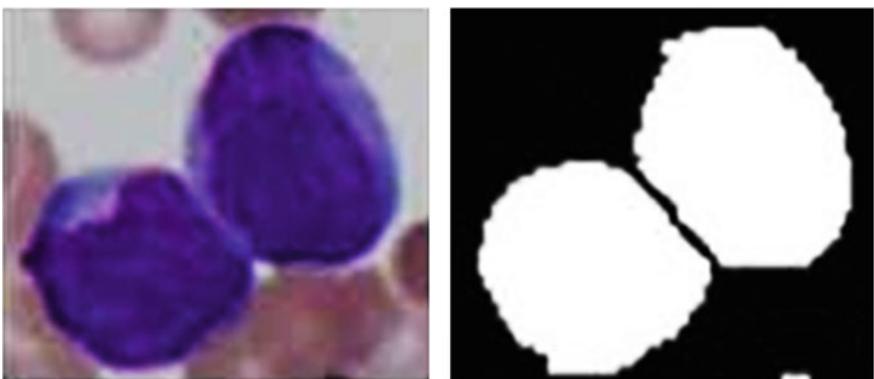
It consists of removing the background of the image, i.e., everything except the leukocytes. The threshold that we obtain from Zack algorithm is used for background removal.

**Fig. 4** Example Zack algorithm



### 3 Identification and Separation of Grouped Leukocytes

An important problem for the analysis of blood images is the presence of leukocyte agglomerates. Only in this phase, we can detect and separate leukocyte agglomerates, because in the previous phase we produced an image containing only the WBCs. Watershed segmentation is then applied to separate the grouped leukocytes (Fig. 5).

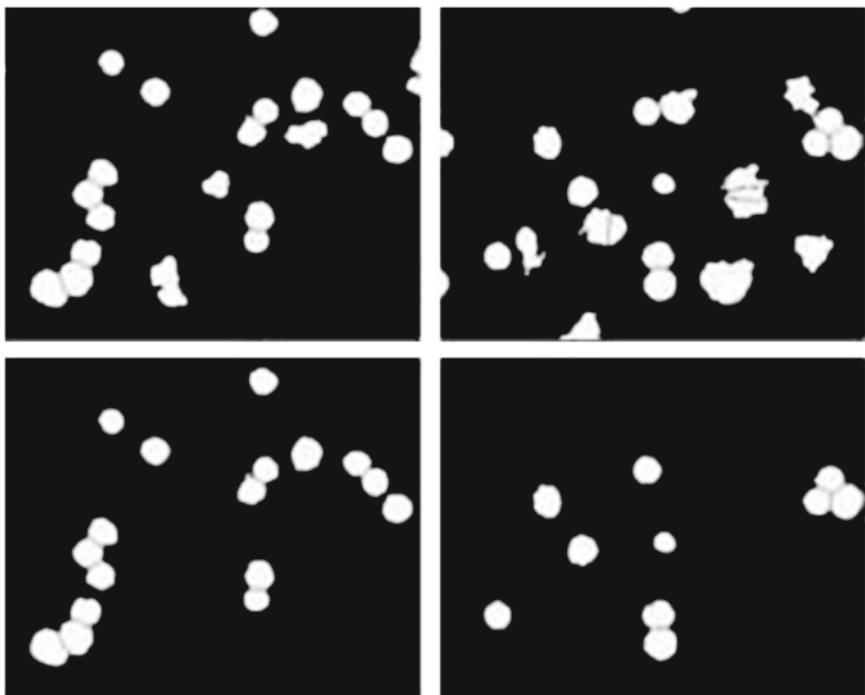


**Fig. 5** Image before and after applying watershed algorithm

### 3.1 Image Cleaning

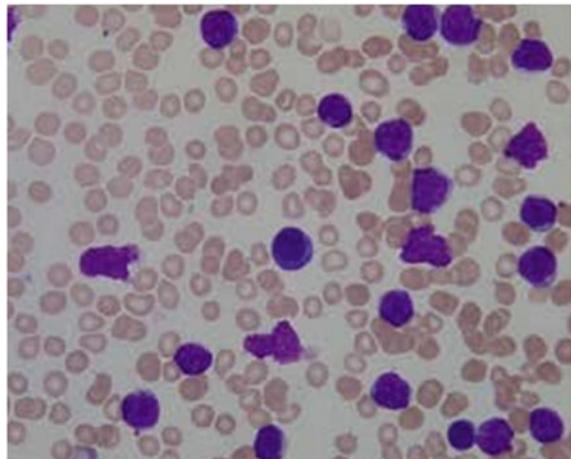
Image cleaning requires the removal of all of the leukocytes located on the edge of the image and all abnormal components (non-leukocytes), which prevents errors in the later stages of the analysis process. Cleaning the image edge is a simple operation. First, the size of the area and the size of the convex area are computed for each leukocyte. The size of the area is used to calculate the mean area, which is necessary to determine and eliminate components with irregular dimensions (Figs. 6 and 7).

In the arrow removal process, we give the arrow same color as the background. Part of arrow in the background is given color as the background and part of arrow on the blood cell is given same color as the blood cell.



**Fig. 6** Final separation results and image cleaning results

**Fig. 7** Image after arrow removal



## 4 System Design

### 4.1 Learning Algorithms

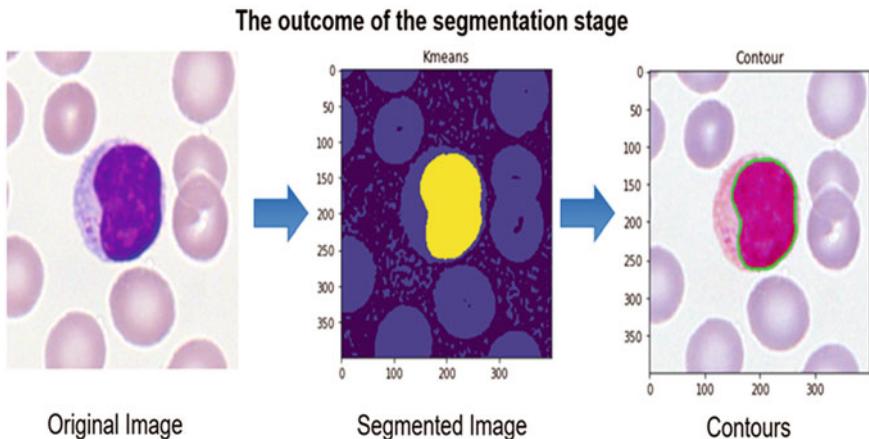
#### 4.1.1 K-means Clustering

It is an iterative algorithm that tries to partition the dataset into K predefined distinct non-overlapping subgroups or clusters where each data point belongs to only one group. It tries to make the inter-cluster data points as similar as possible while also keeping the clusters as different as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid is at the minimum. The less variation we have within clusters, the more homogeneous the data points are within the same cluster.

Feature extraction is an important aspect in the pattern recognition and machine learning tasks in which the visual information, morphological changes in cells, shown in the peripheral blood smear is the first step for diagnosis acute leukemia.

Thus, the morphological features can be used as a magic tool to distinguish between cells:

- *Mean*—Mean is the average value of pixels within the region of interest that represents the brightness of the image.
- *Area*—The space occupied by surface of a blast cell.
- *Perimeter*—The distance around blast cell.
- *Standard Deviation*—By using the standard deviation, we can determine a way of analyzing what is normal, extra-large, or extra small.
- *Roundness*—Roundness is the measure of how closely the shape of an object approaches that of a mathematically perfect circle. It shows the deviation in the shape of the cells.



**Fig. 8** Segmentation stages

- *Entropy*—It is used to measure the randomness or disorder of an image.
- *Skewness*—It is a measure of the lack of symmetry. The zero value indicates that the distribution of the intensity values is relatively equal on both sides of the mean.
- *Kurtosis*—Measures the peak of the distribution of the intensity values around the mean.
- *Variance*—It is defined as the average of the squared differences from the mean (Figs. 8, 9 and Tables 1, 2).

#### 4.1.2 Support Vector Machines (SVM)

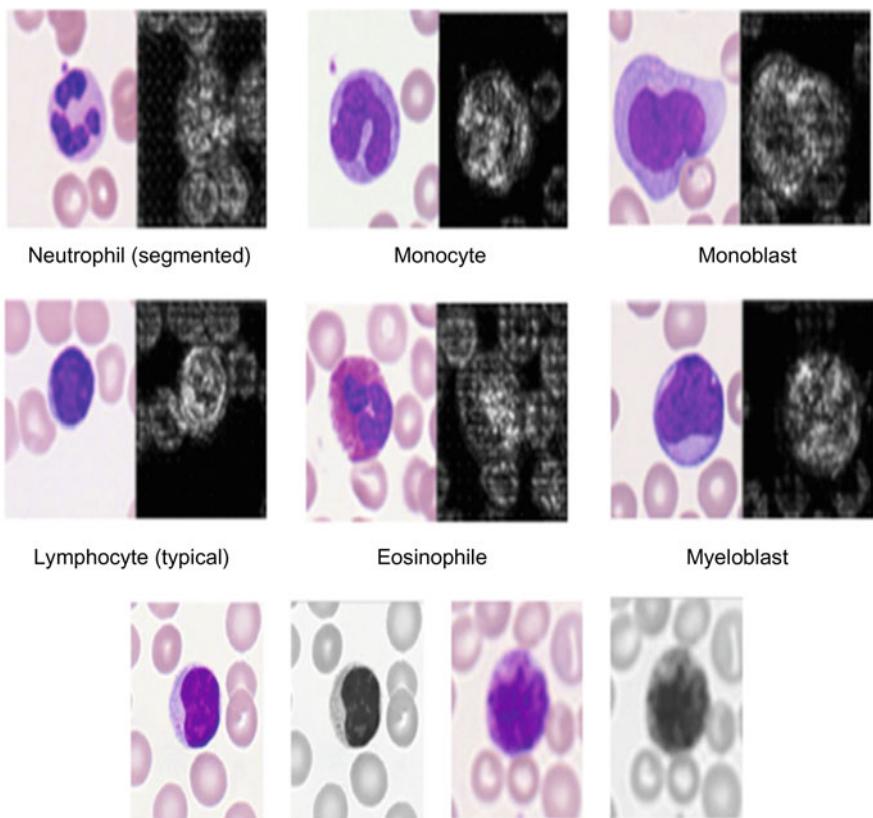
To classify cells, the support vector machine classifier has been used. Support vector machines are a type of supervised machine learning algorithm that provides analysis of data for classification and regression analysis. While they can be used for regression, SVM is mostly used for classification.

The basic principle behind the working of support vector machines is simple—the above extracted table is the result from the K-means clustering algorithm which is the dataset for SVM algorithm, following which will create a hyperplane that separates the dataset into classes (Fig. 10 and Table 3).

## 4.2 Modules

### 4.2.1 Data Acquisition

The dataset that we are using in our project is provided by the TCIA (The Cancer Imaging Archive). It is a service that de-identifies and hosts a large archive of medical



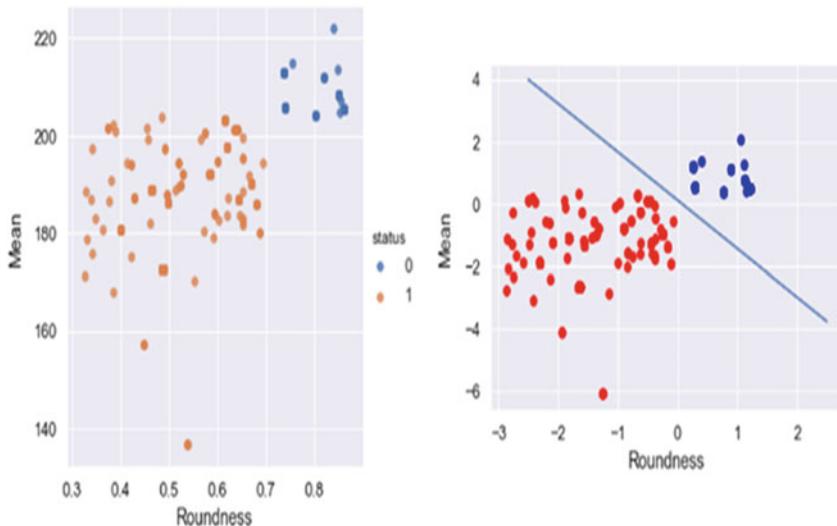
**Fig. 9** Feature extraction

**Table 1** Different cell and its rounded value

Different type of cells	Roundness value
Myelocytes/band	$0.22 < R < 0.33$
Meta myelocytes	$0.34 < R < 0.39$
Eosinophil	$0.40 < R < 0.49$
Pro myelocytes	$0.50 < R < 0.56$
Basophils	$0.57 < R < 0.60$
Monocytes	$0.60 < R < 0.68$
Neutrophils	$0.69 < R < 0.75$
Lymphocytes	$0.81 < R < 0.95$

**Table 2** Feature extracted

	Mean	Area	Perimeter	Roundness	Entropy	Kurtosis	Skew	STD	VAR	Status
0	205.5202188	12997	435.6467495	0.860912067	17.22800406	2.975081768	-1.936645921	52.08100185	2712.430754	0
1	212.895625	12795	467.445739	0.736143551	17.23021883	4.568484838	-2.337812479	51.89678283	2693.276068	0
2	205.5202188	12997	435.6467495	0.860912067	17.22800406	2.975081768	-1.936645921	52.08100185	2712.430754	0
3	205.5202188	12997	435.6467495	0.860912067	17.22800406	2.975081768	-1.936645921	52.08100185	2712.430754	0
4	205.9416563	12996	469.9310203	0.739820067	17.22485077	3.261048173	-1.969678928	53.12740185	2822.520827	0



**Fig. 10** Formation of hyperplane using SVM

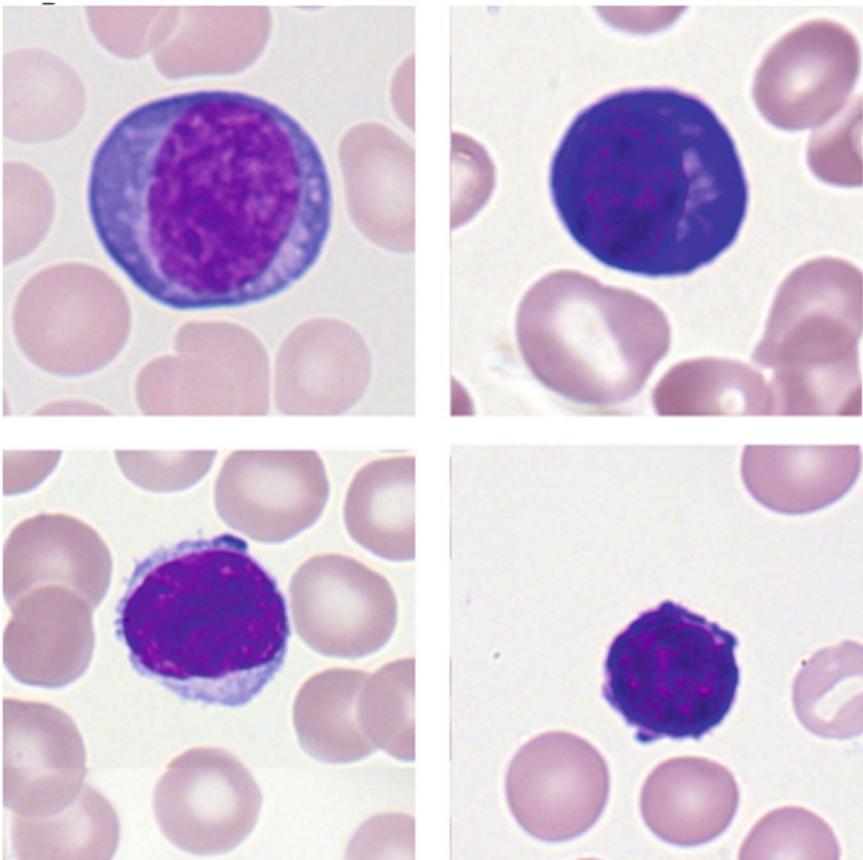
**Table 3** Comparison with other similar algorithms

Algorithm	Accuracy
Support vector machines	99.87341772151899
Linear regression	92.08745622515322
Logistic regression	92.08745622515322
Decision tree classifier	92.08745622515322
Random forest classifier	99.74683544303798

images of cancer accessible for public download. The data are organized as “collections”; typically, patients’ imaging related by a common disease (e.g., lung cancer), image modality or type (MRI, CT, digital histopathology, etc.) or research focus. DICOM is the primary file format used by TCIA for radiology imaging. Supporting data related to the images such as patient outcomes, treatment details, genomics, and expert analyses are also provided when available (Fig. 11).

#### 4.2.2 Data Processing

The image-set that we are using are not enough to begin with as they contain both cells—healthy and infected. We use the K-means clustering to extract the necessary dimensions of the infected cells, i.e., mean, area, perimeter, roundness, etc., and are further classified accordingly into mature or immature ones; it is then stored in a table (csv file) format for further processing (Table 4).



**Fig. 11** Images of affected WBCs from the TCIA

#### 4.2.3 Data Modeling

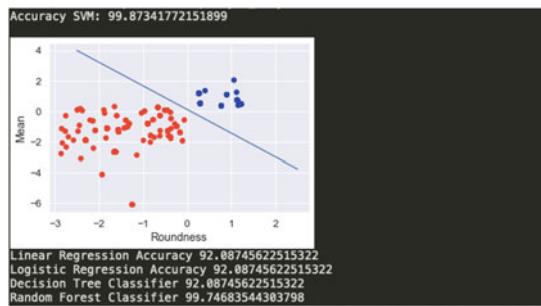
The data is then trained and tested by deploying support vector machine (SVM) algorithm. With the help of SVM, the table is taken as input where a part of it is trained, meaning the algorithm understands the trend in the data, so that the decision it takes next for another part of the data is according to what it learned. Then, it proceeds onto creating a scatterplot or hyperplane dividing the dataset into necessary classes (Fig. 12).

We extracted all single-cell images from the dataset which were produced using the M8 digital microscope/scanner from peripheral blood smears at  $100\times$  magnification and oil immersion, and a coverage of 14.14 pixels per micron is given by the manufacturer; from the Web site TCIA (The Cancer Imaging Archive).

With the help of K-means clustering, the original images are segmented, and then, features of the immature cells are extracted, i.e., mean, area, roundness, etc.

**Table 4** Feature extraction

	Mean	Area	Perimeter	Roundness	Entropy	Kurtosis	Skew	STD	VAR	Status
0	205.5202188	12997	435.6467495	0.860912067	17.22800406	2.975081768	-1.936645921	52.08100185	2712.430754	0
1	212.895625	12795	467.445739	0.736143551	17.23021883	4.568484838	-2.337812479	51.89678283	2693.276068	0
2	205.5202188	12997	435.6467495	0.860912067	17.22800406	2.975081768	-1.936645921	52.08100185	2712.430754	0
3	205.5202188	12997	435.6467495	0.860912067	17.22800406	2.975081768	-1.936645921	52.08100185	2712.430754	0
4	205.9416563	12996	469.9310203	0.739820067	17.22485077	3.261048173	-1.969678928	53.12740185	2822.520827	0

**Fig. 12** Accuracy

The extracted features from all the images are then put into a table and saved as a CSV file (Table 5).

After the creation of features extracted table, we used SVM classifier to classify the cells. The extracted table becomes the dataset for SVM algorithm; which trains and tests itself against the dataset and results to a hyperplane that separates the dataset into classes (Fig. 13).

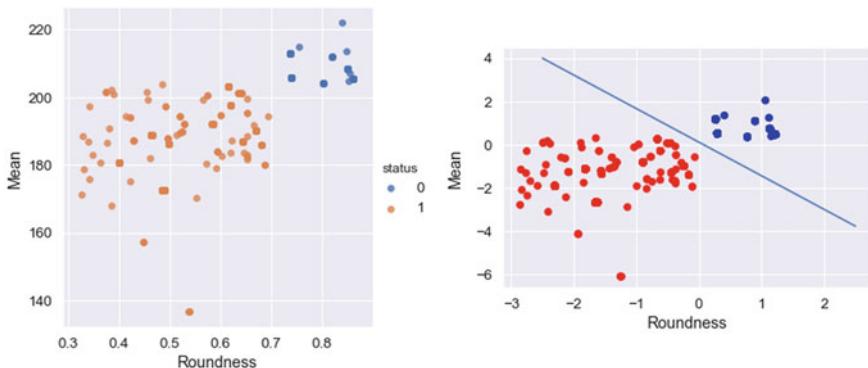
After the classification process, the model is provided with any input of a cell image for detection and prediction of presence of leukemic cell. According to its training and testing process, it predicts whether the cell image is mature or immature.

## 5 Conclusion

Cancer detection has always been a major issue, and as for leukemia's manual identification vary from expert to expert based on their proficiency and may have errors in classifying the biopsy images as healthy or leukemic ones. Getting affected by leukemia can have various sources like tobacco, diet habits, etc. Curing the disease has increased due to combined advancement in medicine and engineering, with the treatment selection depending on its malignancy. The patient's microscopic tissue structure is examined for acute detection, and the malignance level is based on amount of blast cells. Our work simplifies this process of detecting and predicting leukemic (immature WBC) cells, by going through two phases, K-means clustering and support vector machines (SVM). With the help of K-means clustering, we extract the necessary information of the affected cell from the microscopic image dataset that we have acquired from the TCIA, and then, the procured values are trained and tested by SVM which then creates a hyperplane that separates the datasets into necessary classes. After the completion of the two phases, finally, an image is given as input to the program which then further predicts whether the cell is immature or mature. With the assumption that our model is set to replace the older methods in the future, our model provides a steady base for detection and prediction of leukemia and make sound and accurate decision from those insights.

**Table 5** Feature extraction in K-means clustering

	Mean	Area	Perimeter	Roundness	Entropy	Kurtosis	Skew	STD	VAR	Status
0	205.5202188	12997	435.6467495	0.860912067	17.22800406	2.975081768	-1.936645921	52.08100185	2712.430754	0
1	212.895625	12795	467.445739	0.736143551	17.23021883	4.568484838	-2.337812479	51.89678283	2693.276068	0
2	205.5202188	12997	435.6467495	0.860912067	17.22800406	2.975081768	-1.936645921	52.08100185	2712.430754	0
3	205.5202188	12997	435.6467495	0.860912067	17.22800406	2.975081768	-1.936645921	52.08100185	2712.430754	0
4	205.9416563	12996	469.9310203	0.739820067	17.22485077	3.261048173	-1.969678928	53.12740185	2822.520827	0



**Fig. 13** Formation of hyperplane using SVM

## References

1. K.S. Kunwattana et al., Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Trans. Med. Imaging* **35**(5) (2016)
2. G. Litjens et al., A survey on deep learning in medical image analysis. *Med. Image Anal.* **42**, 60–88 (2017)
3. D.C. Cireşan et al., Mitosis detection in breast cancer histology images with deep neural networks, in *Proceedings of Medical Image Computing Computer Assisted Intervention (MICCAI)* (2013), pp. 411–418
4. M.N. Kashif et al., Handcrafted features with convolutional neural networks for detection of tumor cells in histology images, in *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*. (IEEE, April 2016), pp. 1029–1032
5. H. Chen et al., Automated mitosis detection with deep regression networks, in *Proceedings of International Symposium on Biomedical Imaging* (2016)
6. W. Xie et al., Microscopy cell counting and detection with fully convolutional regression networks. *Comput. Methods Biomed. Eng. Imaging Visual.* **6** (2018)
7. Y. Xue, N. Ray, Cell detection in microscopy images with deep convolutional neural network and compressed sensing, August 2017
8. Y. Xie et al., Efficient and robust cell detection: a structured regression approach. *Med. Image Anal.* **44**, 245–254 (2018)
9. M. Tofighi et al., Deep networks with shape priors for nucleus detection, in *2018 25th IEEE International Conference on Image Processing (ICIP)*, October 2018
10. T. Markiewicz, S. Osowski, B. Marianska, L. Moszczynski, Automatic recognition of the blood cells of myelogenous leukemia using SVM, in *Proceedings of IEEE International Joint Conference on Neural Networks*, vol. 4 (2005), pp. 2496–2501
11. A.A. Nasir, M.Y. Mashor, R. Hassan, Classification of acute Leukaemia cells using multilayer perceptron and simplified fuzzy ARTMAP neural networks. *Int. Arab J. Inform. Technol.* **10**(4) (2013)
12. S. Mohapatra, D. Patra, S. Satpathy, An ensemble classifier system for early diagnosis of acute lymphoblastic leukemia in blood microscopic images. *Neural Comput. Appl.* **24**(7–8), 1887–1904 (2014)
13. A. Hadid, The local binary pattern approach and its applications to face analysis, in *First Workshops on Image Processing Theory, Tools and Applications*, Sousse (2008), pp. 1–9
14. V. Singhal, P. Singh, Local binary pattern for automatic detection of acute lymphoblastic leukemia, in *Twentieth National Conference on Communications (NCC)* (2014), pp. 1–5

15. H.T. Madhloom, S.A. Kareem, H. Ariffin, A robust feature extraction and selection method for the recognition of lymphocytes versus acute lymphoblastic leukemia, in *International Conference on Advanced Computer Science Applications and Technologies (ACSAT)* (2012), pp. 330–335
16. S.B. Cho, Exploring features and classifiers to classify gene expression profiles of acute leukemia. *Int. J. Pattern Recognit. Artif. Intell.* **16**(07), 831–844 (2002)
17. S. Ramesh, C. Yaashuwanth, B.A. Muthukrishnan, Machine learning approach for secure communication in wireless video sensor networks against denial-of-service attacks. *Int. J. Commun. Syst.* **33**(12) (2019). <https://doi.org/10.1002/dac.4073>
18. B. Liu, V. Pham, N. Nguyen, A virtual backbone construction heuristic for maximizing the lifetime of dual-radio wireless sensor networks, in *2015 International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)*, Adelaide, SA, Australia (2015), pp. 64–67. <https://doi.org/10.1109/IIH-MSP.2015.20>
19. W. Rong, Z. Li, W. Zhang, L. Sun, An improved CANNY edge detection algorithm, in *2014 IEEE International Conference on Mechatronics and Automation (ICMA)*. (IEEE, 2014), pp. 577–582
20. L. Xuan, Z. Hong, An improved canny edge detection algorithm, in *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*. (IEEE, 2017), pp. 275–278
21. Q. Xu, S. Varadarajan, C. Chakrabarti, L.J. Karam, A distributed canny edge detector: algorithm and FPGA implementation. *IEEE Trans. Image Process.* **23**(7), 2944–2960 (2014)

# Secured Cloud Storage System Using Auto Passkey Generation



Sudha Ayatti, K. Gouri, Pavan Kunchur, and Sadhana P. Bangarashetti

**Abstract** The cloud storage system is an improved online drive storage system with storage on cloud. It can handle all details about a file uploads with categories, preceded by file download whenever required. The file gets uploaded with some encryption strategy and stored safely on the cloud, later on while downloading the file, the decryption is done using a security key provided through the registered mail ID. The details include file name, type of file (Private/Public) and auto generated passkey. In case of system, they need a lot of time for searching the file which is not a secured way. So the accuracy will be maintained in this application. This system is managed by the application in cloud server. It is the work of the cloud application which is headed in place to observe the complete procedure.

**Keywords** Encryption · Passkey · Cloud storage system

## 1 Introduction

### 1.1 Background

Cloud computing is a demand release that computes control, database storage, applications with other IT assets during the cloud service stage using the Internet by pay-as-you-go pricing option. It has various services among which data storage is the main cloud service. This means that the users can keep the records on cloud in its place of storing the limited organization and the contact of that stored information, during the system/Internet connection with client service. Distributed computing

---

S. Ayatti · K. Gouri · P. Kunchur (✉)

Department of Computer Science and Engineering, KLS, Gogte Institute of Technology, Belagavi, India

e-mail: [pnkunchur@git.edu](mailto:pnkunchur@git.edu)

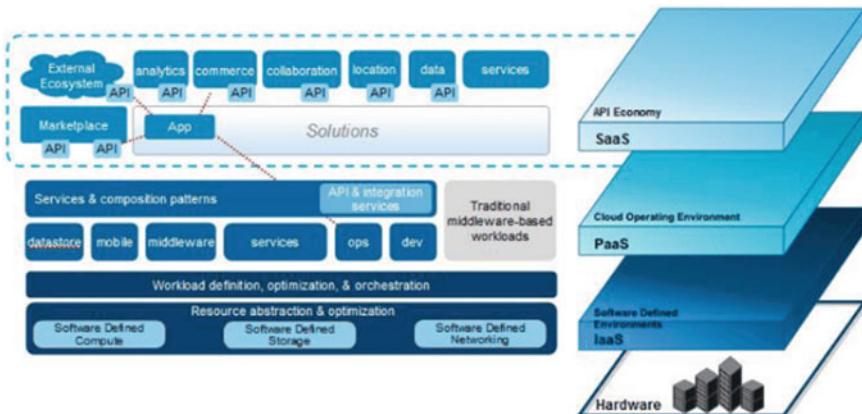
S. P. Bangarashetti

Department of Information Science and Engineering, Baweshwar Engineering College, Bagalkot, India

emerges from the blend of the customary PC innovation and organization innovation, for example, lattice registering, circulated figuring, equal processing, utility registering and virtualization. One of the centre ideas of distributed computing is to diminish the trouble on client's terminals, thereby constantly improving the mists and dealing with the limit. Distributed computing is generally a word intended for anything to facilitate IT administration more a web. The cloud administration is carefully inaccessible and is interested in three lessons, specifically in the information as a service (IaaS), platform as a service (PaaS) and software as a service (SaaS).

## 1.2 Cloud Architecture

The cloud computing architecture is distinct with layer. About three major layers of a cloud computing architecture are as follows: the information as a service (IaaS), platform as a service (PaaS) and software as a service (SaaS). As a total, cloud architecture be meant to provide a user with a large bandwidth, by allowing the users to access the information or data using the application without any interruptions, with on demand agile network to make it possible to move required data, rapidly and economically among servers, and in this process the number of servers and security of data will play an important role in network security. The IaaS provides the infrastructure and hardware running on the cloud, which is available to the users on pay per usage basis as networks, servers, storages, devices, etc. The PaaS gives the users by requesting platform with databases, which is equivalent towards the middleware forces. The SaaS helps in providing the hosting and maintenance work on Internet. SaaS helps the users not to install any software locally which is exposed in Fig. 1.



**Fig. 1** Cloud architecture

### **1.3 Cloud Storage**

The cloud storage is the cloud computing representation which provides storage of a data during the Internet via the cloud computing supplier, and they manage and operate data storage service which means that the users can keep the records on a cloud in its place storing in the neighbourhood structure with contact of that stored records are through the system/Internet connectivity with the customer service. Most of the cloud storage systems are designed as distributed, redundant systems with complicated architecture, but you cannot think of a cloud storage system as just another hard drive. Cloud storage is the key infrastructure to achieve service interaction experience and seamless information sharing across multiple users.

There are many other benefits of cloud storage, which are listed as follows:

- The companies can simply give support of a storeroom they utilize, which makes functioning payment slightly more than principle costs.
- Practically, these are immeasurable storage space capacity. But if a user includes an essential of the more storeroom, it can decrease the cost.
- Clouds vendors give hardware severance with regular storage space field, which helps towards evade services outrage during computer breakdown.

Even after providing reliable services, the cloud storage has disadvantages too. Few disadvantages are listed below:

- Price and reliability—The customers have headed for computed cost effectiveness which is cloud solution next to host to maintain the information.
- Security—Sometimes, there is a possibility that the data might get stolen or is viewed by unauthorized users. The risk increases if the customer does not have fully control over it.

## **2 Problem Statement**

To design and develop a cloud storage system that provides user to upload and download using authorized access, the documents are encrypted during uploading and decrypted during downloading, where the documents are stored categorically.

## **3 Objectives**

It is the network orient function which allows to contact on few kinds of files etc. This request gives practical visit to systematic file storage system. Here we will get the latest information about the files and cloud server activities. This generic application can be used by private organizations, institutions, software companies, etc. It also provides support to store various kinds of files like documents, images, different kind

of videos, and other files. As this a completely cloud-based application, here it will be no administration process or an administrator to handle the process of registration. Thus, file storage is done in such a manner that it can be accessed fast as well as on any mobile device.

### ***3.1 Existing System***

The data in this system is stored in an uncategorized manner, especially after uploading the file, and the files are stored in an online drive without categorization. As the different types of files are stored in the same location, it becomes more overhead for computing process which gives the result that is not time efficient. It becomes more difficult to search files after a good number of months when the user wants to get the file to know some of the details of files or to download. It also takes time to search on the pile of files. Also there are some unknown errors which the system cannot answer which have to be checked manually like if asked to find the file with just an extension in storage location. Users will have to search manually and write down which actually takes a lot of time.

### ***3.2 Proposed System***

The advancement of an original support includes accompanying movements that attempts to mechanize a whole round maintenance considering the advancement in the cloud server. The file uploading, categorized storage and downloading are intended method to provide a cloud structure which is much faster and easier. The files will be encrypted and saved in the storage location and need a passkey to decrypt and download the file.

Some of the advantages of the proposed application are as follows:

- User neighbour lines are furnished in the application with dissimilar reins.
- Structures make a general stockpiling to enter the set simpler with flexibility.
- This can be got towards more as an Internet.
- Different lessons have been utilized towards furnishing record transport with download and post things to see.
- Around the refusal peril, all together mishandle on some point as an endeavour improvement is beneath series.

## **4 Methodology**

The universe of register around endures well-being issue on the side of taking care of data inside cloud. Towards ensure about data inside cloud AES encryption system be

used inside task. Advanced encryption rule is the square code by the square degree 128 pieces. That awards three distinctive info lengths: 256, 192 and 128 or, more than likely pieces. Cryptography is one among the most recognizable and needed techniques to protect the data from the aggressors by using the two principal cycles. These cycles are recorded on the grounds encryption with decryption. Encryption is a way towards changing a data over the protected or ending the attackers to examine the principal data clearly. Encryption incorporates the difference in plain substance to incoherent association. It is known as code text. The customer cannot scrutinize the above association. Consequently, the accompanying cycle that is finished by the customer is decryption. AES is symmetric key code. This infers a comparative secret key used for both encryption and unscrambling, and both the shipper and recipient of the data need a copy of the key. Then again, deviated key systems use a substitute key for all of the two strategies. Deviated keys are best for outside report moves, while symmetric keys are more able to do internal encryption. The advantage of symmetric structures like AES is their speed. Since a symmetric key estimation requires less computational power than an uneven one, it is speedier and continuously viable to run.

## 5 Literature Review

Considering in the time of 2011, creators are familiar with an article related to course handling associations like systems, limit, labourers, associations and applications without really receipt them. As a perception, the bigger frameworks have decreased the danger of information spillage. Fundamentally of distributed computing, it requests web administrations. Information is developing quickly; however, security and protection of an open and rather fluidly available assets are as yet a contention which managed the security from distributed computing climate, cloud conveyance model, partners and attributes too. Regardless, circulated figuring structures and organizations have become critical concentrations for digital assailants. Since the cloud system is with a particular goal in mind as an open and shared stage, it is reliant upon toxic attacks from the two insiders and untouchables.

### Mainly focus on security of cloud computing:

- **User's authentication:** User authentication process must be optimized with clear verification to get rid of malicious users from accessing powerful cloud system.
- **Leakage of data:** Data would be under risk if unwanted users gain access and delete or modify data, problems will be arisen if there is no backup.
- **Hijacking of session:** This happens when legitimate client is the level towards troubled solicitation interface which holder endure discouraged with aggressor.

The ‘Cryptographic Public Verification of Data Integrity for Cloud Storage Systems’, Yuan Zhang, Chunxiang Xu, Hongwei Li, with Xiaohui Liang moreover deals with a security as disseminated extra room benefits yet since another point:



**Fig. 2** Cloud data storage architecture's

affirmation is data genuineness. Various open check plans use an outcast monitor to affirm the reliability of data reallocated to convey capacity organizations, but they are feasible just assuming a strong doubt holds: that the inspectors are clear and trustworthy. To experience promising security results, we ran over utilizing encryption technique AES-256 calculation, and the advanced encryption standard (AES) estimation is one of the square figure encryption computations to be disseminated close to the National Institute of Standards with development (NIST) in 2000. The essential marks of this estimation were to override DES calculation (Fig. 2).

Programming plans towards a significant construction of produced structure with an order of designer's creation with systems. Each association includes preparing system, family members encompassed by the, with property of two parts with family members. Advanced encryption standard methodology acts in consent to both hardware and programming stages under a wide scope of conditions. It fuses 64-cycle and 8-digit stages. Its basic parallelism works with useful use of processor resources achieving commonly great programming execution. The computation has various ideal conditions like less memory conveyances on the side of execution and sensible for restricted opportunity conditions. Plans of computation have fine reach on the side of benefitting as of direction point parallelism.

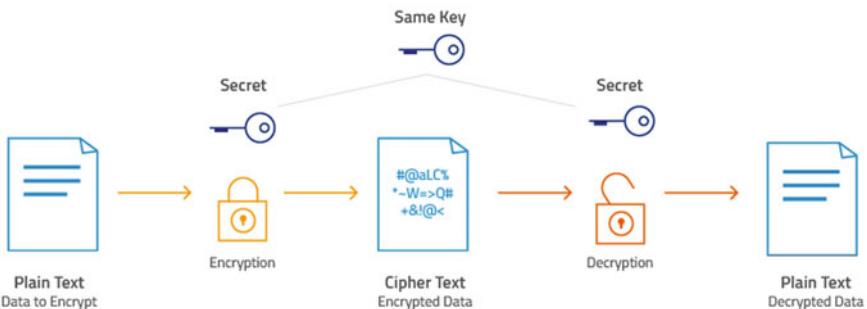
## 6 Implementation of AES Algorithm

Programming configuration secure to the chief designs of an item structure and in solicitation to ensure about data inside cloud AES encryption strategy can be used inside the endeavour. Advanced encryption model is square shape code by the square degree 128 pieces. That awards three particular information lengths: 256, 192, 128 or bits (Fig. 3).

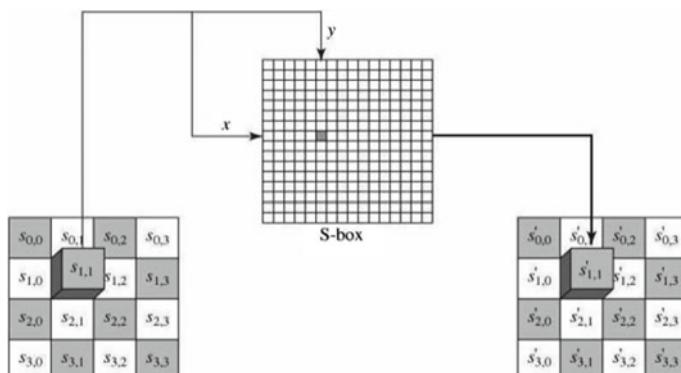
In AES, the organization  $4 \times 4$  including the 128 bytes enter block is perceived since a condition bunch. A cycle as encryption turns around four stages explicitly mix, segments, sub bytes, embed round key with move lines.

Sub-bytes are described since substitution activity. That is nonlinear. Each byte is restored through an extra according to S-box. This movement gives an underhanded degree in code. The resultant grid includes four segments and four lines (Fig. 4).

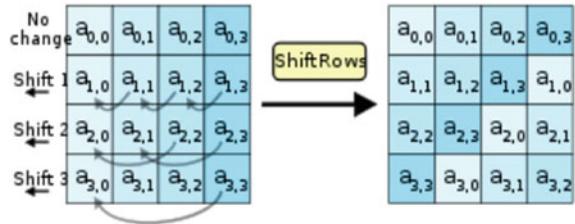
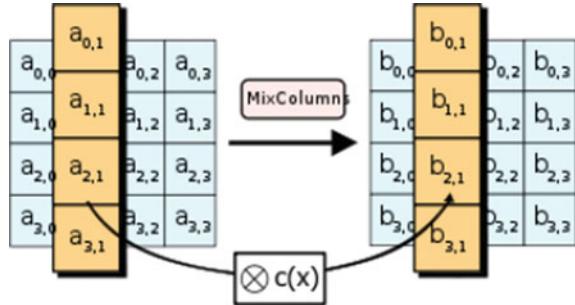
**Shift Rows**—This is a point wherever every segment is turned slowly into an undeniable number of occasions. It is generally called change. The four sections in the



**Fig. 3** Encryption and decryption mechanism



**Fig. 4** Byte substitution

**Fig. 5** Shift rows**Fig. 6** Mix columns

grid are turned fittingly. The sections are moved aside. Move is finished as Row 1 is not turned. Row 2 is moved one-byte spot aside. Line three is moved two spots towards one side. Column four is moved three spots close aside. This resultant system involves the 16 bytes anyway turned with respect to each other (Fig. 5).

**Mix Columns**—In this movement, each portion is changed using lattice duplication. Each segment contains four bytes. The significant structure involves 16 bytes. This data is occupied with the help of each segment. These take four bytes. This yield produces four bytes that are through and through not exactly an equivalent as the four bytes concurred as data mix columns (Fig. 6).

**Add Round Key**—This, with regard to enter, is restricted towards each nibble and is state. Inside the particular development, an organization is XORed by an encompassing key. The four  $\times$  four lattices address the first info. These contain 128 bits. It has four style inputs anywhere placed such that four bytes are changed over the 43 composing keys. An underlying 4.

Expressions speak to W [0], W [1], W [2] and W [3].

## 7 Results and Discussions

### Snapshots:

#### 1. The Homepage

This is the main page of our project, which has further guided the user with two options:

- **Register**, if they are visiting the site for the first time or have not yet created their account.
- **Login**, for the regular users of the site who have registered previously.

## 2. Registration Page

This page allows the users to register them to the site, where the user needs to fill in all the details required for the registration.

## 3. The Login Page

This is the login page, where the users login themselves to the site. The login requires the registered email ID and password.

## 4. File Upload

The users can upload the files here. The ‘File Upload’ is further provided by dropdown options as

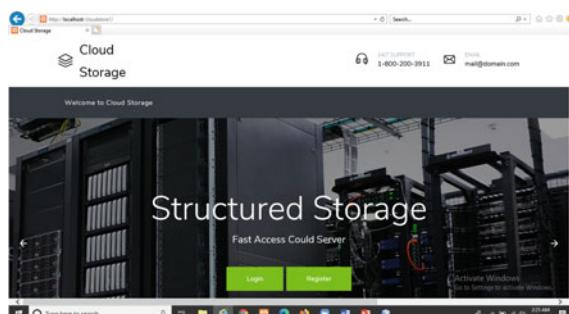
- Private File**, where the files can be accessed only by the user individually.  
**Public File**, where all the registered users of the system can view and download the file.

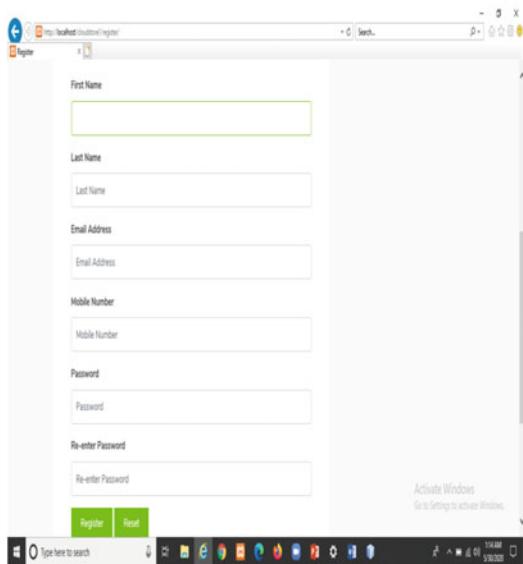
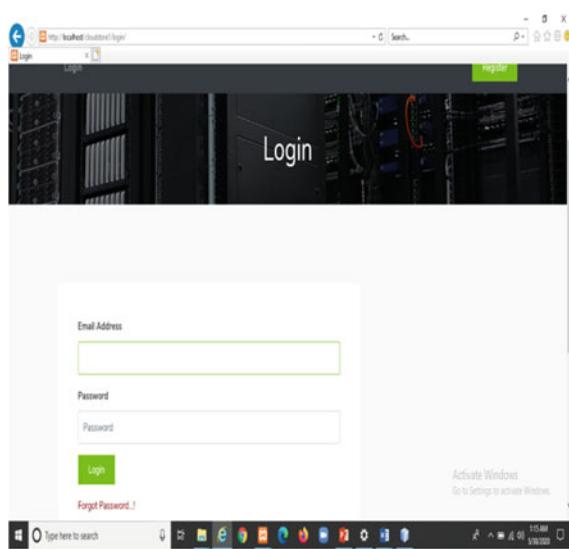
## 5. File Download

The ‘File Download’ has similar dropdown options as that of ‘File Upload’. The users can download the files here.

The user has the click on the ‘Request’, on the corresponding row of the file name, in the ‘Get Key’ column. Further he/she has to click on the ‘Download’, in the ‘Decrypt and Download file’ column and enter the secret key. The secret key is mailed to the users registered email ID, once he requests for the key (Figs. 7, 8, 9, 10, 11 and 12).

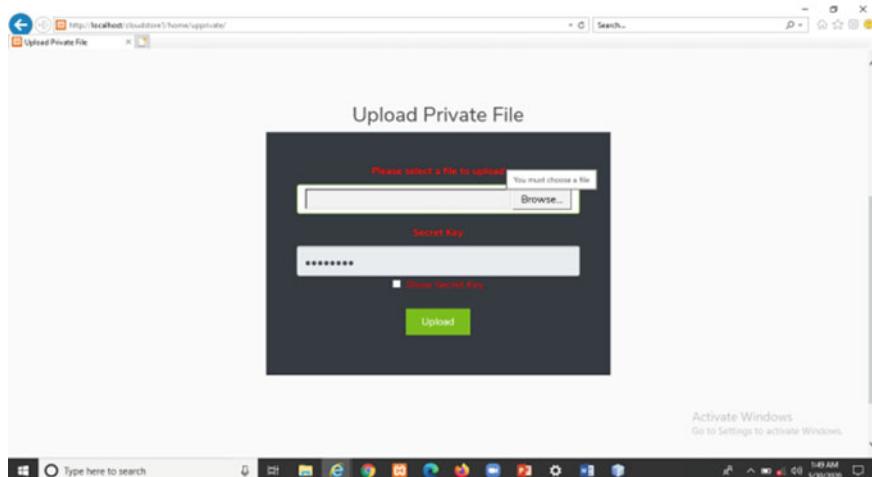
**Fig. 7** Homepage



**Fig. 8** Registration page**Fig. 9** Login page

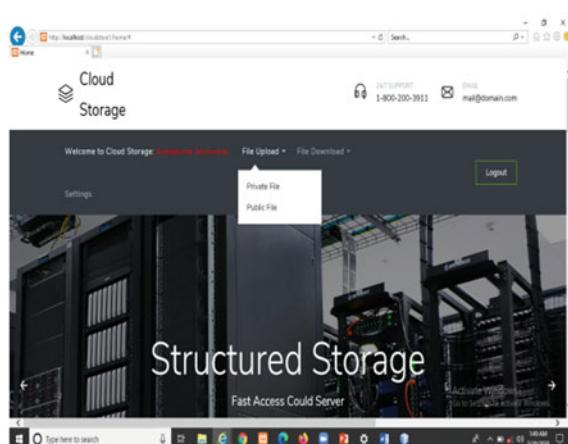
## 8 Conclusion

Recently, it circulated that figuring is an establishment of a quick new development, in any case, the security issues have become obstructions to make the appropriated processing more standard which ought to be tended to. Data open into the cloud

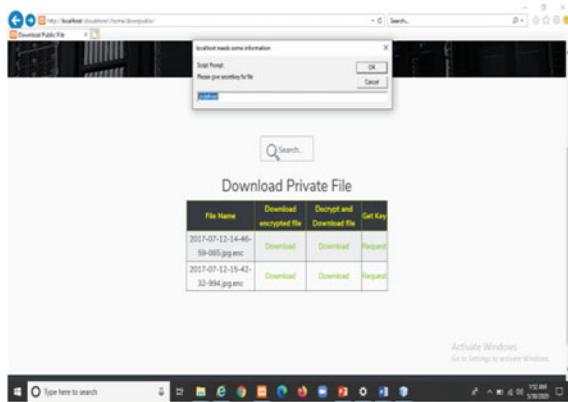


**Fig. 10** Update private file

**Fig. 11** File upload



can be at serious risk if not guaranteed genuinely. This paper examines security and dangers to information in the cloud. Individually, the fundamental concern is that task is data security with its terrorizing notwithstanding arrangements in distributed computing. Data inside an assortment of states have been analysed and closed by the strategies which are useful for scrambling the data which are effective in encoding the information in the cloud. Usefulness over utilizing AES-256 likewise enthusiastic is the way in which information is put away in an organized manner, to work on the proficiency and access information in an enhanced way. Dispersed figuring well-being is not just the particular subject, that moreover incorporate standardization, regulating structure, laws and rules, with a few disparate points, appropriated

**Fig. 12** File download

processing be associated with improvement opportunity with issues, along an assurance theme will be there which is addressed bit close to digit, conveyed to grow, an accommodation which will resolve likewise towards genuinely increasingly more normally.

## References

1. R. Buyya, C. Yeo, S. Venugopal, J. Broberg, I. Brandic, Cloud computing and emerging IT platforms: vision hype and reality for delivering computing as the 5th utility. *J. Future Gener. Comput. Sci.* **25**(6), 599–616 (2009)
2. P. Mell, T. Grance, *The NIST Definition of Cloud Computing*. (NIST Special Publication, 2011), pp. 800–145
3. M. Ali, S. Khan, A. Vasilakos, Security in cloud computing: opportunities and challenges. *Inf. Sci.* **305**, 357–383 (2015)
4. C. Wang, Q. Wang, K. Ren, N. Cao, W. Lou, Towards secure and dependable storage services in cloud computing. *IEEE Trans. Cloud Comput. Date Publ.* **5**(2) (2015)
5. J. Wu, L. Ping, X. Ge, Y. Wang, J. Fu, Cloud storage as the infrastructure of cloud computing, in *2010 International Conference on Intelligent Computing and CognitiveInformatics* (2010)
6. [https://www.researchgate.net/publication/305508410\\_Cloud\\_Storage\\_Advantages\\_Disadvantages\\_and\\_Enterprise\\_Solutions\\_for\\_Business](https://www.researchgate.net/publication/305508410_Cloud_Storage_Advantages_Disadvantages_and_Enterprise_Solutions_for_Business)
7. G. Suciu, S. Halunga, A. Apostu, A. Vulpe, G. Todoran, Cloud computing as evolution of distributed computing—a case study for SlapOS distributed cloud computing platform. *Informatica Economică* **17**(4), 109–122 (2013)
8. M.A. Sharkh, M. Jammal, A. Shami, A. Ouda, Resource allocation in a network-based cloud computing environment: design challenges. *IEEE Commun. Mag.* **51**(11), 46–52 (2013)
9. G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z. Paterson, D. Song, Provable data possession at untrusted stores, in *Proceedings of the ACM Conference on Computer and Communications Security (CCS'07)*, 29 October–2 November 2007, pp. 598–610
10. C. Erway, A. Kupcu, C. Papamanthou, R. Tamassia, Dynamic provable data possession, in *Proceedings of the 16th ACM conference on Computer and communications security (CCS)* (2009), pp. 213–222
11. C. Wang, Q. Wang, K. Ren, W. Lou, Privacy preserving public auditing for secure cloud storage. *IEEE Trans. Comput.* **62**(2), 362–375 (2011)

12. A. Juels, J. Burton, S. Kaliski, Proofs of retrievability for large files, in *Proceedings of the 14th ACM Conference on Computer and Communications Security (CCS)* (2007), pp. 584–597
13. M.A. Shah, M. Baker, J.C. Mogul, R. Swaminathan, Auditing to keep online storage services honest, in *Proceedings of the 11th workshop on hot topics in operating systems (HotOS'07)*, ‘HotOS’, USENIX Association (2007), pp. 1–6
14. Q. Wang, C. Wang, J. Li, K. Renand, W. Lou, Enabling public verifiability and data dynamics for storage security in cloud computing, in *Proceedings of 14th European Symposium Research in Computer Security (ESORICS '09)* (2009), pp. 355–370
15. P. Prasadreddy, T. Srinivasa, S. Phani, A threat free architecture for privacy assurance in cloud computing, in *Proceedings of the IEEE World Congress on Services*, 4–9 July 2011, USA. (IEEE Xplore Press), pp. 564–568. <https://doi.org/10.1109/SERVICES.2011.11>
16. D. M’Raihi, S. Machani, M. Pei, J. Rydell, TOTP: time-based one-time password algorithm, Request for Comments (RFC) 6238, 13 July 2011
17. K. Kiran, K. Padmaj, P. Radha, Automatic protocol blocker for privacy-preserving public
18. S.A. El-Booz, G. Attiya, N. El-Fishawy, A secure cloud storage system combining time-based one time password and automatic blocker protocol. Faculty of Electronic Engineering, Menoufia University, Menouf, Egypt. *IEEE Trans.* (2015). <https://doi.org/10.1109/ICENCO>
19. G. Suciu, S. Halunga, A. Apostu, A. Vulpe, G. Todoran, Cloud computing as evolution of distributed computing—a case study for SlapOS distributed cloud computing platform. *Inform. Econ.* **17**(4), 109–122 (2013)
20. P. Mell, T. Grance, The NIST definition of cloud computing. National Institute of Standards and Technology, Information Technology Laboratory, 7 October 2009. <http://www.nist.gov/itl/cloud/>
21. M.A. Sharkh, M. Jammal, A. Shami, A. Ouda, Resource allocation in a network based cloud computing environment: design challenges. *IEEE Commun. Mag.* **51**(11), 46–52 (2013)
22. c. Wang, Q. Wang, K. Ren, and W. Lou, “Privacy-Preserving Public Auditing for Secure Cloud Storage.” *IEEE Transactions on Computers*, Vol. 62, No. 2, pp. 1–12, 2013.
23. M. Venkatesh, M.R. Sumalatha, C. Selva Kumar, Improving public auditability, data possession in data storage security for cloud computing, in *Proceedings of the International Conference on Recent Trends in Information Technology (ICRTIT)*, 19–21 April 2012, pp. 463–467
24. S. Bhagyashri, Y.B. Gurave, A survey on privacy preserving techniques for secure cloud storage. *Int. J. Comput. Sci. Mob. Comput. (IJCSMC)* **3**(2), 675–680 (2014)
25. G. Caronni, M. Waldvogel, Establishing trust in distributed storage providers, in *Third IEEE P2P Conference*, Linkoping 03, 2003
26. P. Golle, S. Jarecki, I. Mironov, Cryptographic primitives enforcing communication and storage complexity, in *Proceedings of Financial Crypto 2002*, Southampton, Bermuda
27. F. Sebe, J. Domingo-Ferrer, A. Martinez-Balleste, Y. Deswart, J.-J. Quisquater, Efficient remote data possession checking in critical information infrastructures. *IEEE Trans. Knowl. Data Eng.* **20**(8), 1034–1038 (2008)
28. G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z. Peterson, D. Song, Remote data checking using provable data possession. *ACM Trans. Inform. Syst. Sec.* **14**(1), 12.1–12.34 (2011). Article 12
29. A. Juels, J. Burton S. Kaliski, PORs: proofs of retrievability for large files, in *Proceedings of CCS '07*, Alexandria, Va, USA (2007), pp. 584–597
30. G. Ateniese, S. Kamara, J. Katz, Proofs of storage from homomorphic identification protocols, in *Proceedings of ASIACRYPT '09*, Tokyo, Japan (2009), pp. 319–333

# A Novel Multi-Objective Memetic Algorithm for Mining Classifiers



K. R. Ananthapadmanaban, S. Muruganandam, and Sujatha Srinivasan

**Abstract** Genetic algorithm is a nature-inspired evolutionary technique which uses natural evolution methods to explore large solution spaces to find better solutions. Memetic algorithm is a special kind of genetic algorithm that uses a local search technique to exploit the solution space. In the current study, the nature-inspired technique of memetic algorithm is combined with optimization techniques to create a hybrid multi-objective memetic algorithm (MOMA). In the current study, the genetic algorithm is extended into a memetic algorithm by adding a memory base to store the intermediate knowledge and exploit them in the future generations. The algorithm is applied to rule mining as a problem with multiple objectives. The results of testing the algorithm on different data sets is encouraging and throws light in the direction of adding memory to the existing evolutionary computing to get better classifiers.

**Keywords** Genetic algorithm · Memetic algorithm · Multi-objective optimization · Association rules · Data analytics · Data mining

## 1 Introduction

Evolutionary computing techniques which have been applied to data mining are found abundant in the literature. However, there is a still lot more to explore and exploit in this area inspired from nature. “If–Then” type rules that form a classifier to classify unknown future instances are most sought out since they are more user friendly and

---

K. R. Ananthapadmanaban

Faculty of Science and Humanities, SRM Institute of Science and Technology, Vadapalani Campus, Chennai, India

e-mail: [kranantp@srmist.edu.in](mailto:kranantp@srmist.edu.in)

S. Muruganandam · S. Srinivasan (✉)

Department of Computer Science, SRM Institute for Training and Development, Chennai, India

S. Muruganandam

e-mail: [ssmanand@yahoo.co.in](mailto:ssmanand@yahoo.co.in)

understandable form of knowledge. Mining rules with desired properties are a multi-objective optimization problem. Combining evolutionary computing techniques with optimization techniques would be a good solution in the rule mining arena. In the current study, a memetic algorithm which uses optimization techniques to select individuals is proposed. A meme is a piece of information obtained from previous generations of evolution and which is passed on to future generations. The proposed algorithm uses this evolutionary knowledge by incorporating a memory element to the GA. Thus, extended MOMA or E-MOMA is proposed for rule mining. The rest of the paper is organized as follows. Section 2 brings out a review on existing algorithms for rule mining. Section 3 explains the methodology of the proposed or E-MOMA. Section 4 explains the experiments and summarizes the results and adds a discussion of the results obtained. Comparative performance of the proposed MA with the existing algorithms is discussed in terms of accuracy and complexity of the classifier. Section 5 concludes with a list of future research directions.

## 2 Literature Survey on Multi-objective Evolutionary Computing Techniques for Rule Mining

Multi-objective evolutionary techniques have been extensively used in the literature for solving a variety of problems. The following survey brings out the features of multi-objective genetic algorithms that have been proposed for rule mining found in the literature. Three popular multi-objective metaheuristics have been applied in medicine for survival prediction in burnt patients in [1]. A mono-objective GA is applied to the variable selection problem for classification of biodiesel samples to detect adulteration in [2]. A genetic fuzzy system is introduced for threat detection which uses pair-wise learning to improve detection accuracy by [3] and has been applied to the benchmark KDDCUP99 dataset. The authors of [4] discuss a novel genetic algorithm that induces compact chain classifiers that assign multiple class labels to an individual object. While in [5], the authors have introduced a new method that uses distance from average solution as the evaluation criteria for inventory classification. However, the problem of dealing with imbalanced data set for classification rule mining is considered in [6]. Swarm-based clustering, oversampling technique, and multi-objective algorithm for tackling imbalanced data sets are proposed by [7], while Antonelli et al. [8] have proposed a multi-objective evolutionary design of granular rule-based classifier. In the work by [9], email messages are classified as suspicious or doubtful or borderline, and thus, spam classification is considered as a three way classification problem. A multi-objective GA combined with neural network for prediction of water quality is proposed in [10]. While a multi-objective genetic algorithm has been used to train a neural network-based model to classify structural design and predict structural failure, the authors of [11–13] propose a surrogate assisted multi-objective Evolutionary algorithm for many objective problems in optimization. A multi-objective memetic algorithm that combines

different single objective ones using decomposition and local guided search has been proposed by [14]. A memetic algorithm-based support vector machine (SVM) using emperor penguin optimization and social engineering optimization techniques has been proposed in [15] for feature set selection and optimization of SVM parameters. A multi-objective evolutionary algorithm is proposed in [16] for preprocessing data for selection of training set to feed to a support vector machine, while the authors of [17, 18] have proposed a novel memetic algorithm for discovering communities from signed network.

### 3 The Extended Multi-objective Memetic Algorithm

#### 3.1 *The Problem*

In the current study, rule mining is taken as a multi-objective optimization problem. The usual GA is extended to store not only elite but also less fit individuals to add an extended memory base to be exploited by the algorithm. The problem here is to mine classification rules from the given data sets in an evolutionary environment and studying the influence of the proposed extended MOGA or E-MOMA on its performance in providing the users with accurate and compact classifier.

#### 3.2 *Methodology*

The pseudocode of the proposed MOMA is given in Table 1, and the representation of the solution space, fitness evaluation, population creation are explained below.

#### 3.3 *Representation of Genotypes, Phenotypes, and the Fitness Vector*

In the current study, it was decided to represent the genes as discrete numbers for uniformity and reduce complexity. Discrete values are stored as such and real values are discretized using equal width binning. Phenotypes are the representation of the individuals in the solution space. Solution space consists of the best vectors along with their fitness vector. The fitness vector consists the metric values associated with a rule.

**Table 1** Pseudocode of the algorithm E-MOMA

---

**Algorithm: E-MOMA**


---

**Input:** Train Data; Evolutionary Parameters; Optimization Metrics;
**Output:** Classifier

---

Create Initial Population  
 Evaluate individuals  
 Select individuals for next generation  
 Store best metric values  
 Do  
     Generate probability  
     Select individuals for reproduction  
     Perform Crossover or Mutation  
     Evaluate fitness of children  
     Compare fitness vectors of children to choose better  
         individuals  
     Store best children in Elite List and  
         non-elite in separate population  
 While not termination condition  
 Output Classifier

---

### **3.4 Population Creation, Fitness Evaluation, and Optimization**

The initial population is created as follows. Two data instances from the training data are chosen at random. These are the initial parents that undergo the reproduction operators of crossover and/or mutation to create children. This procedure terminates when the required population size is reached. The individuals in the population are evaluated based on the objective metrics. The metrics chosen in the current study are support and confidence. These two metrics are calculated for each individual and stored as vectors. The fitness of the children is evaluated and compared using Pareto optimization for selection of candidates for next generation.

## **4 Experiments, Results, and Discussion**

### **4.1 Experiments**

Experiments were designed to test the algorithm in creating a compact set of rules so as to reduce complexity and maximize accuracy. Thus, the ultimate objective of the study was to minimize complexity and maximize accuracy. With this aim, seven benchmark data sets from the UCI machine learning repository were chosen and the algorithm tested on it. The data sets include iris data set that classifies flowers, Ljubljana breast cancer data set (Ljb), Wisconsin breast cancer data set (WSBC), liver disorder data set (Bupa), German and Australian credit card approval data, (Crx and Aus), and mammography data set represented by mamo.

**Table 2** Summary of the performance of the algorithm on different data sets

S. No.	Data set	Time taken (s)	Accuracy (%)	Average No. of rules in the classifier
1	Iris	2.85	96.80	6.07
2	Ljb	13.10	84.95	7.00
3	Bupa	3.12	96.00	13.7
4	Crx	4.05	73.50	13
5	Aus	11.31	92.38	21
6	WSBC	15.97	93.86	14
7	Mamo	13.96	87.73	6.9

## 4.2 Results

Table 2 summarizes the performance of the algorithm on different data sets, Table 3 gives the comparative performance of the proposed extended MOMA with a simple MOMA without memory, and Table 4 gives the comparative performance of the proposed algorithm with that of other algorithms found in the literature.

## 4.3 Discussion

Table 2 gives the overall performance of the proposed E-MOMA in terms of accuracy, number of rules presented to the user, and the time taken to mine rules. From the table, it can be observed that there is slight performance degradation in terms of accuracy as well as time taken when dealing with imbalanced data sets. Use of preprocessing techniques like fuzzy discretization and use of sampling techniques to balance data instances can be used to deal with this problem as justified later in the paper.

Table 3 gives a comparative summary of a simple MOGA and the proposed E-MOMA with respect to accuracy, number of rules which defines the complexity criteria of the classifier, and the time taken to mine and present the rules to the user. It can be observed from the table that combining an optimization strategy with evolutionary computing techniques improves the performance of the classification algorithm in all respects including maximization of accuracy, minimizing complexity of the classifier in terms of number of rules, and also the time taken to mine the rules. This establishes the fact that adding memetic data and hybridization of techniques from diverse areas can be exploited better toward improving the performance of simple genetic algorithms.

Table 4 gives the comparative performance of the proposed E-MOMA with other multi-objective algorithms found in the literature. The proposed E-MOMA outperforms most of the other algorithms in most of the data sets. However, the algorithm proposed by [16] which uses fuzzy discretization approach to preprocess the data set outperforms the proposed E-MOMA in two (credit card approval and Wisconsin

**Table 3** Comparative performance of the proposed E-MOMA with simple MOGA on different data sets in terms of accuracy, number of rules in the classifier, and time taken by the algorithm

Data set	Accuracy (%)		No. of rules		Time taken in seconds	
	MOGA	E-MOMA	MOGA	E-MOMA	MOGA	E-MOMA
Bupa	73.48 ± 1.1	96.00 ± 1.18	151 ± 7.42	13.7 ± 1.38	10.113 ± 2.19	3.117 ± 1.75
Iris	90.40 ± 4.56	96.80 ± 1.10	109 ± 7.08	6.07 ± 3.42	11.28 ± 5.70	2.85 ± 1.87
Aus	76.52 ± 3.81	92.38 ± 2.44	387 ± 16.28	21 ± 5.68	21.68 ± 4.51	11.31 ± 3.44
Crx	63.00 ± 5.48	73.50 ± 3.09	164 ± 9.53	1.3 ± 4.31	29.18 ± 7.75	4.05 ± 3.09
Mamo	74.25 ± 8.01	87.73 ± 1.62	142 ± 11.37	6.9 ± 4.48	27.02 ± 5.03	13.96 ± 2.58
Ljb	78.49 ± 3.04	84.95 ± 1.39	136.2 ± 3.52	7 ± 0.96	20.21 ± 4.32	13.10 ± 3.35
WSBC	75.28 ± 5.64	93.86 ± 1.31	335 ± 11.00	14 ± 3.37	25.51 ± 6.27	15.97 ± 2.58

**Table 4** Comparative performance of the proposed E-MOMA with other techniques on different data sets in terms of accuracy

S. No.	Data set	Accuracy	
		Multi-objective techniques	E-MOMA
1	Bupa	PAES-R: 67.65, PAES-RT: 64.84 PAES-RG: 68.48, PAES-RGT: 64.24 [8]	96.00 ± 1.18
		D-MOFARC: 70.1, FARC-HD: 66.4 [16]	
2	Iris	PAES-R: 94.85, PAES-RT: 93.64 PAES-RG: 95.36, PAES-RGT: 95.07 [8]	96.80 ± 1.10
		D-MOFARC: 96, FARC-HD: 95.3 [16]	
3	Australian credit	PAES-R: 84.66, PAES-RT: 84.11 PAES-RG: 85.27, PAES-RGT: 85.85 [8]	92.38 ± 2.44
		D-MOFARC: 86, FARC-HD: 86.4 [16]	
4	Credit card approval	D-MOFARC: 84.7, FARC-HD: 84.9, [16]	73.50 ± 3.09
5	Mammography	95.6 ± 0.02 [7]	87.73 ± 1.62
		PAES-R: 82.47, PAES-RT: 82.37 PAES-RG: 82.96, PAES-RGT: 82.59 [8]	
6	Wisconsin breast cancer	PAES-R: 95.98, PAES-RT: 95.78 PAES-RG: 96.31, PAES-RGT: 96.51[8]	93.86 ± 1.31
		D-MOFARC: 96.8, FARC-HD: 96.2 [16]	

breast cancer data sets) out of the six data sets considered for comparison. This encourages further study in using fuzzy logic for discretization as future research. Again using sampling technique as in [7] for dealing with imbalanced data sets seems to work well as observed from the high accuracy rate obtained by their algorithm in the mammography data set. Granularity-based algorithms for discretization of attributes proposed by [8] perform better in the case of the imbalanced Wisconsin data set. These suggest some interesting future research directions for further consideration.

## 5 Conclusion

An extended multi-objective memetic algorithm (E-MOMA) was proposed to mine rules from data sets. The incorporation of memory element made it possible for a local search criterion to be exploited by the proposed technique. Using a higher probability for the elite individuals and lower probability for weaker individuals is to be chosen

for reproduction enabled both exploration and exploitation of the population space. The algorithm was applied to bench mark data sets from the UCI machine learning repository. The performance of the algorithm in terms of accuracy, complexity, and time taken to mine the classifier was studied by comparing the E-MOMA with a simple MOGA. E-MOMA outperforms GA in all the data sets in terms of accuracy and the number of rules presented to the user. This was used to highlight the importance and power of combining multi-objective optimization and evolutionary systems in mining accurate and compact rules in less time. The proposed E-MOMA was also compared with other multi-objective evolutionary systems found in the literature. E-MOMA outperforms all the algorithms in terms of accuracy except in a couple of imbalanced data sets by other algorithms that have used sampling techniques, fuzzy techniques, and granularity techniques for data preprocessing stage before applying rule mining techniques. This motivates further research explorations in the following directions. As future research, we would like to consider different sampling techniques to deal with imbalanced data sets, and other discretization techniques need further exploration in increasing accuracy of classification.

## References

1. F. Jiménez, G. Sánchez, J.M. Juárez, Multi-objective evolutionary algorithms for fuzzy classification in survival prediction. *Artif. Intell. Med.* **60**, 197–219 (2014). <https://doi.org/10.1016/j.artmed.2013.12.006>
2. L. De Almeida Ribeiro, A. Da Silva Soares, T.W. De Lima, C.A.C. Jorge, R.M. Da Costa, R.L. Salvini, C.J. Coelho, F.M. Federson, P.H.R. Gabriel, Multi-objective genetic algorithm for variable selection in multivariate classification problems: a case study in verification of biodiesel adulteration. *Procedia Comput. Sci.* **51**, 346–355 (2015). <https://doi.org/10.1016/j.procs.2015.05.254>
3. S. Elhag, A. Fernández, A. Bawakid, S. Alshomrani, F. Herrera, On the combination of genetic fuzzy systems and pairwise learning for improving detection rates on Intrusion Detection Systems. *Exp. Syst. Appl.* **42**, 193–202 (2015). <https://doi.org/10.1016/j.eswa.2014.08.002>
4. E.C. Gonçalves, A. Plastino, A.A. Freitas, Simpler is better : a novel genetic algorithm to induce compact multi-label chain classifiers, in *Proceedings of the 2015 Annual Conference on Genetic and Evolutionary Computation (GECCO 2015)* (2015), pp. 559–566. <https://doi.org/10.1145/2739480.2754650>
5. M.K. Ghorabae, E.K. Zavadskas, L. Olfat, Z. Turskis, Multi-criteria inventory classification using a new method of evaluation based on distance from average solution (EDAS). *Informatica* **26**, 435–451 (2015). <https://doi.org/10.15388/Informatica.2015.57>
6. J. Jacques, J. Taillard, D. Delerue, C. Dhaenens, L. Jourdan, Conception of a dominance-based multi-objective local search in the context of classification rule mining in large and imbalanced data sets. *Appl. Soft Comput. J.* **34**, 705–720 (2015). <https://doi.org/10.1016/j.asoc.2015.06.002>
7. J. Li, S. Fong, Y. Sung, K. Cho, R. Wong, K.K.L. Wong, Adaptive swarm cluster-based dynamic multi-objective synthetic minority oversampling technique algorithm for tackling binary imbalanced datasets in biomedical data classification. *BioData Min.* **9**, 1–15 (2016). <https://doi.org/10.1186/s13040-016-0117-1>
8. M. Antonelli, P. Ducange, B. Lazzerini, F. Marcelloni, Multi-objective evolutionary design of granular rule-based classifiers. *Granul. Comput.* **1**, 37–58 (2016). <https://doi.org/10.1007/s41066-015-0004-z>

9. V. Basto-Fernandes, I. Yevseyeva, J.R. Méndez, J. Zhao, F. Fdez-Riverola, T.M. Emmerich, A spam filtering multi-objective optimization study covering parsimony maximization and three-way classification. *Appl. Soft Comput. J.* **48**, 111–123 (2016). <https://doi.org/10.1016/j.asoc.2016.06.043>
10. S. Chatterjee, S. Sarkar, N. Dey, S. Sen, T. Goto, N.C. Debnath, Water quality prediction: Multi objective genetic algorithm coupled artificial neural network based approach, in *2017 IEEE 15th International Conference on Industrial Informatics (INDIN)* (2017). <https://doi.org/10.1109/INDIN.2017.8104902>
11. N. Nguyen, B. Liu, V. Pham, T. Liou, An efficient minimum-latency collision-free scheduling algorithm for data aggregation in wireless sensor networks. *IEEE Syst. J.* **12**(3), 2214–2225 (2018). <https://doi.org/10.1109/JSYST.2017.2751645>
12. G.R. Nitta, B.Y. Rao, T. Sravani, N. Ramakrishiah, M. Balaanand, LASSO-based feature selection and Naïve Bayes classifier for crime prediction and its type. *SOCA* **13**(3), 187–197 (2019). <https://doi.org/10.1007/s11761-018-0251-3>
13. L. Pan, C. He, C. He, Y. Tian, H. Wang, X. Zhang, Y. Jin, A classification based surrogate-assisted evolutionary algorithm for expensive many-objective optimization. *IEEE Trans. Evol. Comput.* 1–15 (2018). <https://doi.org/10.1109/TEVC.2018.2802784>
14. A. Alhindi, A. Alhindi, A. Alhejali, A. Alsheddy, N. Tairan, H. Alhakami, MOEA/D-GLS: a multiobjective memetic algorithm using decomposition and guided local search. *Soft Comput.* **23**, 9605–9615 (2019). <https://doi.org/10.1007/s00500-018-3524-z>
15. S.K. Balarisingh, W. Ding, S. Vipsita, S. Bakshi, A memetic algorithm using emperor penguin and social engineering optimization for medical data classification. *Appl. Soft Comput. J.* **85**, 105773 (2019). <https://doi.org/10.1016/j.asoc.2019.105773>
16. F. Cheng, J. Chen, J. Qiu, L. Zhang, A subregion division based multi-objective evolutionary algorithm for SVM training set selection. *Neurocomputing* **394**, 70–83 (2020). <https://doi.org/10.1016/j.neucom.2020.02.028>
17. S. Che, W. Yang, W. Wang, A memetic algorithm for community detection in signed networks. *IEEE Access* **8**, 123585–123602 (2020)
18. M. Fazzolari, R. Alcalá, F. Herrera, A multi-objective evolutionary method for learning granularities based on fuzzy discretization to improve the accuracy-complexity trade-off of fuzzy rule-based classification systems: D-MOFARC algorithm. *Appl. Soft Comput. J.* **24**, 470–481 (2014). <https://doi.org/10.1016/j.asoc.2014.07.019>

# An Intelligent Road Transportation System



S. Muruganandam, K. R. Ananthapadmanaban, and Sujatha Srinivasan

**Abstract** Traffic congestion is caused by inefficient road operations and by excess demand of the travelers. Various researchers have submitted different recommendations for resolving the traffic congestions and for providing optimum routes to travel from a particular origin to the specified destination with less time and less fuel consumption. Historical traffic data has been used in traffic recommendation systems to present optimum travel routes to road users. The process of finding the optimum navigational path for a particular route is the personalization of the transportation system. The personalization makes the transportation into an intelligent transportation system (ITS). The intelligent transportation system stores the frequent traveler's optimal path to travel from any source to the destination on a particular day at a particular time in the database. The artificial neural network (ANN) is then used for data analysis for making predictions in various domains. The optimum path is the path which takes the least travel time with least fuel consumption. The personalized intelligent system assists travelers with the optimum path using the PrefixSpan algorithm.

**Keywords** PrefixSpan algorithm · Intelligent transportation system · Artificial neural network · Personalization systems

## 1 Introduction

Traveling has become a new kind of freedom from stress, and it has become a necessity of life to commute to different places for work and other chores. Due to the competition among car manufacturers, the cost of cars has come down to low

---

S. Muruganandam · S. Srinivasan (✉)

Department of Computer Science, SRM Institute for Training and Development, Chennai, Tamilnadu, India

K. R. Ananthapadmanaban

Faculty of Science and Humanities, SRM Institute of Science and Technology, Vadapalani Campus, Chennai, Tamilnadu, India

e-mail: [kranantp@srmist.edu.in](mailto:kranantp@srmist.edu.in)

prizes that the number of cars on the roads have increased. In spite of more number of highways coming up in almost all the world countries, most of the cities in the world face traffic congestion problems that curtail the traffic flow. In the USA, statistics state that traffic congestion budgets to 5.48 billion delayed hours and 2.87 billion gallons of wasteful fuel daily. The total loss suffered in the cost is estimated as \$120 billion. In a study at IIT-Delhi, vehicles in the Indian capital crawl at snail pace for a sizeable amount of time during a daily commute. 24% of the vehicle run less than 4 km/h which results in the wastage of time and fuel consumption.

Congestion can be abridged in any one of the following approaches:

- i. Refining the infrastructure like intensifying the road capacity which requires enormous expenditures
- ii. Promoting the public transport system
- iii. Managing the traffic using intelligent traffic recommendation Systems (ITRS) that suggest optimal paths for the traveler.

The present paper proposes such an intelligent traffic recommendation system for providing better travel experiences for road commuters. The rest of the paper is organized as follows. Section 2 reviews the latest literature on such ITRS systems. Section 3 gives an overview of the problem and explains the methodology of the PrefixScan algorithm for intelligent travel recommendations. Section 4 presents the results while Sect. 5 concludes with future research directions.

## 2 Literature Survey on Intelligent Transportation System

Kim et al. [1] proposed a predictive control system to predict and prevent traffic congestion by developing an advanced routing algorithm using machine learning. Hamdi et al. [2] suggested a model in which a network called as VANET, vehicular ad hoc network is applied for an intelligent transportation system which reduces the travel time to reach from a particular source to the destination. Kum et al. [3] proposed a model called as multi-task learning to foresee the network-wide traffic speed and to progress the prediction performance. Kammoun et al. [4] proposed an adaptive multi-agent system constructed on the ant colony behavior and the hierarchical fuzzy model. It consents fine-tuning competently the road traffic according to the real-time challenges in the road networks by assimilating an adaptive vehicle route supervision system. Viahogianni et al. [5] suggested a model with an intelligent transportation system (ITS) research application to model traffic characteristics and to produce projected traffic conditions. Aloysius et al. [6] suggested an innovative model that describes how the goods in the retail marketing are sequenced according to the customer's behavior of purchasing pattern of the goods with less ingestion of energy and time. Wang et al. [7] proposed an emerging platform-based intelligent transportation system to explore a crowd sensing-based framework to deliver timely response for traffic management in heterogeneous communal internet of vehicles. Muruganandam and Srinivasan [8–10] have mounted a personalized

model to personalize the e-learning courses adapting to the user's requirements or wishes with the application of PrefixSpan algorithm to mine the learning patterns. Skiy et al. [11] suggested a model in which new ITS technologies such as 'Integrated Corridor Management System' are promoted with comprehensive traffic quantification system. Chowdhary et al. [12] proposed a system that accomplishes the road traffic in a city deprived of smearing the decision-making process by human personnel and also track the vehicles that intrude upon signals at crossing points. Liu and Yu [13] suggested a noel route recommendation system to overcome the strain of traffic jams and long queuing problems in tourist spots. The system personalizes the visiting paths on the predilections of the tourists. Lanke [14] has recommended a model to succeed the traffic congestion problem with a new technology called as radio frequency identification (RFID). The system acts as the key to smart traffic management to lessen the traffic congestion to save time and money.

### 3 Problem Statement

Various researchers have used the historical traffic data in their recommendation systems to yield the optimum journey taking less time and less fuel consumption to travelers. The process of discovering the finest navigational path for a particular route is the personalization of the transportation system. The personalization brands the transportation into an ITS. The intelligent transportation system retains the historical data of the frequent traveler's optimal path to travel from any source to the destination on a particular day at a particular time in the database and analyzes the data using ANN. The ANN is the efficient method to incorporate the data analysis for making predictions in various IT-related fields. The optimum path is the path which takes the least travel time with least fuel consumption. The personalized intelligent system innovates and assists travelers by giving optimum path recommendations. Apriori and PrefixSpan algorithms are the latest data mining sequential patterns used by various researchers in optimizing travel itineraries. The usage and rich features of PrefixSpan algorithm is superior over all the data mining sequential pattern algorithms as found in the literature.

#### 3.1 *PrefixSpan Algorithm*

The recommended system assists the traveler to choose the optimum path from the location he/she begins his/her travel to the location where he/she has decided to reach. PrefixSpan algorithm is applied to mine the corresponding sequential pattern with respect to the factors such as time and fuel consumption.

The steps in the PrefixSpan algorithm are as follows:

*Step 1.* The sequential database that keeps the history of the routes used by the previous travelers is scanned to find the frequent item (frequent path used by the frequent travelers to travel) with respect to minimum support value (minimum support threshold).

*Step 2.* Partition the sequential database into multiple subsets of sequential patterns.

*Step 3.* Mine the subsets of sequential pattern.

*Step 4.* Generate the projected database to mine the  $(s + 1)$  sequence for every frequent k-sequences,

The following is the PrefixScan algorithm

```

Algorithm PrefixSpan (Seqp, Seq1, Seqp | lenseq)

// Seqp is the Sequential pattern, Seq1 is the length of
// the sequential pattern
// Seqp | lenseq is the projected database , Seqdb is the
// sequence data base

Do
    Scanning Seqp | lenseq once, to identify the set of
    frequent patterns fsi such that fsi is accumulated
    to Seqp as the last item ;

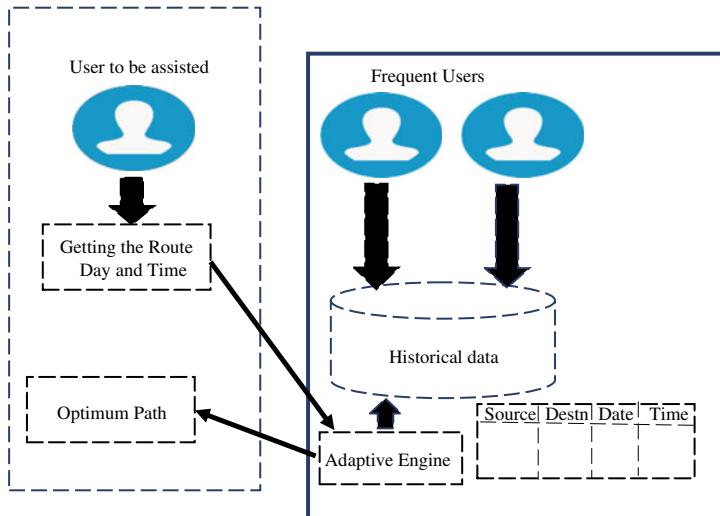
    Make the Sequential pattern otherwise fsi is
    affixed to seqp;

    To form the sequential pattern Seqp1 and
    output Seqp1, each frequent item is appended to
    Seqp ;

    Build the Seqp- projected database with each
    value of Seqp1 and do recursive call
PrefixSpan (Seqp1, Seq1, Seqp | lenseq):
```

### 3.2 Methodology

The above algorithm is called with the sequence database Seq<sub>db</sub> and the minimum threshold support min\_threshold\_support. There are various routes from various origins to the destinations which are followed by the traditional travelers to travel. Irrespective of the day and time they travel, they follow the path which is unguided. The unguided or assisted path used by the travelers consumes more fuel and more



**Fig. 1** Flow diagram for the proposed personalized model

time to travel from the specified origin to the specified destination. If the travelers are provided with the guidance formulated by the suggested model, they are able to reach the destination in an optimal path. The optimal path takes less time-consuming and less fuel consumption compared to the traditional path used by the unguided travelers.

The proposed model is developed with the application of PrefixSpan algorithm in which historical data of the travelers is used to assist the traveler to choose the premium path from various paths from the origins to the destinations. Figure 1 shows the architecture of the proposed model for personalizing the transportation system.

The sequencing of paths classification and the subsequent resultant classifications are mined with the help of PrefixSpan algorithm. Table 1 gives the comparison data of the normal traveler and the assisted traveler to travel from a particular source  $S_k$  to a particular destination  $D_{l,l}$ , route  $R_{k,l}$  on a particular day (here it is Monday).

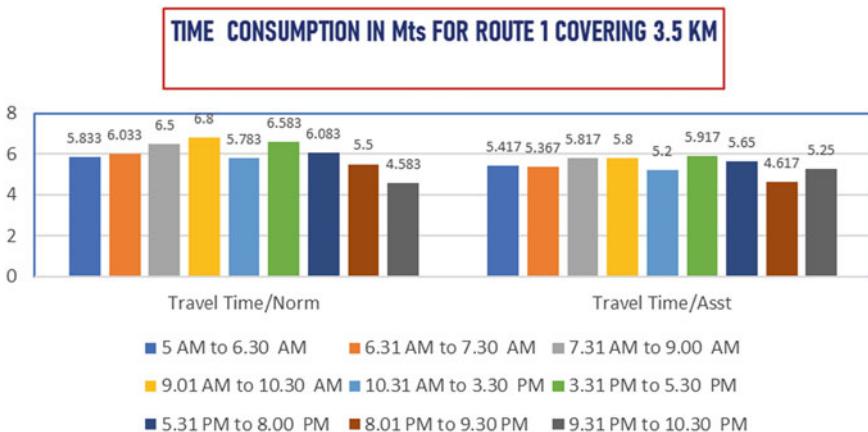
## 4 Results and Discussion

Figure 2 shows the comparison chart on the basis of analysis data for fuel consumption of the traditional traveler and the assisted traveler for the route covering 3.5 KMs on a particular day (it is on Monday).

Figure 3 shows the comparison chart on the basis of analysis data for time consumption of the traditional traveler and the assisted traveler for the same route on that day.

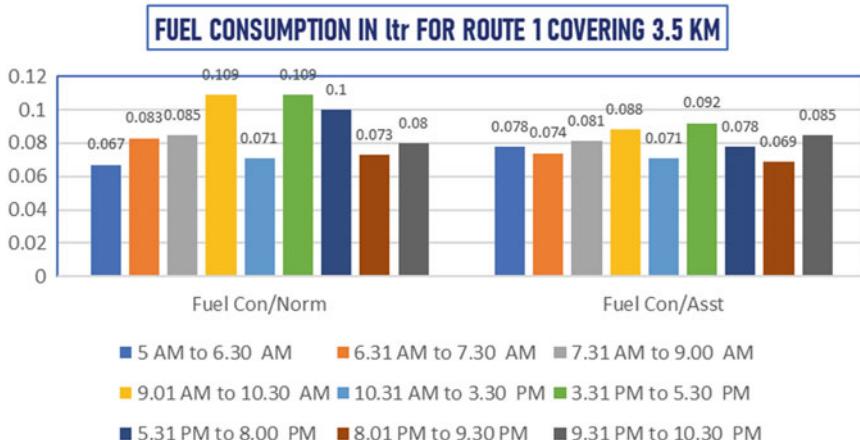
**Table 1** Comparison data of the normal and assisted traveler

Time slots	Travel time (in min)		Fuel consumption (in L)	
	Normal	Assisted	Normal	Assisted
5 AM to 6.30 AM	5.833	5.417	0.067	0.078
6.31 AM to 7.30 AM	6.033	5.367	0.083	0.074
7.31 AM to 9.00 AM	6.5	5.817	0.085	0.081
9.01 AM to 10.30 AM	6.8	5.8	0.109	0.088
10.31 AM to 3.30 PM	5.783	5.2	0.071	0.071
3.31 PM to 5.30 PM	6.583	5.917	0.109	0.092
5.31 PM to 8.00 PM	6.083	5.65	0.1	0.078
8.01 PM to 9.30 PM	5.5	4.617	0.073	0.069
9.31 PM to 10.30 PM	4.583	5.25	0.08	0.085

**Fig. 2** Comparison chart of the normal and assisted traveler on time consumption

#### 4.1 Discussion

The performance of the personalized and proposed model is compared with the traditional model. The performance is measured in accordance with the consumptions of time and fuel to travel from the sources ( $S_i$ ,  $i = 1$  to  $m$ ) to the destinations  $D_j$  ( $j = 1$  to  $n$ ). Table 1 shows the comparison data between the traditional and the assisted travelers to travel from a particular source ( $S_k$ ) to a particular destination ( $D_l$ ) over a specific period of time intervals between 5 AM and 10.30 PM on a particular day (here, it is Monday) for the route ( $R_t$ ,  $t = 1$  to  $m * n$ ) on the basis of travel time and fuel consumption. Chart 1 is the comparison graph of traditional and assisted travelers. The chart clearly confirms that the travel time and fuel consumption for different period of time intervals on the day for the particular route are minimum for the



**Fig. 3** Comparison chart of the normal and assisted traveler on fuel consumption

assisted travelers than the traditional traveler. The comparison of the data in respect of travel time and fuel consumption for various routes of different time intervals on different days (except Sunday) are also analyzed. The analysis has proved that the assisted travelers who are assisted with the personalized recommendation model are benefitted from the model by taking less travel time and less consumption of fuel.

## 5 Conclusion

The experiment shows that the travelers who are supervised with the intelligent transportation model are benefited with the assistance provided on the basis of fuel consumption and time consumption. The experiment was conducted between a particular source and the destination. In the future, it may be extended to various sources and destinations. Its application can be improved by implementing mobile GPS systems.

## References

1. S. Lee, Y. Kim, Intelligent traffic control for autonomous vehicle systems based on Machine Learning. *Exp. Syst. Appl.* **144** (2020). <https://doi.org/10.1016/j.eswa.2019.113074>
2. M.M. Hamdi, L. Audah, A.J. Rashid, A review of applications, characteristics and challenges in vehicle ad hoc network (VANETS). *IEEE Xplore, HORA* (2020). INSPEC Accession Number: 19970560. <https://doi.org/10.1109/HORA49412.2020.9152928>
3. Z.K. Peng, Z. Liang, A deep Learning based multitask model for network—wide traffic speed prediction. *Neuro Comput.* **396**, 438–450 (2020). <https://doi.org/10.1016/j.neucom.2018.10.097>

4. H.M. Kammoun, I. Kallel, J. Casillas, A. Abraham, A.M. Alimi, Adapt-Traf: an adaptive multi agent road traffic management system based on hybrid anti-hierarchical fuzzy model. *Transp. Res. Part C Emerg. Technol.* **42**, 147–167 (2014). <https://doi.org/10.11016/j.trc.2014.03.003>
5. E.I. Vlahogianni, M.G. Karlaftis, J.C. Golias, Short-term traffic forecasting: where we are and where are we going. *Transp. Res. Part Emerg. Technol.* **43**(Part 1), 3–19 (2014). <https://doi.org/10.1016/j.trc.2014.01.005>
6. G. Aloysius, D. Binu, An approach to product placement in supermarket using the Prefix span algorithm. *J. King Saudi Univ. Comput. Inform. Sci.* **25**, 77–87 (2013)
7. X. Wang, Z. Ning, X. Xu, E.C.-H. Nagai, A city—wide real-time traffic management systems: enabling crowdsensing in social internet of vehicle. *IEEE Commun. Mag.* **56**(9), 19–25 (2018).<https://doi.org/10.1109/MCOM.2018/1701065>
8. B. Maram, J.M. Gnanasekar, G. Manogaran, M. Balaanand, Intelligent security algorithm for UNICODE data privacy and security in IOT. *SOCA* **13**(1), 3–15 (2018). <https://doi.org/10.1007/s11761-018-0249-x>
9. B.H. Liu, N.T. Nguyen, V.T. Pham, An efficient method for sweep coverage with minimum mobile sensor, in *2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, Kitakyushu, Japan (2014), pp. 289–292. <https://doi.org/10.1109/IIH-MSP.2014.78>
10. S. Muruganandam, N. Srinivasan, Personalizing e-learning system for courses using prefix span algorithm. *J. Theor. Appl. Inform. Technol.* **74**(2) (2015). ISSN 1992-8645, [www.jatit.org](http://www.jatit.org), E-ISSN 1817-3195
11. A.A.K. Skiy, P. Xariya, Traffic management: an outlook. *Econ. Transp.* **4**(3), 135–146 (2015). <https://doi.org/10.1016/j.jecotra.2015.03.002>
12. P.R. Chowdhary, S. Das, Automatic road management system in a city. *STM J.* 38–46 (2014). TTEA
13. L. Lu, J. Yu, A real-time personalized route recommendation system for self-drive tourist based on vehicle to vehicle communication. *Exp. Syst. Appl.* **4**(7), 3409—3417 (2014)
14. S.N. Lanke, Smart traffic management system. *Int. J. Comput. Appl.* **75**(1), 109 (2013) (0975-8887)

# Securing Data from Active Attacks in IoT: An Extensive Study



C. Silpa, G. Niranjana, and K. Ramani

**Abstract** The Internet of Things (IoT) is one of the most important technologies aimed at improving the quality of human lives. IoT plays a promising role in many business applications including healthcare, automotive, agriculture, and education. But the crucial task is to address and analyze the IoT security issues because the IoT environment has a heterogeneous nature of data. In order to achieve end-to-end secure environment, it is necessary to change the design of IoT application's architecture. Therefore, a detailed study of security issues of IoT devices, its limitations, requirements of security, and possible solutions are presented in this research study. In addition, various sources of security threats at different layers of IoT environment are presented. The main contribution of this study is to classify threats and potential IoT security challenges from an architectural perspective. After discussing various security issues, a number of emerging and existing technologies are presented that are focused on achieving higher user confidence in IoT applications.

**Keywords** Agriculture · End-to-end secure · Heterogeneity · Internet of Things · Security · Threat

## 1 Introduction

Nowadays, Internet of Things (IoT) gained more interest among various researchers and applied in different domains namely transportation, health care, smart buildings, power grids, and entertainment, because it connects physical devices to the Internet and provides data to people. In the upcoming years, IoT is highly utilized in the revolutions of future technical applications due to its significant role [1]. However,

---

C. Silpa (✉) · G. Niranjana  
Department of CSE, SRMIST, Chennai, India

G. Niranjana  
e-mail: [niranjag@srmist.edu.in](mailto:niranjag@srmist.edu.in)

K. Ramani  
Department of IT, SVEC, Tirupati, India

security is one of the major issues that occurred in IoT because of vast amount of data are formed and associated with a greater number of IoT devices [2]. The data are transmitted to/from cloud-based resources by using standard communication protocols such as Hyper Text Transfer protocol (HTTP), which is connected with global Internet and IoT devices. In the foundations of Internet, those devices have various targeted attacks namely malicious modification, man-in-the-middle, buffer overflow, Distributed Denial-of-Service (DDoS), password cracking by brute force attack, and so on, where these attacks are formed due to multiple vulnerabilities in IoT devices [4]. Moreover, the Wireless Sensor Networks (WSN), Internet, and cellular networks have other challenges along with security issues such as authentication issues, information storage, privacy issues, management issues, and so on. For instance, the location of individual can be tracked by targeting IoT devices and providing false data by attackers [5]. There are four important layers presented in any IoT environment, where different functionalities are performed by perceiving the information/data with the help of different actuators and sensors in first layer. According to collected data from first layer, that information are transmitted using communication network in second layer [6]. A middleware layer is defined as third layer, which acts as bridge between the application and network layer and most of the IoT applications are deployed in this third layer [7]. The end-to-end applications based on IoT devices namely smart factories, smart grids, smart transport, etc. are presented in the fourth layer, where the problems of security are available in all of these four layers [8, 9]. In addition, the data are effectively moved by connecting these layers using different gateways, and different security threats are also occurred in these gateways as well [10].

The organization of the paper consists of: the major application areas of IoT are depicted in Sect. 2, the IoT requirements in terms of security are provided in Sect. 2. The different security threats with various layers are presented in Sect. 3. The survey of various existing techniques with its limitations is described in Sect. 4. The conclusion and future directions of IoT environment are illustrated in Sects. 5 and 6 respectively.

## 2 Application Areas of IoT in Terms of Security

Security is one of the major important tasks in almost every IoT application, because many industries are rapidly increased using connected IoT devices. Even though the traditional networking technologies provide security to the applications, various IoT environments require security in the most important applications of IoT that are discussed as follows:

## 2.1 Smart Cities

In order to improve the overall quality of human life, new computers and communication sources are widely used in smart cities. Smart traffic management, smart utilities, smart disaster management, and smart homes are the major areas included in smart cities. The governments and various private agencies are developing different IoT applications around [11]. Even though the human life's quality is improved by the usage of intelligent applications, the citizens' privacy is under threat. A sensitive data may be presented in some of the buildings or homes, which must be protected from unauthorized access or hackers. The location of users are tracked and monitored by using smartphone applications and it may leak sensitive information. There are many different apps that are followed by parents to track their child and if those apps are accessed by unauthorized users, the safety of children is at risk.

## 2.2 Intelligent Environment

This application consists of preventing landslides, detecting the early earthquakes and forest fires, monitoring the pollutions and snow levels in high altitude regions, which are closely related to the animals and human life in those areas. Government agencies involved in these areas also rely on information that is obtained from those applications [12–14]. Security breaches and vulnerabilities can have serious consequences in all areas related to these IoT applications. A disastrous result is obtained by using false positives and false negatives in this context for smart environment applications. Hence, it should avoid data tampering, security breaches and also should be highly accurate.

## 2.3 Security and Emergencies

It is one kind of application in IoT environments, where the application includes allowing only an authorized person in prohibited areas. In industrial zones, the leakage of dangerous gases is tracked and identified for emergency applications. In the areas of cellular base stations or nuclear reactors, the radiations levels are monitored and created an alarm, when the level reaches its normal limit. In order to prevent breakdowns or corrosions in those sensitive buildings, various liquids must be detected by IoT applications. Moreover, different consequences are available in such buildings that lead to security breaches.

## 2.4 Smart Agriculture and Farming of Animals

Irrigation selections in dry zones, tracking the moistures in soils, humidity, microclimate conditions, and temperature controls are the major applications in smart agriculture. Using these advanced features of agriculture will help farmers to achieve higher yields and save their money from losses. Regulating the temperature and moisture content of various cereal and vegetable crops can help to prevent fungal and other microorganisms [15]. Controlling climatic conditions can help to increase the yield and this leads to improving the quality of vegetables and crops. In addition, the health conditions and activities of animals are tracked and monitored by using the attached sensors in those farm animals is called animal farming. If such farming environments are compromised, this could lead to stealing animals from the farm and the opponents destroying crops.

## 2.5 Home Automation

In the environment of IoT, this automation is one of the most used and deployed application areas, where it consists of automatically controlling the electronics to save energy and identifying the intruders by attaching the sensors on doors and windows. In order to save resources and costs, users tracked the consumption of water supply and energy by using home automation applications [16]. However, hackers may intrude the systems and harm the users by gaining unauthorized access at home.

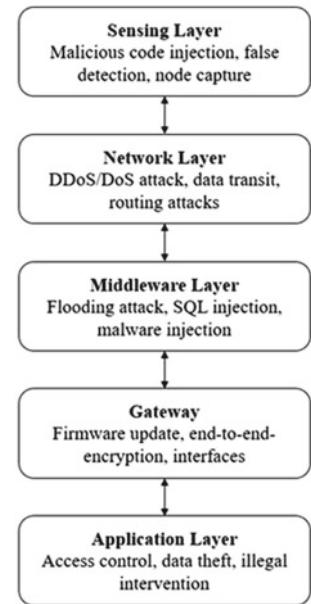
# 3 Security Threats in IoT Applications

There are four major layers such as sensing, network, middleware, and application layers are presented in any IoT application. A vast amount of issues and threats are occurred because diverse technologies are used by every layer in an application. Different possible security threats of four layers are studied in this section, where Fig. 1 illustrates the four layers of IoT systems with its security threats. Moreover, a gateway is used to connect these layers that also have special security issues are presented in the following section.

## 3.1 Security Issues in Sensing Layer

The physical IoT actuators and sensors are dealt with using sensing layers, where the physical phenomenon is sensed by the sensors and according to the sensed data, a

**Fig. 1** Different types of attacks in IoT at various layers



certain action on the environment is performed by actuators. Different kinds of data are sensed by various sensors namely smoke detection sensors, ultrasonic sensors, camera sensors, temperature and humidity sensors. The security threats included in this layer such as node capturing, false data injection attacks, malicious code injection attacks, side channel attacks, sleep deprivation attacks, booting attacks, eavesdropping and interference. Some of the major threats that occurred in sensing layers are discussed as:

*Attack by malicious code injection:* In the node memory, some malicious codes are injected by attackers, which is involved in this kind of attack. The IoT node's software or firmware is updated using gateways and this leads to injecting malicious codes on gateways by attackers. Some intended functions are performed or try to access the complete systems by intruders using such malicious codes.

*Injection attack using false data:* An attacker may inject vast amount of data on the IoT system, once the capturing of data is carried out. Therefore, IoT applications are malfunctioned or the results may be incorrect due to these attacks. Moreover, a DDoS attack is caused by using this attack in IoT systems.

*Interference and Eavesdropping:* In open environments, different nodes are deployed in IoT applications. Therefore, eavesdroppers have resulted in those IoT systems. During various operations like authentication and transmission of data, those data are captured by attackers using eavesdrop.

### 3.2 Security Issues in the Network Layer

The information obtained from the sensing layers is transmitted to the computational unit for processing by network layer, which is the major function of the layer. Phishing site attack, data transit attack, routing attack, DDoS/DoS attack, and access attack are the security threats included at network layer. In the below section, the major security threats are depicted as follows:

*Data Transit Attacks:* In everyday human life, data are exchanged and stored in the environment, which is the main focus of the attackers and other competitors. The cloud or local servers are used to store sensitive information, however, those data are transmitted from one server to others and are more vulnerable to attacks. Massive data are transmitted between cloud, sensors, actuators, etc. in the environment. Hence, data breaches are the most vulnerable attack for IoT devices.

*Routing Attacks:* During data transmission, the routing paths are redirected by injecting the malicious nodes in the system. An artificial shortest routing path is introduced by adversary and leads the nodes to traffic is defined sinkhole attacks. When combining other attacks with sinkhole attacks, a serious security threat is occurred, which is defined as worm-hole attack. In order to speed up the packet transfer, band connection between two nodes is formed by warm-hole. Therefore, basic security protocols becomes vulnerable by creating the wormholes between a device and compromised node.

*DDoS / DoS Attack:* A massive amount of unwanted requests are sent to the target cloud servers by attackers in DDoS/DoS attacks. The services are interrupted by the authorized users by disturbing the target server. The attackers flood the target servers by using multiple sources is defined as DDoS attack, which is not specific to IoT applications due to the complexity and heterogeneity of IoT environment. The attackers easily launch DDoS attacks on the servers, because devices are not strongly configured with each other.

### 3.3 Security Issues in the Middleware Layer

An abstraction layer is created between the application layer and network layer, which is the major role of the middleware in IoT systems. High storage and computing capabilities are provided by middleware and the demands of application layers are fulfilled by presenting APIs in this layer. The queuing systems, brokers, machine learning, persistent data stores, etc. are included in the middleware layer. Even though reliable and robust IoT applications are provided by this layer, a different attacks namely SQL injection attack, signature wrapping attack, cloud malware injection, man-in-the-middle attack, and flooding attack are susceptible. The middleware is injected by attackers and entire IoT systems are accessed by intruders. Therefore, some important attacks are described as follows.

*Middle-in-the-man Attack:* The MQTT broker or proxy uses publish-subscribe modes for the communication between subscribers and clients in the MQTT protocol. The clients of publishing and subscribing are decoupled with each other and the messages are sent without the destination knowledge. Suppose, the broker is controlled by attackers and becomes a man-in-the-middle for accessing the complete control of user's communication without the client's knowledge.

*Signature wrapping attack:* The signature of XML is used in the web services that are presented in the middleware. The signature algorithms are broken for executing or modifying the operation/eavesdropped message by attackers using vulnerabilities in simple object access protocol.

*Cloud Flood Attack:* The quality of service in cloud is affected by this flooding attack, which is same as DoS attack. Multiple requests are continuously forwarded to the cloud service providers by attackers for deleting the cloud resources. This leads to an increase in the load on the servers and privacy issues of user's data.

### 3.4 Security Issues in Gateways

The connection between cloud services, humans, and multiple devices are conducted by gateways, which plays a major role in IoT applications. The hardware and software solutions for devices are also provided by gateways. While transmitting the protocols for communication between layers, the IoT data are encrypted and decrypted by using gateways. In general, IoT systems are heterogeneous in nature that consists of TCP/IP, Z-Wave, and ZigBee stacks with various gateways. The major threats for the gateways are presented below.

*Secure Onboarding:* The encryption keys must be protected when a new sensor or devices are installed in a system. The communication between managing services and new devices is carried out by using gateways, which acts as an intermediary and pass all keys between them. In order to capture the encryption keys, the eavesdropping, and man-in-the-middle attacks have occurred at gateways during the process of onboarding.

*Additional interfaces:* Reducing attack surfaces is an important strategy to consider when installing IoT devices. An IoT gateway manufacturer should only implement the necessary interfaces and protocols. End users must restrict certain services and features for preventing data breach or backdoor authentication.

*End-to-end encryption:* The confidentiality of data must be ensured by securing the end-to-end application layers. The encrypted messages are decrypted in the application layers only by allowing the authorized users rather than other users. The researchers used the Z-wave and Zigbee protocols for supporting the encryption, but end-to-end encryption is not supported by these protocols. The reason is that due to

the requirement of gateways for decryption and re-encryption messages while transmitting information from one protocol to another. Therefore, information breaches may occur at the gateway levels, when decrypting the data.

*Updates on Firmware:* There is no computation power or user interface for downloading and installing the updates on firmware, because resource constrained are presented in many IoT devices. Normally, the operation of updating and downloading the firmware is conducted by gateways. In order to secure the firmware updates, the signature's validity must be verified and a new version of the firmware must be recorded.

### 3.5 Security Issues in Application Layer

The end users achieved the services from the IoT environment using the application layer, which includes smart grids, smart cities, smart homes, etc. While comparing with other layers, specific problems such as privacy issues and data theft are presented in this application layer. There is a sublayer between application layer and network layer is available in various IoT applications, where the sublayer is defined as middleware layer or application support layer. The below section discusses the major issues obtained in this layer that are as follows:

*Data Theft:* A lot of private and important data are available in the IoT applications and it is even more vulnerable to attacks when the data is transmitted from one device to another. The users will hesitate to store their sensitive data when the applications are vulnerable to this kind of attack. Therefore, the researchers designed privacy management, data encryption, network as well as user authentication techniques and protocols for securing the user's data against thefts in IoT.

*Service Disruption Attacks:* the existing techniques, these attacks are also defined as DDoS attacks or illegal interruption attacks. Artificially, an attacker makes the network/servers too busy for responding to the application services and then, authorized users are prevented by such attacks.

## 4 Comprehensive Study of Existing Systems

In this section, a discussion of various existing techniques that are developed to detect various attacks in IoT is presented. The key benefits of the existing techniques with limitations are provided in Table 1 from the year 2017 to 2020.

**Table 1** Key benefits of the existing techniques with limitation

Author with year	Methodology	Major Attacks	Advantage	Limitation	Performance metrics
Challa et al. [17] (2017)	Developed a signature-based authentication scheme for detecting various attacks	The attacks include user impersonation, denial-of-service, replay, stolen smart card, man-in-the-middle, offline password guessing, privileged-insider attacks are considered	The simulations are conducted on NS-2 simulator, where the results proved that the developed scheme is high security with less communication and computation costs	However, the impact on end-to-end delay increases with more number of users that leads to exchange of more messages, and congestions have occurred in the network	The two major parameters such as end-to-end delay and throughput are used to evaluate the impact of the scheme
Yin et al. [18] (2018)	Implemented a framework that contains SD-IoT switches and SD-IoT controllers	An algorithm with cosine similarity of the vectors is designed to identify and mitigate the DDoS attacks	The occurrence of DDoS attack is identified by using the threshold value, where real IDDoS attackers are determined and block those attacks at source itself	However, the method required loading balancing techniques for effective identification of DDoS attacks in the controller pool	Number of packetin messages, controller-switch channel's bandwidth, and number of received packets are used to test the efficiency of the developed method with other existing techniques
Mehmood et al. [19] (2018)	Proposed an algorithm called Naive Bayes classifier that is applied in IDS	The algorithm majorly concentrated on DDoS attacks that are generated by intruders	The nodes actions and irregular traffic are sensed or identified by deploying the IoT in the form of multiagents	But, the number of agents and its type, nature are advanced, hence the Naive Bayes classifiers must be replaced with lightweight pattern algorithm	End-to-end packet forwarding rate, packets drop with time intervals, delays, detection rate with various throughput and detection probability are the major parameters used for testing the developed classifier

(continued)

**Table 1** (continued)

Author with year	Methodology	Major Attacks	Advantage	Limitation	Performance metrics
Rathore et al. [20] (2018)	In order to provide defense and optimal security against cyberattacks, blockchain technology with decentralized security architecture is developed for IoT network	The developed scheme considered three major attacks as ICMP flooding, DDoS attack, and TCP flooding attacks	The time for detecting and migrating the attacks are minimized because the developed scheme is based on layered structure for packing and committing the transactions into new blocks in the developed blockchain	During the blockchain operation, high CPU resources and memory are utilized by fog nodes for packing and committing the transactions into new blocks in the developed blockchain	The evaluation metrics namely Mathew Correlation coefficient, F-score, detection time, accuracy, detection rate, Area under curve, and positive predictive value are used for validating the performance of developed blockchain technique
Dorri et al. [21] (2019)	The end-to-end security is provided in the IoT system by implementing Lightweight Scalable Blockchain (LSB) with Distributed Throughput Management Algorithm (DTM)	Six security attacks are studied in the LSB scheme, where those specific attacks included DoS, DDoS, dropping attack, sybil attack, 51% attack, and consensus period attack	The overhead of mining processing and delay are highly minimized by Distributed Time-based Consensus algorithm (DTC), where the public blockchain is managed by the cluster heads	In the network, the packet overhead and processing time are increased for the new blocks, when the percentage of transaction verification (PTV) value is large	Attack Success Percentage, average processing time, and packet overhead are used for the simulation experiments. Moreover, a case study on smart homes with high and low resource devices is presented

(continued)

**Table 1** (continued)

Author with year	Methodology	Major Attacks	Advantage	Limitation	Performance metrics
Si et al. [22] (2019)	Designed a lightweight sharing security framework with blockchain technology for IoT information, where blockchain is a reliable architecture for multinode information redundancy	Ten important attacks are considered in the developed scheme, where the attacks include DDoS, DosS, additional attack, link drop, device injection, modify attack, destruction time interval, public block modification, and consensus cycle attack	The destruction of collected data or human tampering is prevented by using consensus mechanism, where information transactions and collection of source data are protected by double-chain mode and consensus cycle attack	The problem of IoT information sharing is resolved by the developed scheme, but the leakage of private information has occurred, when the developed scheme is applied in specific industries	Route Success time, weighted time overhead, delay time, and block generation time are used for testing the performance of the developed method <sup>a</sup>

#### **4.1 Observations from Study/Challenges in IoT Security**

From the study of various existing techniques, there are some important issues presented in the security of IoT devices, which are discussed in this section.

*Blockchain Technology in IoT:* According to the system implementation, blockchain security will be enhanced, where the usage of hardware and software are highly presented in those implementations. There is a possible leakage of user's private information when the transactions are occurred by users in public blockchain. The size of blockchain will be increased automatically when more miners increased and this leads to minimizing the distribution speed, chances of issues namely availability, scalability, and high storage cost over the whole network.

*Machine Learning Algorithms in IoT:* Several machine learning algorithms have been introduced into existing techniques. But, the poor accuracy and less effectiveness are highly obtained by choosing the inappropriate algorithms, which produces garbage output, and also, the incorrect results are achieved by using inefficient data. Thus, the selection of diverse data as well as efficient data are the major factors for the successful solutions of machine learning algorithms, it is important to remove and clean the data, however, the process of pre-processing is a crucial task for achieving efficient data accurately. In order to secure the IoT data, the researchers used machine learning techniques by developing the different features namely multiple regression, creation of attributes, linear regression, elimination of redundancies, and compression of data.

The DDoS attacks are raised in transportation layer due to the complexity and diverse nature of the networks, which is also vulnerable. The better solution is to use efficient detection and prevention algorithms for DDoS attacks, where the systems must be upgraded by using those algorithms. But, better detection algorithms are not available recently for solving the DoS/DDoS attacks.

### **5 Conclusion**

With the increasing use of IoT devices in personal lives and many applications, security concerns play an important role. Due to resource constraints and component heterogeneity across different IoT environments, there is a wide range of vulnerabilities have occurred. Much of this vulnerability provided system failures in the IoT environment. Recently many researches have been carried out in IoT, but the challenges of security are presented due to insufficient predetermined standard for the environment of IoT. In this research study, a number of security threats are introduced in different layers of an IoT application. The research study covers the issues related to four layers with gateways are discussed and also some major threats are presented. The privacy of the user information is leaked in the application layer, sensing layer, and network layer during the DDoS and device injection attacks. The complexity and

memory usage of existing techniques are high in the middleware layer during flooding attacks. In addition, existing techniques with its drawbacks and future solutions to IoT security threats including machine learning and blockchain are discussed. These future research directions lead the researchers for enhancing the security levels in IoT's layers.

## 6 Future Directions for Information Security in IoT

Some of the important future directions are described in the section that is derived from the issues of existing techniques.

- In blockchain technology, a collection of addresses and garbage data is formed from the features of tamper-proof. Hence, a vast of unwanted data such as destructed smart contracts' addresses are not removed from the technology that degrades the overall application's performance. In order to effectively handle those garbage data, researchers should design a better blockchain technology.
- In order to reach consensus among nodes, reliable and more effective consensus systems should be developed for preventing more usage of computing power. Because the available consensus algorithms are inefficient and have insufficient resources.
- In the IoT system, a different layer has gateways and this must be secure because attacks can easily intrude into the systems by using gateways. When comparing with particular encryption algorithms for various protocols, end-to-end encryption techniques are highly used for securing the passing data via gateways. For protocol translation, the decryption process will be carried out only at intended destination and didn't occur at gateways.

## References

1. H. Zakaria, N.A.A. Bakar, N.H. Hassan, S. Yaacob, IoT security risk management model for secured practice in healthcare environment. *Procedia Comput. Sci.* **161**, 1241–1248 (2019)
2. S. Rathore, J.H. Park, Semi-supervised learning based distributed attack detection framework for IoT. *Appl. Soft Comput.* **72**, 79–89 (2018)
3. K. Mabodi, M. Yusefi, S. Zandian, L. Irankhah, R. Fotohi, Multi-level trust based intelligence schema for securing of Internet of Things (IoT) against security threats using cryptographic authentication. *J. Supercomput.* 1–25 (2020)
4. X. Li, Q. Wang, X. Lan, X. Chen, N. Zhang, D. Chen, Enhancing cloudbased IoT security through trustworthy cloud service: an integration of security and reputation approach. *IEEE Access* **7**, 9368–9383 (2019)
5. D. Wang, B. Bai, K. Lei, W. Zhao, Y. Yang, Z. Han, Enhancing information security via physical layer approaches in heterogeneous IoT with multiple access mobile edge computing in smart city. *IEEE Access* **7**, 54508–54521 (2019)

6. B. Mukherjee, S. Wang, W. Lu, R.L. Neupane, D. Dunn, Y. Ren, Q. Su, P. Calyam, Flexible IoT security middleware for end-to-end cloud–fog communication. *Futur. Gener. Comput. Syst.* **87**, 688–703 (2018)
7. M.D. Alsbehri, F.K. Hussain, A fuzzy security protocol for trust management in the Internet of Things (Fuzzy-IoT). *Computing* **101**(7), 791–818 (2019)
8. S.K. Sood, Mobile fog based secure cloud-IoT framework for enterprise multimedia security. *Multimedia Tools Appl.* 1–16 (2019)
9. Y. Meidan, M. Bohadana, Y. Mathov, Y. Mirsky, A. Shabtai, D. Breitenbacher, Y. Elovici, N.-baito—network-based detection of IoT botnet attacks using deep autoencoders. *IEEE Pervasive Comput.* **17**(3), 12–22 (2018)
10. Y. Zong, G. Huang, A feature dimension reduction technology for predicting DDoS intrusion behavior in multimedia Internet of Things. *Multimedia Tools Appl.* 1–14 (2019)
11. Y. Liu, C. Yang, L. Jiang, S. Xie, Y. Zhang, Intelligent edge computing for IoT-based energy management in smart cities. *IEEE Netw.* **33**(2), 111–117 (2019)
12. B. Maram, J.M. Gnanasekar, G. Manogaran, M. Balaanand, Intelligent security algorithm for UNICODE data privacy and security in IoT. *SOCA* **13**(1), 3–15 (2018). <https://doi.org/10.1007/s11761-018-0249-x>
13. P.N. Hiremath, J. Armentrout, S. Vu, T.N. Nguyen, Q.T. Minh, P.H. Phung, MyWebGuard: toward a user-oriented tool for security and privacy protection on the web, in *Future Data and Security Engineering, FDSE 2019. Lecture Notes in Computer Science*, edited by T. Dang, J. Küng, M. Takizawa, S. Bui, vol. 11814. (Springer, Cham, 2019). [https://doi.org/10.1007/978-3-030-35653-8\\_33](https://doi.org/10.1007/978-3-030-35653-8_33)
14. R. Arridha, S. Sukaridhoto, S. Pramadihanto, N. Funabiki, Classification extension based on IoT-big data analytic for smart environment monitoring and analytic in real-time system. *Int. J. Space-Based Situated Comput.* **7**(2), 82–93 (2017)
15. O. Elijah, T.A. Rahman, I. Orikumhi, C.Y. Leow, M.N. Hindia, An overview of Internet of Things (IoT) and data analytics in agriculture: benefits and challenges. *IEEE Internet Things J.* **5**(5), 3758–3773 (2018)
16. S. Pirbhulal, H. Zhang, E. Alahi, M. Eshrat, H. Ghayvat, S.C. Mukhopadhyay, Y.T. Zhang, W. Wu, A novel secure IoT-based smart home automation system using a wireless sensor network. *Sensors* **17**(1), 69 (2017)
17. S. Challal, M. Wazid, A.K. Das, N. Kumar, A.G. Reddy, E.J. Yoon, K.Y. Yoo, Secure signature-based authenticated key establishment scheme for future IoT applications. *IEEE Access* **5**, 3028–3043 (2017)
18. D. Yin, L. Zhang, K. Yang, A DDoS attack detection and mitigation with software-defined Internet of Things framework. *IEEE Access* **6**, 24694–24705 (2018)
19. A. Mahmood, M. Mukherjee, S.H. Ahmed, H. Song, K.M. Malik, NBCMAIDS: Näive Bayesian classification technique in multi-agent system-enriched IDS for securing IoT against DDoS attacks. *J. Supercomput.* **74**(10), 5156–5170 (2018)
20. S. Rathore, B.W. Kwon, J.H. Park, BlockSecIoTNet: blockchain-based decentralized security architecture for IoT network. *J. Netw. Comput. Appl.* **143**, 167–177 (2019)
21. A. Dorri, S.S. Kanhere, R. Jurdak, P. Gauravaram, LSB: a lightweight scalable blockchain for IoT security and anonymity. *J. Parallel Distrib. Comput.* **134**, 180–197 (2019)
22. H. Si, C. Sun, Y. Li, H. Qiao, L. Shi, IoT information sharing security mechanism based on blockchain technology. *Futur. Gener. Comput. Syst.* **101**, 1028–1040 (2019)

# Building an Enterprise Data Lake for Educational Organizations for Prediction Analytics Using Deep Learning



Palanivel Kuppusamy and K. Suresh Joseph

**Abstract** Nowadays, educational institutions are one of the biggest producers of data. The rise of e-Learning contents, digital libraries, webinars, learning management systems, online classes and examinations, video surveillance, sensors, and wearables devices contribute to this data explosion. Learning management systems can index millions of students' data, their interactions, course registrations, social networks, and their Internet research results. Besides, the potential to learn from this population-scale data is massive. By building analytic dashboards using machine learning and deep learning approaches on these datasets, educational organizations can improve the learning experience, teaching skills, and learning environment and drive better teaching and learning outcomes. Some real-world examples are students' dropouts, students' behavior, employee and student's health, prevention fraud data and abuse, etc. In present legacy systems, the data silos from the data warehouse could not handle unstructured data. It increases the complexity and cost of transferring data between multiple disparate data systems. Also, there is a performance bottleneck with data throughput while managing multiple data copies in different locations. This paper aims to store all educational data in a central location and handle all structured and unstructured data without any performance bottlenecks. It is proposed to design an enterprise data lake solution for academic organizations using deep learning to predict the outcomes.

**Keywords** Artificial intelligence · Predictive analytics · Machine learning · Educational analytics · Deep learning · Data lake · Educational organizations

## 1 Introduction

Modern technologies such as artificial intelligence (AI), machine learning (ML), and deep learning (DL) are very critical terms to understand each other. They have their own definition, requirements, complexity, and limitations. They require

---

P. Kuppusamy ( ) · K. Suresh Joseph  
Pondicherry University, Puducherry 605014, India

access to various datasets for data analytics in many business organizations. These modern technologies also enable educational institutions to leverage large amounts of information from multiple sources—social media interactions, video conferencing, Internet of Things (IoT), sensor, biometric, learning management system (LMS), Internet search, and other sources to enable data-driven decisions across educational institutions. Many advanced technology organizations can leverage ML for many years. ML has complemented the growing work in predictive analytics by ensuring the results and recommendations with more accuracy and personalized.

DL is an ML method that uses neural networks for predictive analytics [22]. DL is broadly used for many AI applications. DL approaches can be used for various tasks, including object detection, synthetic data generation, user recommendation, and much more. DL delivers more accuracy on many AI tasks that require high computational complexity. Hence, DL applications have to be efficiently designed toward wide deployment for embedded applications such as mobile, IoT, and drones.

## ***1.1 Challenges for Educational Organizations***

The data silos could not handle unstructured data collected from social networks in traditional educational systems [17]. This problem increases the complexity and capital cost of transferring data between multiple disparate data systems. Besides, the proprietary data formats (SQL) prevent direct data access with other tools and increase lock-in risk. The non-SQL use cases require new copies of data for ML, DL, and predictive analytics. It leads to performance bottlenecks with data throughput and slows down data team agility and productivity. It can also increase the cost and management challenges for managing multiple data and security copies in a centralized location.

Despite the opportunity to improve the quality of education with analytics, both ML and DL face the following classical, big data challenges in educational organizations [12]:

- **Variety of data.** The delivery of learning content produces many multidimensional data from a variety of data sources. The educational administrators need to run queries across students, admission, payment, attendance, course registration, and time to build a holistic view of their experience. This process is compute-intensive for legacy analytics platforms. Learning data is unstructured (e.g., notes, documents, imaging, audio, video, animated gifs, etc.).
- **The velocity of data.** With a constant flow of data, LMS may need to be updated to fix coding errors. A transactional model must exist to allow for updates.
- **The volume of data.** Educational organizations generate a variety of data from various sources. It leads to big data.

The traditional data warehouses of educational organizations do not support unstructured data collected from sensors, IoT, and social networks. In traditional data warehouses, the query engines can struggle if the data volume is too large. This

leads to a simple ad-hoc analysis that can take hours or days. This is too long to wait when adjusting for researchers' needs in real-time. In many instances, educational organizations depend on IT experts to create new datasets precisely for ML/DL purposes. These instances slow down analytical progress.

The data flows from various sources in an educational environment and is preprocessed and cleansed in traditional data analytics [25]. The preprocessed data in the form of datasets propagates through several data stores to run diverse applications and then visualize the stakeholders. Therefore, the data scientists and analysts spend more time building and manage data analytics pipelines. The traditional analytical architectures often fail to keep up with the changes in the centralized data. This leads the business organizations to make decisions inefficiently. Therefore, data analytics architectures simplify data analytics pipelines and make data production teams more productive. A centralized data repository can help the data scientists to reduce time to predict the results. Fortunately, data lake is a central repository that can help to solve the above issue.

## 1.2 *The Solution*

Data lake technology [16, 29] provides a centralized transactional data store that supports fast multidimensional queries on diverse datasets along with rich data analytics capabilities. The centralized data store must support data scientists and analysts to run ad-hoc transformations or build predictive insights with learning approaches. Data lake provides a continuous data availability feature and increases overall business organizations' overall speed, including educational institutions. With the data lake approach, academic institutions can take business decisions quickly with quality.

Data lake technology handles all organized data, unorganized data, and semi-organized data in a central repository. It has data pipelines [27] to refine reliable data to the analytical applications gradually. It ensures that data is accessible across all tools and teams and reduces lock-in risk. SQL and ML/DL technologies can work together on a data lake with data from a central data repository. They build once, access many times across use cases for a combined administration and self-service. It processes data in a faster manner for streaming analytics, data science exploration, and model training.

In the data analytics architecture, instead of loading data in one large batch, it might seek to load streaming learning management system (LMS) data to allow for real-time analytics. Instead of using a dashboard, the data scientist and analyst may advance to ML and DL use cases, such as training an ML and DL model that uses data from the dataset to predict students' progression.

The objective is to design a data lake for educational organizations for predicting the results using ML and DL approaches. Learning analytics [21] using the DL method, educational administrators can better understand each student's learning

level, style, preferences, and ability and then tailor each student's learning experience. This allows educational administrators and teachers to identify each student's preferences and make decisions most effectively.

## 2 Background Details

This section introduces the necessary information required to write this article, including ML, DL, predictive analytics, and data lake technology.

In the competitive world, educational organizations are developing educational applications that meet the market need [28]. Educational organizations generate vast massive data every day from educational applications, students, content developers, faculty, learning processes, sensors, and devices. The educational organizations could analyze this gigantic data to make critical operational decisions to improve student retention, assessments, learning outcomes, and the institution's goals and educational services. Recently, AI, ML, DL, and learning analytics technologies have been used to analyze the above massive data to make better decisions.

### 2.1 Artificial Intelligence

AI is a concept that envisioned machines having a trait somewhat similar to human intelligence. Most devices either required a manual operator to perform the designated task or input instructions. AI aims at giving the machines the capability to think and act for themselves. It makes them better at performing tasks autonomously. ML and DL assist AI perceptions.

### 2.2 Machine Learning

Machine learning (ML) algorithms can assist machines in learning patterns from the datasets and predict predictions [23]. It is defined as "*the study of computer programs that leverage algorithms and statistical models to learn through inference and patterns without being explicitly programmed.*" Many different real-life use cases of ML—voice assistants, dynamic pricing, e-mail filtering, product recommendations, process automation, fraud detection, etc., are widely used by business organizations.

In general, ML tasks are classified as classification and regression. In the classification task, the predictive models are trained to organize data into different classes, whereas, in the regression task, models are built to expect nonstop data. A classification model in ML can be created with a set of steps. It includes import data, explore

data, training and testing data, building ML mode, testing model, evaluation model, and visualizing data.

**Life cycle.** ML life cycle obtains the power of large volumes and variety of data, abundant computing, and open-source ML tools to build intelligent applications [8]. ML life cycle has data acquisition, data preparation, model, and model deployment steps.

- i. The data acquisition and preparation step make the input data complete and of high quality.
- ii. The model includes training, testing, and selection with the highest prediction accuracy.
- iii. The model deployment step provides application development, inferencing model monitoring, and the management step to measure business performance and address potential production data drift.

**Challenges.** It leads that ML workloads become gradually more predominant. Hence, there are some significant challenges [15] in the scalability and deployability of ML workloads. The workload challenges [15] are listed below.

- i. *Data Collection.* The more data into ML workloads can give better results. Hence, data collection is a tedious job for many organizations.
- ii. *Data Exploration.* After data collection, data exploration requires a data catalog in the central repository.
- iii. *Data Experimental.* ML workloads take multiple tests to check the ML model's status. Hence, a disposable infrastructure is needed for ML.
- iv. *Data Flexibility.* During experimentation, the decision-makers can quickly install different tools, frameworks, and many new technologies.
- v. *Open Data.* Keeping the data in the open data format to support most open-source applications.
- vi. *Data Store.* Data store, in their original form, becomes complex.

Data scientists ensure that the selected ML model provides the highest prediction accuracy. Some of the critical challenges faced by the data scientists are as follows:

- i. Select and deploy the right ML tools (e.g., Apache Spark, TensorFlow, PyTorch, etc.).
- ii. Train and retrain the ML model that provides the highest prediction accuracy.
- iii. Hardware acceleration leads to slow execution of ML modeling.
- iv. Dependency on IT operations to manage ML infrastructure
- v. Collaborate with software developers to ensure input data hygiene for a successful ML model.

Having all the data in the same location leads to find the data easily. Today, many business organizations have many data repositories scattered on premise platforms, data centers, cloud, etc. This leads to traditional ML algorithms taking much time (ranging from a few hours to a few days) to train. The scenario is completely reverse in the testing phase. These issue leads to apply an alternate solution called deep learning.

### 2.3 Deep Learning

Deep learning (DL), also known as deep neural learning or deep neural network, is an AI process in data processing and forming patterns unsupervised from data (i.e., unstructured or unlabeled) for use in decision making [14]. DL algorithms use neural network algorithms to high-level model concepts in educational content (i.e., massive data) requirements [19]. DL denotes multiple transformation tiers and data representation levels between the inputs and outputs [3]. The DL has input, hidden, and output modules. The DL algorithm takes the dataset from the input module, assign labels, and sends it to the output module. A set of hidden module are placed between these two modules for computing complex functions that are more specific.

The challenges of DL approach are the massive data and the high-end computing power. Big data and graphics processing units (GPUs) facilitate DL to train larger and deeper models. The feature engineering process [3], a complicated and time-consuming process in traditional ML, is avoided by DL that performs *feature-learning* technique for the task automatically.

**Deep Learning Architectures.** DL architectures have transformed the analytical process for massive data among the deployment of sensory networks. They have been applied to various AI-enabled applications [11] performing supervised and unsupervised algorithms with diverse datasets, i.e., videos, images, documents, text, audios, etc.). The DL architectures [2, 31–33] include autoencoder, convolutional neural networks (CNN), feedforward neural networks (FNN), long short-term memory (LSTM), bidirectional LSTM (BLSTM), multilayer perceptron (MLP), recurrent neural networks (RNN), and variants (VGG16). The baseline methods are decision tree (DT), K-nearest neighbors (K-NN), linear regression (LinReg), logistic regression (LogReg), Naïve Bayes (NB), random forest (RF), singular value decomposition (SVD), support vector machine (SVM), support vector regression (SVR), etc.

The deep learning framework [13] with measurement tools and collaboration process enables educational systems to shift practice.

### 2.4 ML and DL in Education

Analytics methods are used in educational institutions to add visibility to students' academic progress and proactively react to their development. Analytics can be **descriptive, predictive, and prescriptive analytics**. ML and DL methods can be applied in educational organizations to enhance students' campus experience [4].

ML gives each student a personalized educational experience in the learning environment. The ML applications in the education field [7] are adaptive learning, learning analytics, predictive analytics, personalized learning, evaluating assessments, etc. Also, DL provides students with the advanced skills necessary according to their interests. DL recommends [26] teaching strategies that have long been considered good practice.

Both ML and DL can predict [3] student performance, detect undesirable student behaviors, their profile, their groups, social network analysis, provide reports, create alerts for stakeholders, plan, schedule, create courseware, develop concept maps, generate recommendations, adaptive systems, and evaluate scientific inquiry, etc. Most educational organizations perform the above tasks for analytics with the data store, secure data, reliable data, and data catalog.

Most educational organizations have become vital to harness the generated data to achieve better decision making and productivity [1]. They accumulate and accomplish all the generated data in a central repository (also called a principal data management architecture), more generally denoted as a data lake.

## 2.5 Data Lake Technology

Data lakes often consolidate educational institutions' data in a central location without a schema or structure on its upfront [5]. Data in all stages can be stored in a data lake as raw data can be consumed, stored in a structured format and tabular data sources, and intermediate data tables generated in refining raw data. Data lakes can process all data formats and types (e.g., audio, documents, images, text, and video) and run different analytics in the form of a dashboard, graphical representation, real-time analytics, predictive analytics, guide suggestions/recommendations and better decisions.

The definition [10] is that "*A data lake is a storage repository that holds a vast amount of raw data in its native format, including structured, semi-structured, and unstructured data. The data structure and requirements are not defined until the data is needed.*" The other definition [20] is that "*A data lake is a collection of storage instances of various data assets additional to the originating data sources; stored in a near-exact, copy of the source format.*" Data lakes can be classified as Hadoop-based and relational. Data lake stores and caters the data to multiple stakeholders for data consumption.

Data lakes allow business organizations to import any amount of data from sensor devices, biometric devices, LMS, blogs, mobile apps, log data, social media interaction, and ERP software and store them in non-relational and relational data. It allows stakeholders to access data using analytic tools with ML and DL workloads and make decisions faster. Data lakes break down the data silos generated by educational organizations, centralize, consolidate the organization's entire batch, and stream data for analytics. A data lake architecture opens up a wide range of new use cases for enterprise analytics with BI, ML, and DL algorithms.

The technical benefits are scalability, flexibility (as structured, semi-structured, or unstructured), availability, metadata, and quality. Data lakes empower advanced analytics, prediction, and ML and DL approaches. Some of the drawbacks are immature governance, user skills, integration, security, and system landscape. The challenges are the reliability of data and query performance. The traditional query

engines' performance has traditionally become slower when the size of the information increases in the data lake. This leads to some bottlenecks that include metadata management, improper data partitioning, and others.

Data lake can be used to consolidate data across all the stages of the pipeline in one place. The effective way of simplifying data pipeline design can be done as follows:

- It eliminates data movement in an analytics environment.
- It simplifies data governance in analytic data applications.
- It accelerates analytics sandbox productivity.
- It promotes the flexibility of analytics platforms.

Building a dashboard allows us to identify conditions across a population of educational data. LMS dataset may be made available through modern data analytics platforms such as Databricks datasets using AWS and Azure. This dataset may be stored in a structured or unstructured file format (e.g., CSV). It will load the CSV files before masking protected educational information and joining the tables to get the data representation it needs for our downstream query. Once the data has been refined, it will build a dashboard that allows us to explore and compute standard health statistics on the dataset interactively.

### 3 Literature Review and Methodology

To build a data lake, educational organizations have to assess data strategy, infrastructure, and workflows. Both ML and DL focus on personalized learning and pleasurable experience. Hence, the literature review may include ML, DL, and learning analytics for educational organizations.

The author reviewed [35] the factors that influence DL and discussed how environmental educators could encourage students to use DL strategies. This has identified [13] three waves of development of deep learning. The scholarly work provided [6] DL approaches to predict educational performance aspects. For huge data, big data and AI [34] helped HEIs understand student backgrounds more precisely. HEIs use the big data approach to improve student performance, teachers' effectiveness, reducing administrative workload, and increase flexibility.

The data lake provides a framework with various enterprise data strategies [40] that leads the organization to grow its data infrastructure [37]. Constructing a data lake architecture [38] was critical for laying down a strong data foundation. Data lake technology [36] might be combined with traditional data warehousing to provide great flexibility for faster data discovery and analysis. DL could be explored [9] using educational data mining and learning analytics datasets. Data lake [39] leverages to store the necessary data for predictive analytics and learning workloads.

A proper data lake architecture enables the ability to

- Allow the users to transform raw data into structured data for analytics with low latency. Raw data can be retained indefinitely for future use.
- Eliminate problems with data silos.
- Collect all data types and retains indefinitely for batch and streaming, documents, video, image, binary files, and more.
- Provide a landing zone for new data and always up to date.
- Enable end users with entirely different skills, tools, and languages to perform various analytics tasks all at once.

**Methodology.** The methodology includes a literature survey of articles in the last few years with a systematic approach. The literature search had full articles, research publications, research reports and importance to the subject, i.e., “data lake” (i.e., “model” and “architecture”) and “educational” (i.e., “educational organizations” or “educational institutions”). This was done in the following steps.

- i. Step 1. It searched the research articles with combinations above keywords. They were scanned for keywords relevant to “data lake,” “model,” “architecture” to “educational organizations.”
- ii. Step 2. Then, it selected and extracted the material related to the title.
- iii. Step 3. The extracted data was then analyzed and produced in *Sect. 4*.

## 4 Data Lake Architecture

The data lake design takes special considerations and provides advanced capabilities that include elasticity, automated recovery, availability, and analytical platform.

### 4.1 Requirements

The educational organizations handle a range of data/information that includes structured, unstructured, and semi-structured. The critical requirements considered to design a data lake architecture are durability, metadata, scalability, separate storage from computing, and supporting various data.

- *Data Types.* It is capable of storing all kinds of data in a single data repository.
- *Durability.* It allows exceptional data robustness to high-availability designs.
- *Metadata.* It enforces the requirement of metadata creation.
- *Scalability.* It must be significantly scalable without running into fixed limits.
- *Storage.* It decouples storage from computing for enabling independent scaling.

**Key Considerations.** The key considerations to design a data lake are the metadata requirement and metadata creation. The data lake design [30] may use a framework for various operations with metadata access in the central repository. The proposed architecture must be tied with the leading use cases being run on the analytical

platform. The key design factors that used data lake design are operational aspects, scalability, use cases, and performance.

- ***Operational Aspects.*** They force to ensure that the system is maintainable by the educational organizations.
- ***Scalability.*** As educational institutions generate massive data, the data lake is able to scale without replacement.
- ***Use Case.*** It allows appropriate prioritization of different analysis engines and data integration.

In addition to the above factors, the data lake architecture may consider the advanced capabilities, automation, data access and retrieval, and security. The ***advanced capabilities*** can quickly enable the teams to stimulate new analysis and reports. An automated data lake function can lower the operational burden for any updation and management for deployment and recovery. The ***data access and retrieval*** provide a multitude of tools for accessing and storing. The ***security factor*** meets the data lake threat levels and needs.

**Functional Areas.** Building a data lake incorporates technologies to provide access to diverse datasets in the following functional areas that are critical in the data lake deployments:

- ***Data Consumers.*** Data consumers access data and expose them to different application types. They are dashboards, e-Commerce, data science, business intelligence (BI), and mobile apps.
- ***Data Processing.*** The data processing can be automated and highly reliable manner. It includes streaming, rules/matching, ETL, and governance.
- ***Data Retrieval.*** The retrieval can be done in batch, analytical, in-memory, search/index, and OLTP.
- ***Data Storage.*** The data storage allows storing in non-relational data and historical copies of the information for later analysis.

**Analytical Model.** The ML system [18] focuses on either a supervised or an unsupervised learning system—the end users demand predictions. The ML components can be expected to provide some services, typically accessed via APIs. The end user and data analysts request predictions, and the model builder creates and updates models. The application (e.g., mobile app) used by the end user will benefit from the ML system. The model builder component is usually called either by a scheduler (update ML model) or via an API (data pipeline). The model builder component evaluates the ML models and for test predictions. These *model builders* components have APIs that can be directly accessed by the client using *front-end* tools. The end users can decide which model can be integrated on data for visualization. The ML system and the client application can access the dashboard regularly.

In general, ML performs iterative steps to force the analyzed dataset to training exact models for execution on unknown datasets. The data analysts will train the models when leveraging ML. They can then leverage in conjunction with various analytical tools (e.g., R, SAS, and Python) and cloud providers (e.g., AWS and

Google) that provide trained models and custom models for use against proprietary datasets.

The educational data lake enables flexibility through an analytical platform for the analysis of compound datasets. In general, the analytical workflow begins with the data process that includes data ingest, data cluster, data index and analyzes data within the data lake. These steps ensure that high-quality information is brought together to enable data scientists and analysts to examine the preprocessed data.

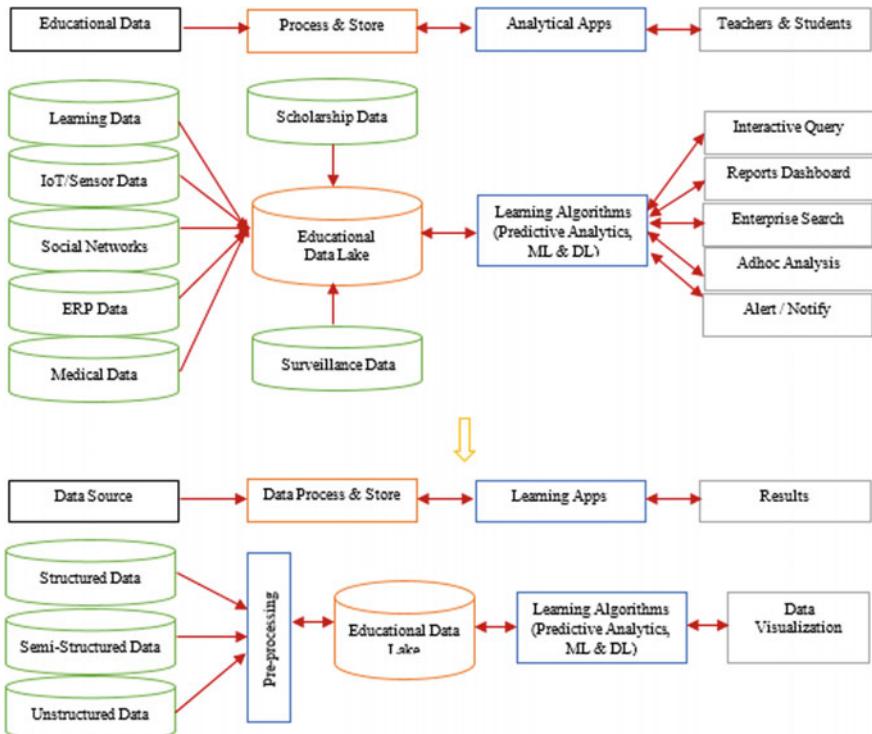
**Datasets.** Deep learning tasks need different educational datasets for data analysis. They are training and for evaluation datasets. For example, how students learn and how students interact. The features are student performance, student behaviors, grouping students, social network interaction, reports, alerts, planning and scheduling, courseware, concept maps, recommendation, adaptive, evaluation, inquiry, etc.

## 4.2 Data Lake Functional Model

Designing a data lake for a learning analytics system can integrate many complex technologies to provide consistent access to diverse datasets (e.g., documents, audio, video, gifs, etc.). This requires reliable connectivity to traditional systems. Figure 1 shows the recommended data lake functional model [24] too smart learning analytics. The data lake can seamlessly connect various methods to provide data movement in an automated and highly reliable manner. A data lake's functional model for deployment has *data processing*, *data storage*, and *data consumption*.

**Data Processing.** Various data processing methods are data streaming, rules, ETL, and governance. Data streaming analyzes and makes decisions on data in motion. Rules can execute pattern matching against data identification and duplication operation. ELT integrates data into traditional data platforms. Data governance ensures compliance and adherence to educational organization procedures. For example, the publish/subscribe system provides a seamless experience for sharing data among various processing tools.

**Data Store.** Data store, including data retrieval, enables the data analyst to request data in typical formats (e.g., APIs) from the data lake. They may be batch, analytical, in-memory, search, and OLTP. Batch processing provides high throughput and high latency for data analysis. Analytical processing is used for interactive workloads where the queries change frequently. In-memory processing supports very low latency queries that help low latency. Search (or index) is used to locate information and relationships quickly. These functions support transactional systems (e.g., OLTP) found in business operations. Data storage can be object store and long-term storage—an object store stores non-relational data and historical data for future analysis. Long-term storage is necessary for archiving data and is required to be accessible.



**Fig. 1** Data lake functional model to educational organizations

**Data Consume.** The data lake allows the data consumers and their applications to consume the data through various interfaces to access and expose them to different applications. These applications are dashboards, e-Commerce, predictive analytics, business intelligence (BI), and mobile apps. Dashboards update the data with reporting on specific metrics and changes over time. The e-Commerce applications analyze the transactional data. Predictive analytics allow consumers to model, test theories, and generate new reports. BI tools do data analysis with complex datasets and relationships. API offers the ability to quickly identify patterns in data without any additional servers or services. The mobile apps can be designed for quick access with less response time with more accuracy of data.

#### 4.3 Data Lake Architecture

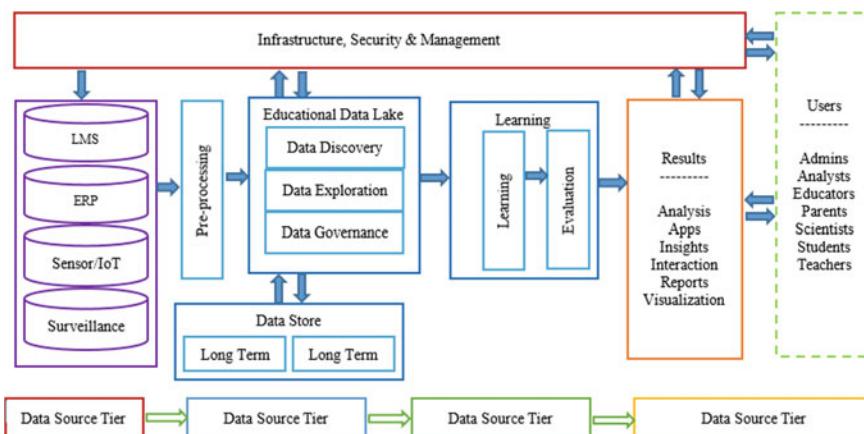
The proposed architecture has the stakeholders of data scientists, data analysts, educators, teachers, parents, and recruiters, and they can have access to educational data (immutable data) for discovering and mining the data for high-value professional

insights. The proposed architecture stores the information in a raw format. The data analysis tool can impose educational business applications to the analysis context on the diverse dataset. The immutable data in the data lake creates multiple processing tiers to enable educational applications, i.e., structured data store. The preprocessing tools process the collected data into labeled, unlabeled and semi-labeled, which could be used for various value-added and structured data. Examples of data lake tools and frameworks are Presto (Facebook), Microsoft Azure Data lake Gen2, Amazon S3, Delta Lake, Databricks, Apache Sparke, etc. These tools run queries on diverse datasets and predict the analytical results.

As shown in Fig. 2, which is derived from Fig. 1, the proposed architecture has a data intake layer, data processing and storage tier, data analytics tier, and data visualization tier.

**Data Source Tier (DST).** In educational applications, the data sources are LMS, ERP, sensors, IoT devices, surveillance, medical, social networks, library, payment gateway, Internet usage, etc. In this layer, the educational data source is encountered and subsequently sent to the other tiers for various operations. The source data can be of any size, format, structured, semi-structured, and unstructured. Examples of our data are database, documents, audio and video, animated GIFs, log file, survey data, Web browser data, or any other text file. The data sources are relational databases, file export, APIs, ZIP files, and big data sources.

**Data Processing Tier (DPT).** This layer accepts the raw data from the data intake tier. The data processing tier has two major modules—*data processing* and *storing*. Data processing receives raw data from DST and processes it into structured data that supports data analytics applications. DPT has a large set of metadata fields. The DPT parses and normalizes into metadata for automated analytics. The DPT has a data indexer that enables the search facility in a scalable manner, i.e., elastic search. The elastic search enables distributed, full-text, unstructured, and access to



**Fig. 2** Data lake architecture to educational organizations

the structured data. It allows a copy of every raw message to an archive that can be recalled, re-processed, and added back into future analysis indexing. Here, the data is cleansed, transformed, and structured. It provides opportunities for educational administrators to use data as required by the educational process.

The data storing module stores the processed data in the data lake system. It deals with the data storage, retrieval, and delivery of the data to the required applications. This module responds to queries. Different applications impose different requirements on this module. For example, one application may require real-time response and another may require historical data access. Hence, it requires huge storage of all the generated data and real-time access to designated data. It may use the *schema-on-read* technique to eliminate the necessity of heavy ETL processing of data.

Various data processing modules are transformation, abstraction, data cleansing, federation, data services, and de-duplications with encryption.

**Data Analytics Tier (DAT).** The DAT performs data-intensive computation, management, and visualization. The DAT depends on an analytics system, **data modeling**, and the data pipeline, which integrates various **data sources**. The DAT interacts with end users and allows them to visualize the data as reports and dashboards. Data scientists, business analysts, and other users can build their analytical models for their prediction. The BI and data visualization components make data easy to understand and manipulate. BI provides visualizations and tools that allow users to make data-driven business decisions. Examples of BI tools are Tableau, Looker, and Microsoft Power.

The HEIs may use student and institutional information for predictive analytics are alert system, recommender system, adaptive technology and enrollment management.

**Data Visualization Tier (DVT).** The data visualization (or presentation) tier allows the end users to interact with the data. It consists of various queries and reporting tools. Query tools extract data and send it to the end user through the graphical representation. Reporting tools are used to obtain data and applied to gather several kinds of information. This layer maintains and views metadata information, system operations, and performance.

The proposed architecture ensures maximum availability, data security, data quality, reliability, and scalability of the data lake platform. For *data quality*, the proposed architecture provides a single data repository with diverse datasets. The best practices to measure data quality are schema on reading, immutable data, identifying data, source of record, relationship mapping, and metadata catalog. For *data security*, the proposed architecture allows educational users to access the data in different methods. The best practices for data integration are **security context**, **identities**, and data policies.

The separation of DST, DPT, DAT, and DVT enables its security and analytics platform to deliver integrated security management at an enterprise scale.

## 5 Discussion and Results

Education technology focused on leveraging emerging technology to improve educational processes and outcomes for institutions. An analytics-driven educational institute can be more prosperous and satisfied the stakeholders. ML and DL approaches can predict students' performance and behaviors, providing recommendations and suggestions with automated evaluation techniques. They offer many theoretical models—teaching, learning, evaluation, etc., developed to benefit educational stakeholders.

Data lake technology combines multiple data sources in one place consistently and makes them available to build business solutions and processes in a single system. It allows for fast and efficient decision making based on updated and editable data of various types. This functionality of data lake is processing advanced analytics. The data lake combines filtered and secured data into the data hubs and downstream applications. The updated data can be available for stakeholders in one toolset. The updated data decides fast and swifts without technological bottlenecks.

The prediction in educational platforms has recently gained more consideration in sensing the stakeholders' unwanted behaviors in the learning environment. Educational institutions can apply an agile approach to their analytical design to optimize data storage and access. They can bring analytics-driven insights to market faster that significantly reduces the cost and complexity of managing their data architecture.

Predictive analytics adds visibility to educational improvement on behalf of students who enable to react to their development. However, analyzing student personal data requires cautious consideration to integrity and privacy.

## 6 Conclusion and Future Enhancements

Educational organizations are looking for ML, DL, and predictive analytics technologies for analyzing effectively at targeting prospects, supporting students, building effective products, and responding to market needs. The educational organization must develop an infrastructure to store data, execute analytical workloads, and protect data from an unexpected change to leverage these technologies. A thorough study of ML and DL with data lake techniques was provided in this paper. The DL models and DL architecture can serve for future analytical applications in educational data.

Educational organizations understand the necessity of agile methodologies in the context of data management. Educational organizations take the lead on screening potential technology and approaches to building data lakes. Data lakes must learn to balance the traditional data oversight against the need for flexibility as data is collected and used by them.

A data lake provides a flexible platform for data storage and processing, scalability, and high-availability levels to educational organizations. These use cases effectively grow capability as the students increase in educational organization. Data lakes can

easily add new capabilities by leveraging deep expertise in scalability and security. A proper recommendation system will be a challenge for the future education system. Here, the challenge is the impossibility of manually structuring a large amount of data from various. Proposing a proper recommendation system will be future work.

## References

1. A. Pan, Unlocking the potential of machine learning in a Data Lake (2019). <https://www.datavirtualizationblog.com/unlocking-the-potential-of-machine-learning-in-a-data-lake/>
2. A. Varangaonkar, Top five deep learning architectures (2018)
3. A. Hernández-Blanco, B. Herrera-Flores, D. Tomás, B. Navarro-Colorado, A systematic review of deep learning approaches to educational data mining, complexity (2019). <https://doi.org/10.1155/2019/1306039>
4. B. Dik, Higher education predictive analytics: the future in the digital age (2019)
5. Capgemini, The technology of the business data lake, consulting technology outsourcing (2020)
6. C. Perrotta, N. Selwyn, Deep learning goes to school: toward a relational understanding of AI in education. *Learn. Media Technol.* **45**(3), 251–269 (2020). <https://doi.org/10.1080/17439884.2020.1686017>
7. DataFlair, How is machine learning enhancing the future of education? (2020) <https://data-flair.training/blogs/machine-learning-in-education/>
8. DataRoot, Machine learning life cycle (2020). <https://www.datarobot.com/wiki/machine-learning-life-cycle/>
9. T. Doleck, D.J. Lemay, R.B. Basnet et al., Predictive analytics in education: a comparison of deep learning frameworks. *Educ. Inf. Technol.* **25**, 1951–1963 (2020). <https://doi.org/10.1007/s10639-019-10068-4>
10. Enterprise, Enterprise data lake architecture: what to consider when designing (2020)
11. F. Shaikh, Ten advanced deep learning architectures data scientists should know! (2017)
12. F.A. Nothaft, M. Ortega, A. Kermany, Building a modern clinical health data lake with delta lake (2020)
13. M. Fullan, J. Quinn, M. Drummy, M. Gardner, Education Reimagined; The Future of Learning. A collaborative position paper between New Pedagogies for Deep Learning and Microsoft Education (2020). <http://aka.ms/HybridLearningPaper>
14. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*. (MIT Press, 2016). <http://www.deeplearningbook.org>
15. H. Sarmah, Can data lakes solve machine learning workload challenges? (2019) <https://analyticsindiamag.com/can-data-lakes-solve-machine-learning-workload-challenges/>
16. H. Vatter, Reality and misconceptions about big data analytics, data lakes, and the future of AI (2019)
17. J. Patel, Overcoming data silos through big data integration. *Int. J. Comput. Sci. Technol.* **3**(1), 1–6 (2019). <https://doi.org/10.5121/IJDMS.2019.1301>
18. L. Dorard, The architecture of a real-world machine learning system (2019)
19. M.A. Peters, Deep learning, education, and the final stage of automation. *Educ. Philos. Theor.* **50**(6–7), 549–553 (2018). <https://doi.org/10.1080/00131857.2017.1348928>
20. M. Schedlbauer, Data Lakes—how to enable advanced analytics and machine learning? (2019)
21. N. Hobart, How learning analytics can make your teaching more effective (2018). <https://www.classmate.com/blog/learning-analytics-make-teaching-more-effective/>
22. O.I. Abiodun, A. Jantan, et al., *State-of-the-art in Artificial Neural Network Applications: A Survey*. Elsevier **4**(11), 1–41 (2018)
23. O. Genç, Notes on artificial intelligence, machine learning, and deep learning for curious people (2019). <https://towardsdatascience.com>

24. K. Palanivel, K. Suresh Joseph, Data lake model to modern educational organizations. *Int. Res. J. Eng. Technol.* **7**(7), 268–276 (2020)
25. R. Dwivedi, Step-by-step building block for machine learning models (2020)
26. S. Briggs, Deeper learning: what is it and why is it so effective? (2015). <https://www.opencolleges.edu.au/informed/features/deep-learning/>
27. S. Rao, How to build a data pipeline for autonomous driving (2020)
28. S. Digumarti, Predictive analytics has a future in education (2013). <https://analyticsindiamag.com/predictive-analytics-has-a-future-in-education/>
29. S. Tiao, Machine learning and modern data lake (2018)
30. S. Bhattacharya, N. Matthews, Enterprise data lake architecture: what to consider when designing (2020). <https://www.cloudtp.com/doppler/how-to-guide-architecture-patterns-to-consider-when-designing-an-enterprise-data-lake/>
31. G. Manogaran, G. Srivastava, B.A. Muthu, S. Baskar, P.M. Shakeel, C. Hsu, P.M. Kumar, A response-aware traffic offloading scheme using regression machine learning for user-centric large-scale Internet of Things. *IEEE Internet Things J.* 1–1 (2020). <https://doi.org/10.1109/jiot.2020.3022322>
32. T.N. Nguyen, B.-H. Liu, S.-Y. Wang, On new approaches of maximum weighted target coverage and sensor connectivity: hardness and approximation. *IEEE Trans. Netw. Sci. Eng.* **7**(3), 1736–1751 (2020). <https://doi.org/10.1109/TNSE.2019.2952369>
33. M. Tim Jones, Deep learning architectures, the rise of artificial intelligence (2017)
34. S. Tsai, C. Chen, Y. Shiao, et al., Precision education with statistical learning and deep learning: a case study in Taiwan. *Int. J. Educ. Technol. High Educ.* **17**, 12 (2020). <https://doi.org/10.1186/s41239-020-00186-2>
35. K. Warburton, Deep learning and education for sustainability. *Int. J. Sustain. High. Educ.* **4**(1), 44–56 (2003). <https://doi.org/10.1108/14676370310455332>
36. D. Kieffer, Building an enterprise analytics and BI practice in higher education (2019)
37. E. Levy, Understanding data lakes and data lake platforms (2018). <https://www.upsolver.com/blog/understanding-data-lakes-and-data-lake-platforms>
38. J. Stephan, The charm of security-driven data lake architecture (2020)
39. J. Jablonski, Building a platform for machine learning and analytics (2019). <https://www.cloudtp.com/doppler/building-platform-machine-learning-analytics/>
40. T. Spicer, Data Lakes? Big myths about architecture, strategy, and analytics (2019)

# Events of Interest Extraction from Forensic Timeline Using Natural Language Processing (NLP)



Palash Dusane and G. Sujatha

**Abstract** Forensic timeline performs a critical role in the investigation of computer forensic incident. It provides a summary of events that happened in and around a specific security incident. The events found in log files can help the investigator to construct a forensic timeline. We can consider events of interest as abnormalities in the timeline. An investigator can manually identify interesting events in a forensic timeline by entering keywords that have a high possibility of occurrence, but this process can be rather time-consuming. Some systems work on predefined rules, but those do not provide any help for previously unseen log messages. There are many useful methods and tools for creating a forensic timeline, but nothing significant for detecting events of interest. This paper presents a methodology that will help an investigator in extracting interesting events from a timeline. We propose the use of natural language processing (NLP) for identifying the sentiments of log messages. After spotting a log message with negative sentiment, the examiner can study the neighboring events recorded in that forensic timeline. In this way, the time required for investigation will reduce. This paper also presents a comparative study of learning algorithms used for sentiment analysis on forensic logs.

**Keywords** Computer forensics investigation · Forensic timeline · Events of interest

## 1 Introduction

In a computer forensics investigation, one of the main challenges is the reconstruction and examination of the forensic timeline to understand digital evidence because of

---

P. Dusane ()

Department of Information Technology, SRM Institute of Science and Technology,  
Kattankulathur, Chennai, India  
e-mail: [pd4399@srmist.edu.in](mailto:pd4399@srmist.edu.in)

G. Sujatha

Department of Information Technology, SRM Institute of Science and Technology,  
Kattankulathur, Chennai, India  
e-mail: [sujathag@srmist.edu.in](mailto:sujathag@srmist.edu.in)

the enormous quantity and variety of log messages [1]. A log file is a valuable piece of information in an investigation for any forensic incident. It contains system actions, events and user activities. All events occurred in a system are logged with the necessary information, such as timestamp, name of the system, process name and ID, and detail description of an event [2]. Usually, the disk image of a compromised system is used by an investigator to build a forensic timeline after an attack. A forensic timeline will be able to give an investigator a summary of the events that happened at the time of a specific security incident [3]. We can consider all the interesting events as abnormalities in the timeline.

There are several tools available for automatic generation of a forensic timeline; one such prominent tool is Log2Timeline which is an open-source python implementation. Log2Timeline creates a forensic timeline, which is a combination of several events and log messages throughout a system [4]. It takes a disk image of the compromised system as an input and produces CSV file as output. A tool like Timeline2GUI can be further used to visualize the output of Log2Timeline. Timeline2GUI is a graphical tool that can read CSV files generated by Log2Timeline, and it supports analysis and appropriate highlighting of a timeline. Timeline2GUI tool is aimed to make parsing of the log files easier for the user. This tool is based on Excel, and its GUI is kept as simple as possible for the users [5].

Sentiment analysis has turned out to be an active research field in NLP from early 2000. It is popularly used to study people's sentiments, opinions and emotions toward entities such as events, services, products, issues and topics [6]. In this paper, we will apply sentiment analysis to log messages from various system log files. Unlike product review or comments, log messages are shorter, simple and straightforward; hence, document-level sentiment analysis is enough for this study. We can also use aspect-level sentiment analysis, which will allow obtaining further specific data from log records. This type of sentiment analysis will result in a complex process of model training [7, 8].

As log files are the principal source to create a forensic timeline, the messages in log files include both positive and negative sentiment [8]. The fundamental intention behind this study is to train models to predict the sentiment of the log messages. In this way, the investigator will be able to identify positive and negative sentiment messages in log files. The log messages with negative sentiment will be considered as events of interest; then, all these messages will be further investigated for the incident. An example of a log message from a public dataset containing negative sentiment such as "failure" is shown in Fig. 1.

Following are the main contributions of this research paper:

```
combo sshd(pam_unix)[1650g]: authentication failure; logname= uid=0 euid=0 tty=nodevssh ruser= rhost=61.153.202.254 user=root
    ↗ Negative Sentiment
```

**Fig. 1** Log message with negative sentiment

1. We present a methodology that will assist the investigator in extracting abnormalities from a timeline. It will help them to investigate a security incident in less time.
2. This work proposes using document-level sentiment analysis on log messages. The message with negative sentiment is deemed as an exception in a log file.
3. Keeping NLP in mind, we also propose a semi-automated investigation process which is a combination of existing forensic tools and trained sentiment model.
4. We have performed various experiments to evaluate the sentiment analysis performance on public forensic datasets. Detailed comparative study of machine learning algorithms used for sentiment analysis of forensic timeline is presented in this paper.
5. The proposed method of document-level sentiment analysis with existing forensic tools produces high performance on the log files. Most of the machine learning algorithms used in this experiment achieve an accuracy of 99% and more for identifying negative messages from forensic log datasets.

The rest of the research paper is arranged as follows: We provide the related research on sentiment analysis, anomaly detection and timeline analysis in Sect. 2. Section 3 covers the details of our proposed methodology. Section 4 presents the comparative result and analysis of the experiments. Section 5 concludes the paper, and Sect. 6 provides its future work.

## 2 Related Work

### 2.1 Sentiment Analysis

Sentiment analysis is one of the popular applications of Natural Language Processing (NLP). It is primarily applied to classify texts like product reviews, comments, etc. Recent work by Studiawan et al. [8] showed how sentiment analysis with deep learning is effective to identify negative messages from the forensic timeline. They proposed an aspect-based sentiment analysis method to identify the events of interest in a forensic timeline. Guzman et al. [9] presented sentiment analysis as a mechanism for identifying emotions shown in commit comments. They analyzed the relationship between these emotions with factors like time and day of the week, programming language, etc. They also found out that commit comments written on Mondays lead to more negative emotion. Ding et al. [10] developed an entity-level sentiment analysis tool known as SentiSW to identify sentiment and entity tuples of project issue comments. They built a dataset of 3000 issue comments from 10 popular GitHub projects. Singh et al. [11] worked on four machine learning algorithms for optimization of sentiment analysis. They found out that Naive Bayes is the fastest in terms of learning, whereas OneR is more accurate than others. A framework for multi-class sentiment analysis is introduced by Liu et al. [12], where they select important features using feature selection algorithm and then train the classifier

using a machine learning algorithm. They used popular feature selection algorithms like CHI Statistics, Gain Ratio, Document Frequency and Information Gain, and machine learning algorithms like Decision Tree, Naive Bayes, etc. According to their study, an improved multi-class sentiment classifier can be developed in future. Rathi et al. [13] combined SVM and Decision Tree to produce finer classification outcomes of Tweets. Their experimental results prove that the proposed approach of using multiple classifiers is more accurate in contrast to individual classifiers.

## 2.2 *Anomaly Detection*

An unsupervised framework is proposed by Vaarandi et al. [14] for anomaly detection in Syslog log files. They have created a baseline using regularly occurring message patterns. The messages related to system failures and errors occur infrequently as per these assumptions. Bertero et al. [15] proposed a linguistic approach to detect anomaly in log files. They represented log files as a set of features and then processed them using a machine learning algorithm. They were able to attain good accuracy on Syslog files. A combined technique for finding anomalies in system log files using K-prototype clustering and k-NN classification algorithm is presented by Liu et al. [16]. They used 10 features from session information to identify user behavior. Wang et al. [17] combined NLP methods, such as Word2vec and TF-IDF, DL algorithms to detect anomalies in log files.

## 2.3 *Forensic Timeline Analysis*

A unique method based on four different levels of abstraction is introduced by Bhandari et al. [1]. It follows a unique structured which gives information related to an activity performed by the user. They use data produced by Log2timeline tool. Results showed that the abstraction-based approach can reconstruct the timeline efficiently. Chabot et al. [18] proposed a new methodology that can help examiners throughout the entire process of timeline creation and the analysis concerning the log messages. They presented a formalized knowledge model which helps to correlate events. The use of pattern matching to automatically reconstruct forensic timeline is shown by Hargreaves et al. [19]. The proposed approach automatically reconstructs high-level activities from a set of low-level activities from the timeline. Studiawan et al. [20] proposed a tool nerlogparser for automatic event log parsing. They used bidirectional LSTM to recognize a series of fields in a log record. This tool allows investigators to prepare their own domain-specific log files.

There are some popular tools which are used in computer forensics for timeline creation, visualization and analysis. One such tool is Encase, which offers a built-in timeline feature used for analysis. Forensic Tool Kit (FTK) is another commercial tool used for timeline construction [21]. Log2Timeline is a command-line tool to

create a timeline from a mounted device like a forensic disk image. It is known for its efficiency and accuracy [5, 21]. Timeline2GUI is a visualization tool that systematically displays a timeline generated by Log2Timeline, and it also supports analysis of that timeline. [5]. CyberForensic TimeLab is a visualization tool created by Olsson et al. [22–24] to display indexed forensic evidence.

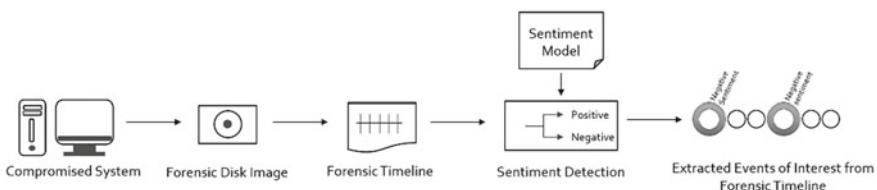
### 3 Methodology

We present a methodology by which a computer forensic investigator can quickly extract events of interest from a compromised system. An overview of the proposed investigation process is given in Fig. 2. This process is not an alternative for the actual process which an investigator follows; it is just an addition to enhance its speed and accuracy.

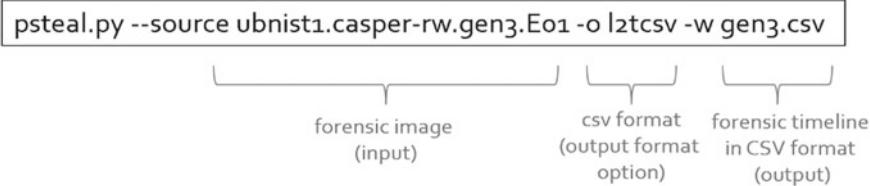
This semi-automated process is a combination of existing forensic tools and proposed Natural Language Processing (NLP) methodology. It includes the following four steps.

#### 3.1 Forensic Image Creation

In forensic image creation, data from a hard drive of a compromised system is copied to a file placed on a different drive. Disk imaging performs a sector-by-sector copy of the hard drive under the investigation. The output usually has an E01 (encase image file format) or DD (disk image file format) format [25]. There is a tool called SafeBack which can perform sector-by-sector copy with CRC checksum verification. DD is a command-line utility for Unix like operating systems which can perform sector-by-sector and file-by-file copy with MD5 checksum verification. We also have a tool known as Encase Imager by Encase, which is used to create forensic images of a hard disk used in investigations. The SnapBack DatArrest is a popular tool to do a fast and exact copy of hard disks [26].



**Fig. 2** Proposed investigation process



**Fig. 3** Psteal command

### 3.2 Forensic Timeline Generation

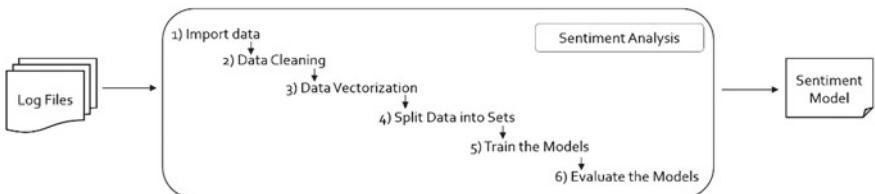
A forensic timeline provides an investigator with an overview of events that took place in and around an incident. A popular open-source tool called as log2timeline can be used to create a forensic timeline [4]. It is built to parse different log files and artifacts from a forensic disk image. It automatically produces a super timeline to help investigators in their timeline investigation. The easiest and quickest way to create a timeline is by using psteal tool. psteal is a command-line tool that merges the functionality of log2timeline and psort [27]. The command shown in Fig. 3 will generate a CSV file which will include all the events from a forensic image.

### 3.3 Sentiment Detection

In this step, the investigator will automatically detect the sentiment of log messages by using pre-trained sentiment model. A log message will either be a positive message or a negative message. We performed sentiment analysis on log files using machine learning algorithms to generate trained sentiment model. An overview of the training phase is shown in Fig. 4.

We performed sentiment analysis in the following six steps:

- Importing Data:* The dataset that we are going to use for training is loaded into a system for processing.



**Fig. 4** Training phase

- b. *Data Cleaning*: We need to clean the log messages before they can be used for training the model. In this step, all unnecessary characters and multiple spaces are removed from log messages.
- c. *Data Vectorization*: Statistical learning algorithm works only with numbers. So first, we must convert text into numeric form.
- d. *Splitting Dataset into Training and Testing sets*: In this step, the dataset is divided into a training dataset and testing dataset. The training set will be used to train the algorithm, while the testing set will be used to evaluate the performance of the machine learning model.
- e. *Training the Model*: In this step, we can use a machine learning algorithm to learn from the training data.
- f. *Evaluating the Model*: We can predict the test data from the trained model and evaluate its performance using metrics like accuracy, f1 score, etc.

### 3.4 Events of Interest Extraction and Analysis

In this step, the investigator will consider log messages with a negative sentiment as events of interest. Investigator will further study the surrounding events of the negative log messages. In this way, the investigator will directly examine the abnormalities present in a forensic timeline without going through numerous log messages. Hence, the time required for investigation will reduce, and the process of finding events of interest will be more accurate.

## 4 Experimental Results

### 4.1 Datasets

We are using two different datasets for conducting these experiments. Details about these datasets are shown in Table 1. The first dataset is from Loghub [28], where they maintain several system logs files. Loghub dataset is freely accessible for research. This unlabeled dataset was collected on Linux server from /var/log/messages for a period of 263.9 days. The second dataset is from Digital Corpora [29], which

**Table 1** Dataset details

Dataset	Size (number of records)	Log parameter used for analysis	Positive logs	Negative logs	Positivity (%)
Loghub	25,547	Log description	10,635	14,912	41.63
Casper	11,086	Log description	9794	1292	88.34

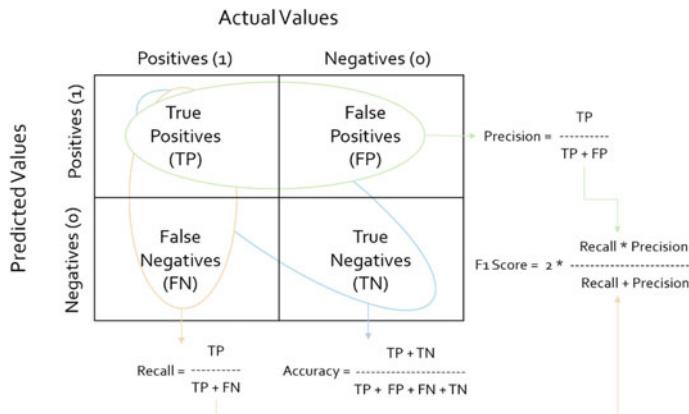
a website used in computer forensics education research. This dataset nps-2009-casper-rw is an ext3 file system from a bootable USB token that had an installation of Ubuntu 8.10. There are 25,547 readable log messages in the first dataset and 11,086 readable log messages in the second dataset.

## 4.2 Experiment Setup

The system which we used for the experiment is an i7, 6-core CPU,  $\times$  64 based system. It has 32 GB of RAM, 19.9 GB GPU memory of NVIDIA Quadro T1000 and 15.9 GB GPU memory of Intel(R) UHD Graphics. We used Jupyter Notebook 6.0.3 with Python 3.8.3 for this experiment. We also used machine learning library Scikit-learn 0.23.1 [30] for performing operations like vectorization, splitting data into train-test data, fitting the model, evaluating the model and calculating performance metrics of model. We divided data into 80/20, 80% data for training sentiment model and 20% data for testing sentiment model.

## 4.3 Evaluation Metrics

We use a confusion matrix to evaluate the performance of sentiment models. Confusion matrix helps us to calculate metrics like precision, recall, f1 score and accuracy. Following Fig. 5 explains the evaluation metrics that we used in this experiment:



**Fig. 5** Confusion matrix

#### 4.4 Comparative Results

To study the performance of sentiment analysis on log messages, we used classifiers like Logistic Regression, Random Forest, Linear Support Vector Classifier, Naïve Bayes, Stochastic Gradient Descent and K-Nearest Neighbors. All these methods were used to predict sentiments of log messages from both Loghub and Casper dataset. The testing dataset was used to calculate metric values in experiment results. These techniques have been already used to detect sentiments in other datasets like Twitter comments, movie reviews and product reviews. We know that the log files are the main reference to build a forensic timeline. Therefore, in this study, we assess these techniques by applying them to Linux system log files.

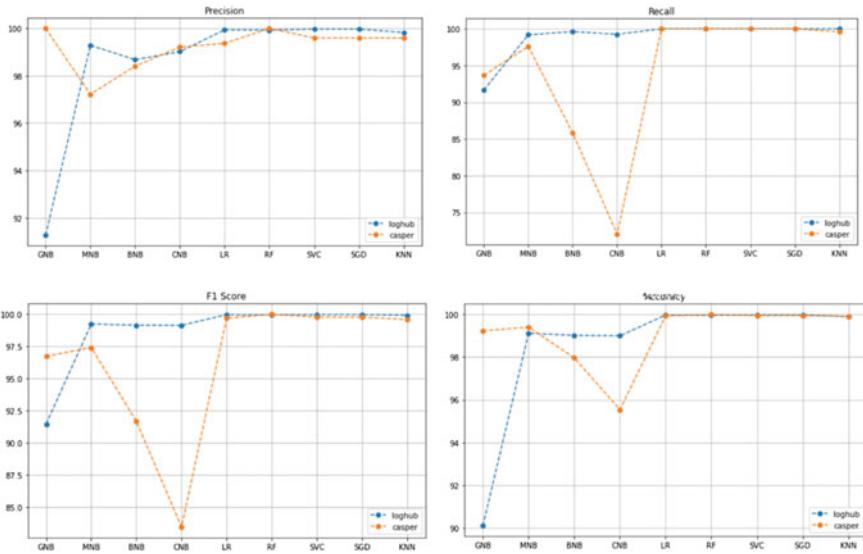
Table 2 presents the outcomes of sentiment analysis on Loghub dataset. We can observe that except Gaussian Naïve Bayes, all other classifiers achieve accuracy more than 99%. In the same way, Table 3 presents the outcome of sentiment analysis on Casper dataset. Here, we can see that besides Complement and Bernoulli Naïve

**Table 2** Loghub dataset result

Method	Types	Precision	Recall	F1	Accuracy
Naive Bayes	Gaussian	91.27	91.67	91.47	90.12
	Multinomial	99.29	99.19	99.24	99.12
	Bernoulli	98.68	99.62	99.15	99.02
	Complement	99.02	99.26	99.14	99.00
Logistic regression	–	99.94	100.00	99.97	99.97
Random forest	–	99.93	100.00	99.96	99.96
Linear support vector classifier (SVC)	–	99.97	100.00	99.98	99.98
Stochastic gradient descent (SGD)	–	99.97	100.00	99.98	99.98
K-nearest neighbors (KNN)	–	99.83	100.00	99.92	99.90

**Table 3** Casper dataset result

Method	Types	Precision	Recall	F1	Accuracy
Naive Bayes	Gaussian	100.00	93.68	96.74	99.23
	Multinomial	97.22	97.61	97.41	99.41
	Bernoulli	98.41	85.81	91.68	97.97
	Complement	99.21	72.05	83.47	95.54
Logistic regression	–	99.37	100.00	99.69	99.93
Random forest	–	100.00	100.00	100.00	100.00
Linear support vector classifier (SVC)	–	99.60	100.00	99.78	99.95
Stochastic gradient descent (SGD)	–	99.60	100.00	99.78	99.95
K-nearest neighbors (KNN)	–	99.59	99.59	99.59	99.91



**Fig. 6** Result comparison of loghub and casper dataset

Bayes, all other classifiers also achieve accuracy more than 99%. The comparison graphs of Loghub and Casper dataset are shown in Fig. 6.

## 5 Conclusion

The large volume and variety of log messages create many challenges for an investigator to investigate a forensic incident. In this paper, we resolved the problem of finding interesting events (i.e., abnormalities) from a computer forensic timeline by using sentiment analysis. This allows us to utilize NLP techniques to extract sentiment of log messages. We proposed a semi-automated investigation process which is a combination of existing forensic tools and pre-trained sentiment model. This process considers the log messages with a negative sentiment as abnormalities or anomalies in a forensic timeline. After spotting an event of interest, the examiner can study the neighboring events recorded in that forensic timeline. This semi-automated method is not proposed as an alternative to the existing forensic investigation process. It is expected that with the addition of this method into an existing process, and the speed and accuracy of the investigation will increase.

## 6 Future Work

In future, this method needs more development to derive detailed information about an event of interest. Sentiment analysis can also be combined with different types of anomaly detection methods to obtain a better result. A system or tool needs to be developed to convert this semi-automated process into a fully automated process.

## References

1. S. Bhandari, V. Jusas, An abstraction based approach for reconstruction of timeline in digital forensics. *Symmetry*. **12**, 104 (2020). <https://doi.org/10.3390/sym12010104>
2. L. Zeng, Y. Xiao, H. Chen, B. Sun, W. Han, Computer operating system logging and security issues: a survey. *Security Comm. Networks* **9**, 4804–4821 (2016). <https://doi.org/10.1002/sec.1677>
3. H. Studiawan, F. Sohel, C. Payne, A survey on forensic investigation of operating system logs. *Dig. Invest.* **29**, 1–20 (2019). ISSN 1742-2876, <https://doi.org/10.1016/j.dii.2019.02.005>
4. K. Gudjonsson, Mastering the super timeline with log2timeline, in *Information Security Reading Room* (SANS Institute, Bethesda, MD, USA, Tech. Rep., 2010)
5. M. Debinski, F. Breitinger, P. Mohan, Timeline2GUI: A Log2Timeline CSV parser and training scenarios. *Digit. Invest.* **28**, 34–43 (2019)
6. L. Zhang, S. Wang, B. Liu, Deep learning for sentiment analysis: a survey. *Wiley Interdisc. Rev.: Data Mining Knowledge Discov.* (2018). <https://doi.org/10.1002/widm.1253>
7. S. Vanaja, M. Belwal, Aspect-level sentiment analysis on E-commerce data, in *International Conference on Inventive Research in Computing Applications (ICIRCA)* (Coimbatore, 2018), pp. 1275–1279. <https://doi.org/10.1109/ICIRCA.2018.8597286>
8. H. Studiawan, F. Sohel, C. Payne, Sentiment analysis in a forensic timeline with deep learning. *IEEE Access* **8**, 60664–60675 (2020). <https://doi.org/10.1109/ACCESS.2020.2983435>
9. E. Guzman, D. Azocar, Y Li, Sentiment analysis of commit comments in GitHub: an empirical study, in *Proceedings of the 11th Working Conference on Mining Software Repositories (MSR 2014)* (Association for Computing Machinery, New York, NY, USA, 2014), pp. 352–355. <https://doi.org/10.1145/2597073.2597118>
10. J. Ding, H. Sun, X. Wang, X. Liu, Entity-level sentiment analysis of issue comments, in *Proceedings of the 3rd International Workshop on Emotion Awareness in Software Engineering (SEmo'18)* (Association for Computing Machinery, New York, NY, USA), pp. 7–13. <https://doi.org/10.1145/3194932.3194935>
11. J. Singh, G. Singh, R. Singh, Optimization of sentiment analysis using machine learning classifiers. *Hum. Cent. Comput. Inf. Sci.* **7**, 32 (2017). <https://doi.org/10.1186/s13673-017-0116-3>
12. Y. Liu, J. Bi, Z. Fan, Multi-class sentiment classification: the experimental comparisons of feature selection and machine learning algorithms. *Expert Syst. Appl.* **80**, 323–339 (2017). ISSN: 0957-4174, <https://doi.org/10.1016/j.eswa.2017.03.042>
13. M. Rathi, A. Malik, D. Varshney, R. Sharma, S. Mendiratta, Sentiment analysis of tweets using machine learning approach, in *Eleventh International Conference on Contemporary Computing* (Noida, 2018), pp. 1–3, <https://doi.org/10.1109/IC3.2018.8530517>
14. R. Vaarandi, V. Blumbergs, M. Kont, An unsupervised framework for detecting anomalous messages from syslog log files, in *IEEE/IFIP Network Operations and Management Symposium* (Taipei, 2018), pp. 1–6, <https://doi.org/10.1109/NOMS.2018.8406283>

15. C. Bertero, M. Roy, C. Sauvanaud, G. Tredan, Experience report: log mining using natural language processing and application to anomaly detection, in *IEEE 28th International Symposium on Software Reliability Engineering (ISSRE)* (Toulouse, 2017), pp. 351–360. <https://doi.org/10.1109/ISSRE.2017.43>
16. Z. Liu, T. Qin, X. Guan, H. Jiang, C. Wang, An integrated method for anomaly detection from massive system logs. *IEEE Access* **6**, 30602–30611 (2018). <https://doi.org/10.1109/ACCESS.2018.2843336>
17. M. Wang, L. Xu, L. Guo, Anomaly detection of system logs based on natural language processing and deep learning, in *4th International Conference on Frontiers of Signal Processing (ICFSP)* (Poitiers, 2018), pp. 140–144. <https://doi.org/10.1109/ICFSP.2018.8552075>
18. Y. Chabot, A. Bertaux, C. Nicolle, M. Kechadi, A complete formalized knowledge representation model for advanced digital forensics timeline analysis. *Digital Invest* **11**(Supplement 2), S95-S105 (2014). ISSN 1742–2876, <https://doi.org/10.1016/j.diin.2014.05.009>
19. C. Hargreaves, J. Patterson, An automated timeline reconstruction approach for digital forensic investigations. *Digital Invest.* **9**(Supplement), S69-S79 (2012), ISSN 1742-2876, <https://doi.org/10.1016/j.diin.2012.05.006>
20. H. Studiawan, F. Sohel, C. Payne, Automatic log parser to support forensic analysis, in *Proceeding 16th Australian Digital Forensics Conference* (2018), pp. 1–10
21. B. Chapin (2013) Timeline creation and analysis guides, in *The Senator Patrick Leahy Center for Digital Investigation (LCDI)* (2013)
22. G. Manogaran, P.M.S., S.B., C. Hsu, S.N. Kadry, R. Sundarasekar, B.A. Muthu, FDM: fuzzy-optimized data management technique for improving big data analytics. *IEEE Trans. Fuzzy Syst.* 1–1. <https://doi.org/10.1109/tfuzz.2020.3016346>
23. P.N. Hiremath, J. Armentrout, S. Vu, T.N. Nguyen, Q.T. Minh, P.H. Phung, MyWebGuard: toward a user-oriented tool for security and privacy protection on the web, in *Future Data and Security Engineering. FDSE 2019. Lecture Notes in Computer Science*, vol. 11814, eds. T. Dang, J. Küng, M. Takizawa, S. Bui (Springer, Cham, 2019). [https://doi.org/10.1007/978-3-030-35653-8\\_33](https://doi.org/10.1007/978-3-030-35653-8_33)
24. J. Olsson, M. Boldt, Computer forensic timeline visualization tool. *Digital Invest* **6**(Supplement, 2009), S78-S87 (2009), ISSN 1742-2876, <https://doi.org/10.1016/j.diin.2009.06.008>
25. Digital forensics Article, <https://www.digitalforensics.com/blog/how-to-make-the-forensic-image-of-the-hard-drive/>
26. M. Saudi, An overview of disk imaging tool in computer forensics, in *Information Security Reading Room* (SANS Institute, 2001)
27. Plaso documentation, <https://plaso.readthedocs.io/en/latest/sources/user/Creating-a-timeline.html>
28. S. He, J. Zhu, P. He, M. Lyu, Loghub: a large collection of system log datasets towards automated log analytics. Arxiv (2020). <https://github.com/logpai/loghub/tree/master/Linux>
29. S. Garfinkel, P. Farrell, V. Roussov, G. Dinolt, Bringing science to digital forensics with standardized forensic corpora. DFRWS 2009, Montreal, Canada. <http://downloads.digitalcorpora.org/corpora/drives/nps-2009-casper-rw/>
30. J. Hao, T. Ho, Machine learning made easy: a review of Scikit-learn package in python programming language. *J. Educ. Behav. Stat.* **44**, 348–361 (2019). <https://doi.org/10.3102/1076998619832248>

# Survey on Voice-Recognized Home Automation System Using IOT



Shreya Mittal, Tarun Bharadwaj, and T. Y. J. Naga Malleswari

**Abstract** In today's world, there is a necessity for innovation that is helpful to the society. IOT is a leading field that encourages the development of products that automate a variety of tasks. IOT refers to a network of devices that are connected through the Internet and are capable of executing tasks, minimizing human involvement. When applied in the right direction, IOT enabled devices are very useful for people with disabilities. The paper focuses on this aspect of IOT and uses the principle to develop a device that will assist individuals with mobility impairment. The proposed device is mainly divided into two segments. The hardware segment includes an Arduino, smart light bulb and a channel relay. The software portion consists of the IOT and cloud platform and an app which is specifically developed for this purpose. The main function of this device is to detect specific words in a person's voice and accordingly activate or deactivate the smart device, creating a smart home environment. Initially, this will be put into practice using a light bulb. The app will be developed using Java programming language. The most distinguishing feature of the application will be the presence of a microphone which will be the means by which commands are given. The light bulb and the mobile phone will also be fitted with Bluetooth and Wi-Fi modules to maximize the speed of transmission. Particular care will also be given to noise sensitivity and speedy transmission. The proposed device can also be customized to include a variety of appliances such as fans, air conditioners etc. A provision for increased safety is considered where the application only responds to a particular voice.

**Keywords** Social innovation · IOT · Speech recognition

---

S. Mittal (✉) · T. Bharadwaj · T. Y. J. Naga Malleswari  
SRM Institute of Science and Technology, Chennai, India

T. Y. J. Naga Malleswari  
e-mail: [nagamalt@srmist.edu.in](mailto:nagamalt@srmist.edu.in)

## 1 Introduction

The world has now become reliant on technology even for day-to-day activities. There are new inventions being developed or improved every day in order to reduce human effort. IOT is a major example of this, and it aims at connecting devices through the Internet in order to create a network of devices with customizable features. IOT also includes a sensor network interconnected through the net, responsible for pushing data and carrying out simple tasks. Smart home is the term given to living spaces that are IOT enabled. All or some of the equipments are connected to the Internet and can be operated hands-free. Smart homes are known to be more accessible and friendly for people with disabilities. The aim of this paper is to present a device that will utilize this technology and help automate homes in a user-friendly, efficient and sustainable way.

Ishan Krishna et al. have created a device that uses ZigBee to automate devices. It is voice activated and can be expanded to a variety of uses. The main drawback being the fact that Zigbee has a relatively shorter range even though it is faster and more effective [1]. Kumar Mandula et al. developed a mobile application-controlled home automotive device that requires 3G Internet feature of the phone to function. Though, the 3G net reduces the reach of concept due to weak mobile signals in many parts of the world, and it is expensive [2]. Ayat Cubuku et al. Developed a voice-control automation system. This device is capable of voice recognition but had the limitation of being only able to recognize two words at a time [3]. Hina Makhija et al. use NodeMCU and Google Assistant to control the device. The entire methodology runs on the webserver; hence, the range varies from device to device. There is no utilization of a phone [4].

The device explained further on will take into account the drawbacks of the already existing products such as short range, noise sensitivity and expensive cost and develop a prototype that has a high noise filtering capacity, longer range as well as a cheaper cost. The main principle is IOT. The sensors fitted onto the light bulb will be connected to the Arduino which acts as the brain of this device. Once the user speaks into the microphone using the app, this will directly be sent to the Arduino which will turn on or turn off the device. Additionally, this can also be monitored through a cloud platform. With the rapid rise of people with disabilities and mobility issues, this device will be of great help and necessity. It can be further improved by adding specialized security with voice detection etc.

## 2 Related Work

Home automation is the future. Hence, there are multiple inventions being developed each day to make life simpler. The range of new devices ranges from voice control and even to devices being controlled by ocular movement. In ‘IOT based Home Automation System with Pattern Recognition’ by Ritvik Iyer et al.[5], the

authors developed an inventive system using a microcontroller to note the frequency and pattern in which the user turned on and off his household devices. The device was also capable of automatically turning on and off the devices at a preset time. ‘Implementing Home automation System using LabVIEW and GSM’ written by Shreenidhi et al. [6] develop a novel idea where the use LabVIEW software and mobile GSM to control local devices. The system simplified complex protocols and is easily implementable. ‘Awareness Home Automation System Based on User Behaviour through Mobile Sensing’ by Nishcan Mafrur et al. [7–9] aim to develop a self-aware HAS using accelerometer and magnetic sensor. This project aims to eliminate the use of buttons or intermediate software. ‘IoT Based Home Automation using Computer Vision’ by Devendra Kumar et al. [10] adopts the use of computer vision and NFC-enabled devices to create an alarm system that notifies a user of any intrusions.

Similarly, ‘Intelligent Home Automation System for Disabled People’ by Husni et al. [11] presents a home automation system specifically for disabled individuals using Zigbee and Raspberry pi. Sensors such as camera and PIR sensors are integrated into the device, and the study is validated. ‘E-MAIL INTERACTIVE HOME AUTOMATION SYSTEM’ by Manohar et al. [12–14] also includes an innovative HAS system involving email. The main component of the project was LED lights.

Table 1 draws a comparison between various devices developed with a similar goal in mind.

It is inferred from the above table that the most preferred technique for the development of such devices is the Arduino Uno due to its high efficiency and low cost. Additionally, speech-to-text conversion software is usually adopted in order to send the signals from the application to the Arduino and eventually to the device. Various noise cancellation algorithms have been used in the pre-existing devices using Python libraries which can be easily implemented. It can be stated that the most used communication protocol is Wi-Fi mainly due to its long range. A serious shortcoming that comes to focus is the lack of security in these devices. While not relevant to small appliances like lights and fans, important appliances like lockers must be more secure. Moreover, some of the developed devices are bulky and are extremely hard to install. The accuracy of the conversion from speech also lags behind in some cases. Hence, if a prototype is being developed, it must overcome most of these difficulties.

### 3 Proposed Architecture

Our proposed system will consist of hardware and software components:

- The main hardware components are an Arduino, the light bulb, dual channel arrays as well as a Wi-Fi module.
- The software mainly consists of a Google Voice API-enabled microphone application developed specifically for this device.

**Table 1** Shows comparison between different IOT devices using different techniques

Title	Authors	Objective	Gaps identified	Advantages	Hardware
'Intelligent home automation system using BitVoicer'	Ishan Krishna K. Lavanya	Uses bit voicer and develops an automated device	The components are too complex and expensive	The mobile device is extremely handy and it is easier to work with	IOT
'Mobile based home automation using Internet of Things (IoT)'	Kumar Mandula Ramu Parupalli CH. A. S. Murty E. Mageesh Rutul Lunagariya	Uses high-speed mobile networks to operate the IoT devices	Bluetooth restricts range	NA, used for smart homes	IOT
'Development of a voice-controlled home automation using Zigbee module'	Aykut Çubukçu Melih Kuncan Kaplan	To develop a device that is capable of responding to voice commands	Zigbee tends to be more expensive	Zigbee offers high transmission speed and reaction time	Zigbee
'Design of a phonebased voice controlled home automation system'	G B Karan Dhananjay Kumar Kiran Pai J Manikandan	To develop a device that can control small appliances like light bulbs with the help of a phone	NA	It can be used to control home devices in an easy way	IOT
'Development of a voice-controlled home automation system for the differently-abled'	Karan Pande Ashirbad Pradhan Suraj Kumar Nayak Pratyush Kumar P atnaikBiswajeet Champay ArfatAnis KunalPal	Developed a device to assist differentially abled people to operate devices reducing the need to move	NA	It has a large social impact by enabling even differentially abled people to do their work	IOT

(continued)

**Table 1** (continued)

Title	Authors	Objective	Gaps identified	Advantages	Hardware
'Voice and touch control home automation'	Sushant Kumar S, S Solanki	Develops a device that uses touch as well as voice to control devices	NA	Is easily implementable and can be used in all homes	Arduino IOT
'Implementation of voice based home automation system using raspberry Pi'	Harshada Rajput Karuna Sawant Dipika Shetty Punit Shukla Prof. Anit Chougule	Uses Raspberry Pi to implement a voice-based system aimed at creation of multiple smart homes	Raspberry Pi is bulky and difficult to use for a beginner	A very novel concept which if used can bring about a change in how homes are designed	Raspberry Pi Arduino microcontroller board
'Voice controlled automation using raspberry Pi'	Amrita Maharaja Namrata Ansari	This paper utilizes Raspberry Pi to create a smart bubble of devices all interconnected and controlled by a single interface	NA	Majorly aimed at differentially abled people for ease of control	Raspberry Pi
'Voice-controlled home automation system'	Hina Makhiya Atul Mathur Manish Kumar	In this project, a prototype of voice-enabled home automation system using NodeMCU and Google Assistant is shown	The process is running on the Web; hence, there is no fixed range	The combinatory offers network, a simple and supple user interface	IOT Arduino
'An overview of basics speech recognition and autonomous approach for smart home IOT low power devices.(main paper)'	Jean-Yves Fourniol et al.	This paper uses a low power device in order to increase the efficiency of smart home devices	Reliability in noisy conditions was less	Increases the level of safety	IOT

(continued)

**Table 1** (continued)

Title	Authors	Objective	Gaps identified	Advantages	Hardware
'Internet of Things (IoT) for building smart home system'	Timothy Malche and Priti Maheshwary	This paper discusses functions of smart home and its applications and introduce FLIP system	Complicated design principle, difficult installation	The transfer loop system prevents accidental actions	IOT
'Arduino based home automation using Internet of things (IoT)'	Lalit Mohan Satapathy et al.	In this paper, they present a home automation system using Arduino Uno microcontroller and module	Controller efficiency was less, delay absence of speech control absence of android phone detection	Will be used to control basic appliances using a Bluetooth remote	Arduino Uno IOT
'Smart Home Automation and Security System using Arduino and IOT'	Siddharth Wadhwanı et al.	The smart home devices can be continuously controlled and monitored on a cloud platform	No specialized app. Very low Wi-Fi range. Speech recognition absent	This will be used to bring apparatus at each side of the house under our control	IOT
'Voice recognition based wireless home automation system'	Humaid AlShu'eli et al.	This paper is about an overview of a smart home system along with its implementation and function using voice control	Absence of Wi-Fi. No specific app outdated components, confirmation commands missing	Use of voice ensures availability to disabled people	Arduino IOT
'IoT architecture for home automation by speech control aimed to assist people with mobility restrictions'	Joceli Mayer	This paper describes research on home automation architecture designed to help people that find it difficult to use a remote control or a smartphone for control purposes	slight time delay	Effective and useful for people with extreme mobility issues	IOT Arduino

Arduino is an electronic device capable of converting an input signal into output signal with the help of coding. It is being used in this application because of its easy operable interface. Mainly, the coding is done using C ++ language. The two-channel relay hardware connection ensures two-way communication that can control the device. The Wi-Fi module ensures faster connectivity between the Arduino and the mobile phone, reducing transmission and action time.

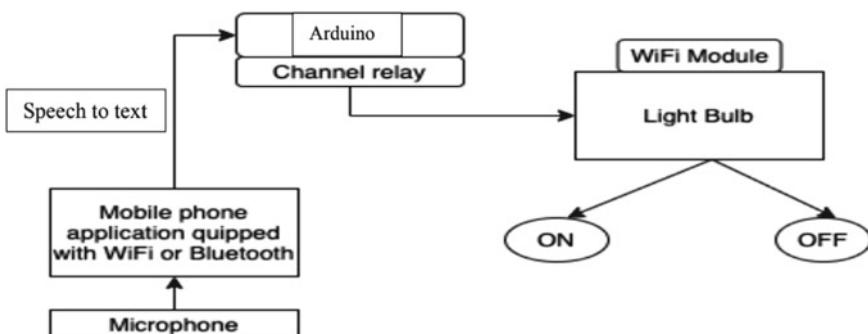
The basic functioning of the prototype starts with the android application.

- (i) The user can open up the app with the mobile device and the light bulb connected to the same Wi-Fi or both connected to each other with Bluetooth.
- (ii) The user says the ON or OFF command into the microphone.
- (iii) This microphone has speech to text converter inbuilt which converts the word into a signal that the Arduino can understand.
- (iv) The Arduino uses the microphone sensor and activates the lightbulb according to command.

According to Fig. 1, the app will operate from the mobile phone. The mobile phone must be connected to the same Wi-Fi as the bulb. The user will have to keep the microphone button pressed and simultaneously speak.

Once he speaks, the trigger words such as ON, OFF or LIGHT the speech-to-text software will be activated and wirelessly transmit this data to the Arduino using Wi-Fi. Special Python library for speech recognition and converts the generated WAV file into speech. The script also continuously monitors the incoming command and directs it to the specific relay channel. The Arduino then in communication with the light bulb will turn it off or on according to the command by cross checking it with the Python library. The Arduino is also connected to a channel relay that is physically connected to the bulb. All the complex computation and speech processing will occur on an online IOT platform. The proposed device is special in the sense that it will be trained using.

Supervised machine learning algorithms to only respond to a particular tone or voice of a person, hereby drastically increasing security. This can be achieved by



**Fig. 1** Work flow of the application

forming a dataset with WAV audio clips and training the model to specifically recognize the frequency of that speech. The module that we will be utilizing is speechrecognition.py. This script in Python uses a library known as speech recognition library and converts the spoken words into text. When the user speaks using microphone in the application, the Python script records it in WAV form and sends it to a platform called wit. Wit also provides a unique API key. WIT converts this WAV format into text and sends it back to the Python Script using a speech recognition algorithm. This text is matched with predefined comments in the code and accordingly operates the device and turns it on or off. Specialized IOT script written in Python called the iot.py monitors the received text and the relay.

## 4 Conclusion

Hence, a detailed survey of the existing techniques was done and a suitable improved prototype was developed. After studying the papers thoroughly, new techniques that are being implemented to change the features and also tweak the command function have been brought to our notice. The following inferences were drawn:

- (i) The users will be able to accomplish turning on or off the appliance by using a very operable application interface using their phones as something technically like a remote.
- (ii) The new prototype is in the initial phases of development and with inputs from the studied papers, new improvements and enhancements will be made. This project provides a suitable arrangement to the need of mechanization at the extremely fundamental level, that is, in our homes.
- (iii) The papers also suggest that usage of Arduino Uno and two-channel relay is the most efficient way to carry out the project.
- (iv) The framework is additionally streamlined by enabling apparatuses to be controlled by our voice only. This device can also be customized to include various different languages. Overall, these devices will be extremely useful in conserving time and specially helping individuals with disabilities.

## References

1. Y. Krishna, S. Nagendram, ZIGBEE based voice control system for smart home. *Int. J. Comput. Technol. Appl.* 03
2. M. Kumar, P. Ramu, C.H.A.S. Murty, E. Magesh, R. Lunagariya, Mobile based home automation using Internet of Things (IoT), pp. 340–343. <https://doi.org/10.1109/ICCICCT.2015.7475301>
3. Ç. Aykut, M. Kuncan, K. Kaplan, H.M. Ertunç, Development of a voice-controlled home automation using Zigbee module, in *2015 23nd Signal Processing and Communications Applications Conference (SIU)* (2015), pp. 1801–1804

4. K. Anitha, B. Theja, B. Thanuja, Home automation using voice via google assistant **11**(4) (2020). ISSN NO:0377-9254
5. R. Iyer, A. Sharma, IoT based home automation system with pattern recognition. Int. J.Rec. Technol. Eng. (IJRTE) **8**(2) (2019) ISSN: 2277-3878
6. Implementing home automation system using LabVIEW and GSM. Int. J. Emerg. Technol. Innov. Res. **4**(5), 287–294 (2017). Available ISSN:2349–5162
7. B. Maram, J.M. Gnanasekar, G. Manogaran, M. Balaanand, Intelligent security algorithm for UNICODE data privacy and security in IOT. SOCA **13**(1), 3–15 (2018). <https://doi.org/10.1007/s11761-018-0249-x>
8. T.N. Nguyen, B. Liu, N.P. Nguyen, J. Chou, Cyber security of smart grid: attacks and defences, in *ICC 2020—2020 IEEE International Conference on Communications (ICC)* (Dublin, Ireland, 2020), pp. 1–6. <https://doi.org/10.1109/ICC40277.2020.9148850>
9. Mafrur, Rischan, Fiqri Muthohar, M., Bang, Gi Hyun, Lee, Do Kyeong, and Choi, Deokjai (2015). Awareness home automation system based on user behavior through mobile sensing. 6th FTRA International Conference on Computer Science and its Applications, CSA 2014, Guam, 17–19 December 2014.(Springer, Dordrecht, Netherlands, 2014)
10. D. Kumar, R.K. Maurya, K. Dwivedi, IoT based home automation using computer vision. Int. J. Innov. Technol. Explor. Eng. (IJITEE) **8**(12) (2019) ISSN: 2278-3075
11. W.N.A.A.W Husni, M. Faisal, P. Jern Ker, D.N.T. How, M.A. Hannan, M.A. Salam, Intelligent home automation system for disabled people. Int. J. Eng. Adv. Technol. (IJEAT) **9**(2) (2019). ISSN: 2249-8958
12. S. Manohar et al., Email interactive home automation system. Int. J. Comput. Sci. Mobile Comput. **4**(7), 78–87 (2015)
13. A. Shrivastav, T.Y.J. Riya Maurya, N. Malleswari, Iot audio analytics for automated safety system using deep learning techniques. Int. J. Adv. Sci. Technol. **29**(05), 9818–9829 (2020)
14. K. Rahul, A. Singh, Throughput optimization for wireless information and power transfer in communication network, in *2018 Conference on Signal Processing And Communication Engineering Systems (SPACES)* (IEEE, 2018), pp. 1–5. <https://doi.org/10.1109/SPACES.2018.8316303>
15. C. Gaurav, A. Singh, R. Kumar, B.K.S. Deo, A. Sehgal, Energy efficient distributed clustering algorithm for improving lifetime of WSNs
16. S. Jacobs, C.P. Bean, Fine particles, thin films and exchange anisotropy, in *Magnetism*, vol. III, ed. by G.T. Rado, H. Suhl (Academic, New York, 1963), pp. 271–350
17. Home Automated Living website. [Cited 2010 14th Oct]. Available: <http://www.homeautomatedliving.com/default.htm>
18. L.R. Rabiner, R.W. Schafer, *Digital Processing of Speech Signals* (Prentice Hall Inc., New Jersey, US, 1978)
19. “XBee-2.5-Manual,” ZigBee RF communication protocol. (2008). Minnetonka: Digi International Inc.
20. B. Yukekkaya, A.A. Kayalar, M.B. Tosun, M.K. Ozcan, A.Z. Alkar, A GSM, internet and speech controlled wireless interactive home automation system, in *IEEE Transactions on Consumer Electronics* (vol. 52, 2006), pp. 837–843
21. F.J. Owens, *Signal Processing of Speech* (McGraw-Hill Inc., New York US, 1993)
22. T.S. Ng, Microcontroller, in *Studies in Systems, Decision and Control* (2016)
23. L. Horsley, M.P. Foster, D.A. Stone, State-of-the-art piezoelectric transformer technology, in *2007 European Conference on Power Electronics and Applications, EPE* (2007)
24. D.R. Toberge, S. Curtis, Arduino Uno, J. Chem. Inf. Model. (2013)
25. C.N.T. Devidas, V.K.N. Ekoskar, HC-05 bluetooth module I interfaced with Arduino. Int. J. Sci. Eng. Technol. Res. (2016)
26. M. Spencer et al., Demonstration of integrated micro-electro- mechanical relay circuits for VLSI applications. IEEE J. Solid-State Circ. (2011)
27. D. Javale, M. Mohsin, S. Nandanwar, M. Shingate, Home automation and security system using android ADK. Int. J. Electron. Commun. Comput. Technol. (IJECCCT) **3**(2) (2013)

28. B. Pandya, M. Mehta, N. Jain, Android based home automation system using bluetooth & voice command. *Int. Res. J. Eng. Technol. (IRJET)* **3**(3) (2016)
29. F. Baig, S. Beg, M.F. KhanInternational, Controlling home appliances remotely through voice command. *J. Comput. Appl. (0975-888)* **48**(17) (2012)
30. S. Akhter, M.F. Arif, M. Nur-Amin, N. Mustafiz, R. Khan, S. Waliullah, K. Hossain, Voice controlled home appliances: the use of android phone. *J. Mod. Sci. Technol.* **4**(1), 179–191 (2016)
31. N. Sriskanthan, T. Karand, Bluetooth based home automation system. *J. Microproces. Microsyst.* **26**, 281–289 (2002)
32. M.I. Ramli, M.H.A. Wahab, Nabihah, Towards smart home: control electrical devices online, in *Nornabihah Ahmad International Conference on Science and Technology: Application in Industry and Education* (2006)
33. Al-Ali, Member, IEEE, M. AL-Rousan, Java-based home automation system R. *IEEE Trans. Consumer Electron.* **50**(2) (2004)
34. G Pradeep, B. Santhi Chandra, M. Venkateswarao, *Ad-Hoc Low Powered 802.15.1 Protocol Based Automation System for Residence using Mobile Devices*, vol. 2 (Department of ECE, K L University, Vijayawada, Andhra Pradesh, India IJCST, SP 1, December 2011)
35. Universal Mobile Application Development (UMAD) On Home Automation, vol 2, no. 2 (Marathwada Mitra Mandal's Institute of Technology, University of Pune, India Network and Complex Systems, 2012). ISSN 2224-610X (Paper) ISSN 2225-0603 (Online)
36. Voice-Controlled Home Automation System. <http://electronicsforu.com/electronics-projects/voice-controlled-homeautomation-system>
37. Arduino and Matlab interfacing via Bluetooth module. <http://crackeconcept.blogspot.in/2014/03/arduinoand-matlab-interfacing-via.html>
38. R. Mafrur, M.F. Muthohar, G.H. Bang, D. Kyeong Lee, D Choi, Awareness home automation system based on user behavior through mobile sensing, in J. Park, I. Stojmenovic, H. Jeong, G. Yi (eds.), *Computer Science and its Applications. Lecture Notes in Electrical Engineering*, vol 330 (Springer, Berlin, Heidelberg). [https://doi.org/10.1007/978-3-662-45402-2\\_135](https://doi.org/10.1007/978-3-662-45402-2_135)
39. S. Sandeep, J. Singh, R. Kumar, A. Singh, Throughput-save ratio optimization in wireless powered communication systems, in *2017 International Conference on Information Communication, Instrumentation and Control (ICICIC)*, (IEEE, 2017), pp.1–6. <https://doi.org/10.1109/ICOMICON.2017.8279031>

# Fingerprinting of Image Files Based on Metadata and Statistical Analysis



J. Harish Kumar and T. Kirthiga Devi

**Abstract** Fingerprinting of digital images is the process of finding evidence from the tampered images by the use of image's metadata and machine learning algorithm. In the recent decades, digital images growing tremendously in the people's life, and digital images have been used in many application areas. Hereafter, digital image integrity should not be taken for granted because now a days popular powerful software used for image editing are available in the Internet at low cost. Because of the technological evolution and the availability of powerful image editing software, digital crimes also increase tremendously. To identify the legitimacy of the image files, we propose a new methodology for the digital image forgery detection by combining the analysis of metadata's exchangeable image file format (Exif) information, statistical features and also with the help of error level analysis method (ELA) that can help the forensic investigators to authenticate the digital image files by distinguishing the difference between the real image and the fake image.

**Keywords** Error level analysis · Metadata · Machine learning · Image forensics · Digital image tampering

## 1 Introduction

In this technological world, huge number of people have become a victim of image forgery and a lot of people use powerful photo editing software to tamper the digital image and using it as an evidence for misleading the court of law. And also, digital images are very popular not only in legal investigation, media, and news but also in research areas. Since the visible traces in the digital images are difficult to find out, so the legitimacy and integrity of the digital images are questionable. Image

---

J. Harish Kumar · T. Kirthiga Devi (✉)  
SRM Institute of Science and Technology, Chennai, India  
e-mail: [kirthigt@srmist.edu.in](mailto:kirthigt@srmist.edu.in)

J. Harish Kumar  
e-mail: [hj5574@srmist.edu.in](mailto:hj5574@srmist.edu.in)

forgery is in existence since the 1860s. There are so many number of cases in the history which affected many organizations and people just because of manipulating the digital image. And also, powerful software used for image editing are made available in the Internet such as Photoshop, Fotor, Pixlr, Photos and POS Pro. There are many tools made available for free in the Internet for editing the digital image. Therefore, digital image forgery detection becomes extremely necessary to find out the picture in question is distorted or not. Hence, it arises the need for the effective image tamper detection method for identifying whether the images are real image or fake image.

Fake images are created by combining two or more image or by making changes of an existing image. Nowadays, image tampering happens at the pixel level of a digital image, so the identification of tampered images is not easy as it was before in the digital era. Current image tampering methods can be classified into two types: (1) facial appearance manipulation and (2) facial identity manipulation. One of the most noticeable facial appearance manipulation strategies is the technique called Face2Face. This technique is used to transfer the facial expression of one person to someone else. Facial identity manipulation is the second most widely used image forgery techniques. Rather than changing expression, these strategies replace the face of an individual with the face of someone else.

Digital image forensics have many numbers of parameters such as image header information, quantization matrix, thumbnails of an image and also the images metadata information for identifying the real sources of an image. Different camera devices have different compression settings, for example, quantization table. Considering all the parameters, we have proposed a method for identifying the fake image from real image by using metadata information of an image combining with error level analysis, and we leverage the recent technical advancement in deep learning, in particular, the ability to learn in-depth analysis of the image features with the convolution neural network (CNN) for classification of fake images from real images.

## 2 Literature Review

There are numerous studies in the field of picture forensics that have already been carried out. Each of the research work provides an insight into a variety of image forensics techniques, with detection and classification of real image from fake images.

Swaminathan et al. [1] provide the analysis on the photo forensics which identifies the distinction of the fake image from real image using the inherent traces. To recognize inborn follows that are deserted in an advanced picture when it experiences different preparing chain. They are using a prototype called detailed imaging model, component analysis model, and higher order statistical model for the fake image detection purpose.

Kee et al. [2] analyzing the images by taking image parameters image dimension, quantization table, and Huffman coding for identifying camera signature. The main

impediment in this work is, it is defenseless against standard rebroadcast attack. So, for overcoming that issue, they are utilizing quantization calculation and Huffman calculation with reconciliation of metadata is utilized for seeing if the picture is phony or unique.

Alani et al. [3] finding out whether the image is original or tampered by analyzing with the forgery features for identification. Here, they are utilizing neural network for finding the phony pictures. Expanding block algorithm, verification algorithm and privacy preserving algorithm were used for classification of digital images.

Chen et al. [4] analyzed based on CNN, here, to create automatic learn feature method combining with the classification of median filtering forensics used to detect fake images. In this work, three models were utilized, they are convolution neural organization, deep Boltzmann, deep auto-encoder. Here, the main algorithm used for the detection process is the back propagation algorithm. This back propagation algorithm is used for supervised learning of artificial neural network.

Lint et al. [5] used one of machine learning model called support vector machine (SVM) model. Here, the existing algorithm assumes that the response function of a camera in the same image is spatially invariant. To overcome the issue, they are using two algorithms called response function recovery algorithm that is used to map the link of pixel irradiance and value, and the algorithm for the detection of doctored images is based on calculating the inverse camera response functions by checking the suitable edge patches.

Caldelli et al. [6] had designed and proposed a method of detecting whether an image originates from a social network and attempting to classify which one was downloaded by extracting the traces left on the digital image. They are using two algorithm called extraction algorithm and bagged tree random forest algorithm for distinguishing the fake images from real images.

Elias et al. [7] has designed and implemented a method for detection of fake images using the compression ratio of digital image. All the pictures are pre-processed and taken care of into the neural network for preparing. Two main models are used, they are deep neural network and multi-layer perception model for the detection purpose.

Kha Tu et al. [8] proposed a novel approach using comparison of features and calculation of sharpness, and using the feature comparison to find the pixel blocks that are close to the forgery of copy shift images and to estimate the sharpness of a digital image to collect the suspicious edges. Extraction in LL sub-groups and sharpness assessment in HH to recognize the duplicate move control model were utilized for high exactness discovery.

Moreira et al. [9] analyzed a way to find the fake image by testing the scope of an image and by combining the applicability of metadata-based inference for constructing the provenance graph for the digital images. CNN model is used for the classification of the fake image.

Mullan et al. [10] have proposed a method to reconstruct the provenance of an image to narrow down the source in which the image was really generated. AI utilizing random forest model is utilized for distinguishing proof of altered pictures.

### 3 Metadata Analysis

Data about data are called metadata. Metadata gives in-depth information about the data. In images metadata, we can find information such as the device make which is used to capture the image, the geolocations like latitude and longitude, resolutions of an image, time stamp, Exif information, model, software, and color space information. Most of the digital images just not contains an image but it also contains so much information about that image called metadata [11]. Image files actually contains different types of files in them. And different types of metadata are stored in different types of files. For example, JPEG image might also contain (1) Exif data which holds camera setting and manufacturer information. (2) IPTC data which has a metadata which is added by the user (3) 8BIM data which is added by the Photoshop (4) ICC data which contains embedded color profile information.

Most of the image's metadata provides information based on how the image file was generated and handled. This information can be used to identify if the image's metadata which appears to be from a digital camera device or altered images [12]. Common information's in an image's metadata are:

(i) **Camera make, software and make used**

This information can be used to identify the device or software which is used to that digital image. In most the image's Exif metadata, the image's make and model will be available and also the software in the camera device.

(ii) **Timestamps**

The timestamps details found on image's metadata information which can be used to identify when a digital image was taken or altered.

(iii) **Metadata types**

There are many metadata types where some of them only generated by device which is used to generate picture, while others are generated by some kind of application software.

#### 3.1 *Metadata Analysis on File System*

File system is the main part which controls how data are stored and retrieved in the computers. Without a record structure, dataset in a storage medium would be one enormous group of data with no way to tell where one bit of information it stops and the next following information starts [13–16]. It controls the access to the content of the files and also the metadata about those files. File system metadata contains the time stamp recorded by the operating system when the file in the system is altered, accessed, or created. It is effectively misinterpreted and can be difficult to comprehend [18–20]. In the file system, metadata can be categorized into two types: (1) Internal and (2) External.

Metadata which is embedded in a computerized object file that it describes is called internal metadata. Internal metadata embedded in the file by itself. For example, ID3

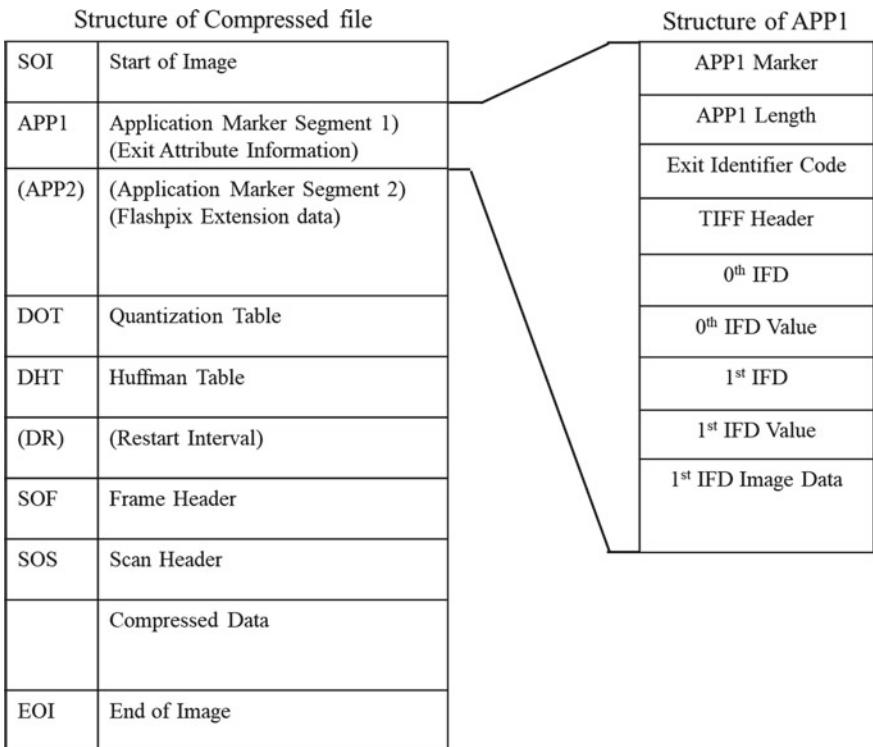
File Name: References for GDF.doc
Title: References for Larry E
Author: Leslie Denton
Comments:
App Name: Microsoft Office Word
Version: 14.0
Date Created (OLE): 1/14/2010 3:08:00 PM
Date Last Printed: 11/17/2010 11:25:00 PM
Date Last Saved: 11/17/2010 11:25:00 PM
Total Edit Time: 1
Template: Normal.dotm
Shared: False
Subject:
Category:
Company:
Keywords:
Manager:
Last Saved By: Larry E. Daniel
Word Count: 131
Page Count: 1
Paragraph Count: 1
Line Count: 6
Character Count: 747
Character Count (with spaces): 877
Byte Count: 0
Presentation Format:
Slide Count: 0
Note Count: 0
Hidden Slides: 0
Multimedia Clips: 0
File Path: E:\My Dropbox\Guardian Documents\Marketing Materials\References for GDF.doc
Created Date (FS): 11/21/2010 8:28:19 PM
Last Modified (FS): 11/17/2010 11:25:09 PM
Last Accessed (FS): 11/21/2010 8:28:19 PM
File Size: 29184
MD5 Hash: D0B77B742AC599A2545AA970945C310A
SHA-1 Hash: B20502C56AB9A96500673231D5FF62E83F8098EC
SHA-256 Hash: 859500878512F40033E20434FE82B7809DF49E1AA16F03AB2A729C9E62808C2A

**Fig. 1** Metadata information of a word document stored

tags, Exif data, and XMP tags are some of internal metadata. Metadata that is stored outside of the document framework that is described is called external metadata [21, 22]. It is simpler to oversee and consume such metadata as it resides in a repository separately from data. The software which are running on the file system like Microsoft Office, Word Perfect Office, Open Office, and Star Office records the metadata in some kind in the word processing document, spreadsheets and presentations created with the programs (Fig. 1).

### 3.2 *Metadata Analysis on Cloud*

With the cloud, these basic description details are put away in a unified area that they can access from anywhere and from any device. The data's stored in the cloud storage have metadata associated with them, which is used to identify the properties of the object, and it specifies the way how the data should be handled [11, 23–26]. The metadata for the cloud storage stored in the format of key:value pair. For example, data class of an object is represented in the metadata entry as dataClass:DATA. Here, the dataClass is the key for accessing the metadata, and all the objects which are having the same keys are associated with them. DATA specifies the value for that particular object.

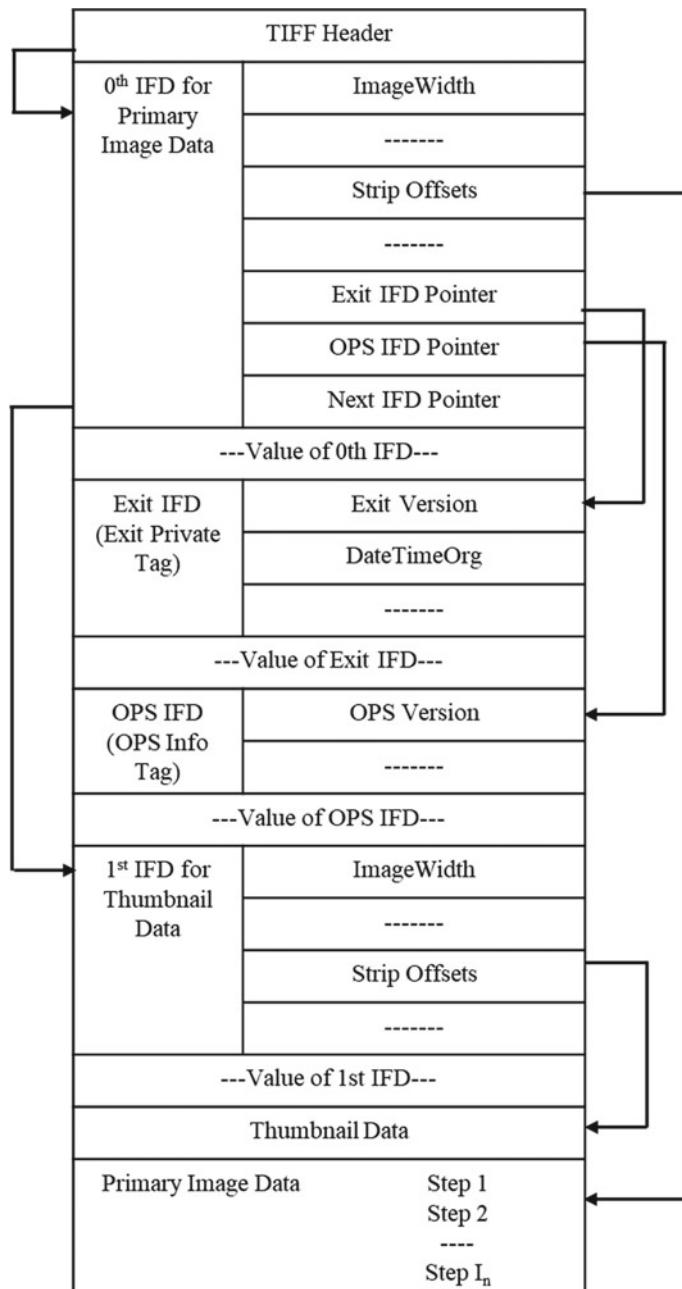


**Fig. 2** Basic structure of compressed image file

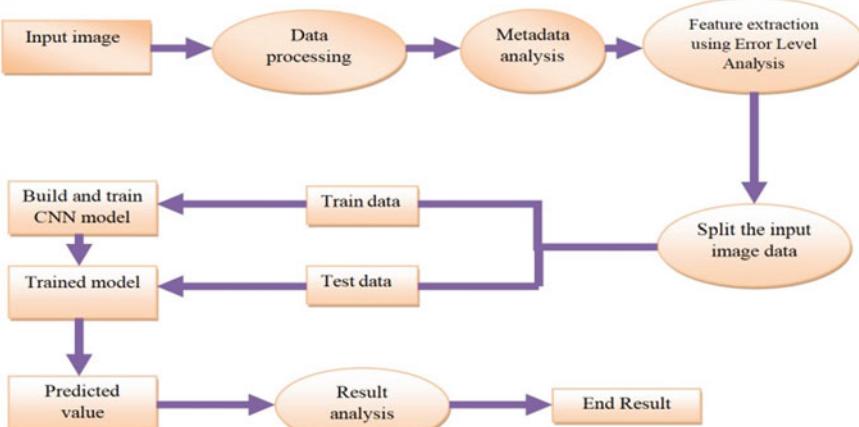
Basically, in cloud platform, the metadata operations are carried out by the help of metadata server cluster and it carried out security strategies in distributed systems. And most importantly, it uses partitioning strategies for the metadata distribution [27, 28]. There are two types of metadata can be found on the cloud. (1) Editable metadata and (2) Non-editable metadata. The editable metadata in the cloud platforms are fixed-key metadata and custom metadata [29–31]. The non-editable metadata's are object metadata like the file stored in the cloud which is not editable (Figs. 2 and 3).

## 4 Proposed Work

The key aim of this project is the implementation of a system to find the distinction between the altered image from real images using the error level analysis which is used to find the compression ratio of the input image and compared with the trained model. If the image is tampered, it has a different compression ratio at the tampered part of the image so that we can easily able to identify the distinction of the tampered



**Fig. 3** Basic structure of uncompressed image file



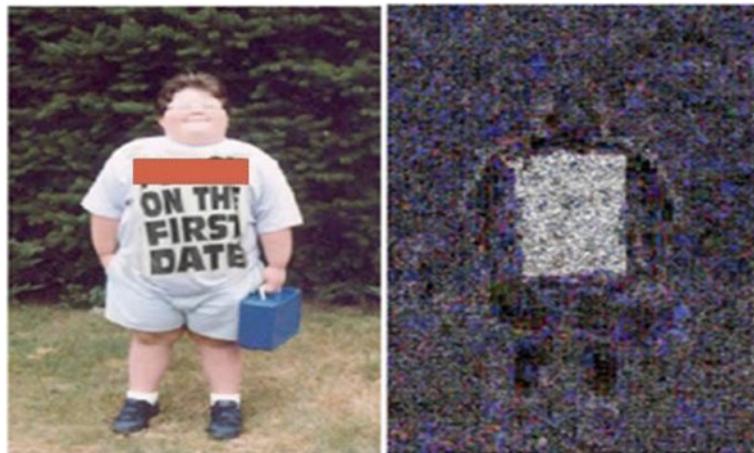
**Fig. 4** Proposed system architecture

part of the digital image. And also, with the use of CNN model for classifying the fake image from real image. The proposed system has multiple layers. They are (1) Conv2d (2) Maxpooling2d (3) Flatten (4) Dropdown (5) Dense (Fig. 4).

## 5 Implementation

Feature extraction process is used to identifying areas of the digital image with different compression ratio. The compression ratio in most of the digital image files should be same, if there are any changes in the compression ratio of the digital images then it is most likely to be altered or modified image. The compression ratio of the real image with the compression ratio of the modified image will be different. The compression ratio part of the modified image will be different from the real image. Each time if the image modified, the amount of error introduced to the digital image will be high that can be easily identified by the error level analysis (ELA). This process works by resaving the digital image at the rate of 95% compression ratio and evaluating the distinctions on the image. Modified areas of the tampered images can be easily identified with the compression ratio of the tampered part of the image. The original image should have an error rate at the same level.

Figure 5 shows the modified image with its error pattern. The picture on the right side shows that the wordings in the t-shirt is modified. The ELA results clearly show the different compression ratio part. The modified part of the image has a high compression ratio than the unmodified part of an image.



**Fig. 5** Digitally modified image with its error pattern

## 6 Decomposition of the System

The main scope of this project is to identify the tampered image by using the efficient metadata analyzer algorithm and error level analysis algorithm and also with the help of CNN model for the identification.

### 6.1 *Giving Image File as Input*

Now a days digital images are very popular in use and also by the technological evolution of mobile phone comes with the high quality camera. Here, digital image is given as an input for identifying whether the given image is real image or fake image. There are different formats of image file are saved; the most popular image formats are JPEG, GIF, PNG, and TIFF. (1) JPEG stands for the Joint Photographic Experts Group, this jpeg image supports up to 24 bits of color depth. The JPEG image uses a lossy compression which is used to reduce the size of the image it stores. JPEG compression produces the high quality photographs with smooth transitions in tone and color. (2) GIF stands for graphics interchange format, this gif image supports 1 to 8 bits color depth. This GIF is used to store multiple bitmap images in a single file for exchange between the platforms and the system. (3) PNG stands for portable network graphics, this png image format supports up to 48 bits color depth. PNGs are mainly designed for transferring the images on the Internet, and it is an improvement on all of the original GIF image's quality except PNG. (4) TIFF stands for Tagged Image File Format, this tiff image format supports up to 24 bits color depth. This TIFF is mainly designed for storing black and white images created by copy machines

and computer applications. Here, the image is given as input, and it goes to the pre-processing phase where the given images is processed so that it enhances the image features important for the further processing which is used for the classification.

## ***6.2 Processing the Input Image File***

In the data processing phase, the given image is processed and converts the images file as an understandable format for the system. It translates the image into numbers, for that it divides the images file into a small portion called pixels. And each translated number describes the property of the given image file.

## ***6.3 Analyzing Metadata Information***

Metadata analysis is the process of analyzing the metadata information of a given input image. This metadata of an image file will give so much information about the image generated source and also in-depth details about the software and camera which is used for taking the picture. Image metadata can be stored in a wide range of ways inside the document itself; and also, there are a few standards for storing and organizing the metadata. The most common metadata formats are exchangeable image file format (Exif), Extensible Metadata Platform (XMP), and International Press Telecommunications Council (IPTC).

This metadata analysis phase extracts metadata information such as compression type, data precision, image height width, make, model, orientation, and software version. After the metadata extraction process, the metadata analyzer algorithm. This metadata analyzer algorithm is basically a tag searching algorithm.

It searches for the Photoshop tags on the image's metadata. If any of the tag present in the image metadata, then the possibility of image got tampered will increase. After the metadata analysis phase, it goes for the another analyzing phase called error level analysis for compression detection (Fig. 6).

The above picture shows the extracted metadata information of a tampered image which has a Photoshop's tag in it. The software which is used for editing the image files leaves some type of intrinsic information on the images metadata. The metadata analyzer first extracts the metadata information from the give file and searches for the Photoshop tag in it. If it finds any of the tags, the possibility of image got tampered increases.

[JPEG] Compression Type - Baseline  
[JPEG] Data Precision - 8 bits  
[JPEG] Image Height - 742 pixels  
[JPEG] Image Width - 572 pixels  
[JPEG] Number of Components - 3  
[JPEG] Component 1 - Unknown (0) component: Quantization table 0, Sampling factors 1 horiz/1 vert  
[JPEG] Component 2 - Unspecified component: Quantization table 1, Sampling factors 1 horiz/1 vert  
[JPEG] Component 3 - Cb component: Quantization table 1, Sampling factors 1 horiz/1 vert  
Exif IFD0--

[Exif IFD0] X Resolution - 300 dots per inch  
[Exif IFD0] Y Resolution - 300 dots per inch  
[Exif IFD0] Resolution Unit - Inch  
[Exif IFD0] Software - Adobe Lightroom 5.4.2 (iOS)  
[Exif IFD0] Date/Time - 2020:10:10 13:04:17 Exif SubIFD--

[Exif SubIFD] Exif Version - 2.31  
[Exif SubIFD] Date/Time Original - 2020:10:10 12:53:24  
[Exif SubIFD] Unknown tag (0x9010) - +05:29  
[Exif SubIFD] Sub-Sec Time Original - 9  
[Exif SubIFD] Color Space - sRGB  
[Exif SubIFD] Scene Capture Type - Standard Exif Thumbnail--

[Exif Thumbnail] Compression - JPEG (old-style)  
[Exif Thumbnail] X Resolution - 72 dots per inch  
[Exif Thumbnail] Y Resolution - 72 dots per inch  
[Exif Thumbnail] Resolution Unit - Inch  
[Exif Thumbnail] Thumbnail Offset - 306 bytes  
[Exif Thumbnail] Thumbnail Length - 17259 bytes XMP--

[XMP] XMP Value Count - 127  
[XMP] Creator Tool - Adobe Lightroom 5.4.2 (iOS)  
[XMP] Metadata Date - 2020-10-10T13:04:17+05:29  
[XMP] Modify Date - 2020-10-10T13:04:17+05:29 ICC Profile--

[ICC Profile] Profile Size - 3144  
[ICC Profile] CMM Type - Lino  
[ICC Profile] Version - 2.1.0  
[ICC Profile] Class - Display Device  
[ICC Profile] Color space - RGB  
[ICC Profile] Profile Connection Space - XYZ  
[ICC Profile] Profile Date/Time - 1998:02:09 00:49:00

Photoshop--

[Photoshop] Resolution Info - 300x300 DPI  
[Photoshop] Thumbnail Data - JpegRGB, 197x256, Decomp 151552 bytes, 1572865 bpp, 17259 bytes  
[Photoshop] Caption Digest - 212 29 140 217 143 0 178 4 233 128 9 152 236 248 66 126  
Adobe JPEG--

[Adobe JPEG] DCT Encode Version - 25600  
[Adobe JPEG] Flags 0 - 192  
[Adobe JPEG] Flags 1 - 0  
[Adobe JPEG] Color Transform - YCbCr File--

[File] File Name - fdgsa.JPG  
[File] File Size - 254774 bytes  
[File] File Modified Date - Sat Oct 10 15:44:26 +05:30 2020

**Fig. 6** Metadata information of a tampered image

#### **6.4 Error Level Analysis**

The next stage is the feature extraction process. In this phase, the important features of the image file are extracted using the error level analysis (ELA). This error level analysis feature extraction process (ELA) is an extremely helpful technique to recognize the control of pictures having a place with a serious picture investigation. This is basically a compression detection algorithm which is used for identifying the areas within an image with different compression levels.

With digital images, the entire parts of the images should be roughly at the same error level, on the off chance that a segment of the picture is at an altogether unique blunder level, at that point it probably shows a computerized alteration. This extraction process works by resaving the digital image at 95% compression ratio and evaluating the distinctions on the image. Modified areas of the tampered images can be easily identified with the compression ratio of the tampered part of the image. The compression ratio is checked with the process of feature extraction; the results will

be splitted and sending it to the train data and test data module for further processing of tamper detection.

### **6.5 Testing the Image with CNN Model**

Convolution neural network (CNN) model is a class of deep learning neural network; it performs the classification operation and learns the features automatically. When compared with other models, CNN gives high accuracy for image detection and classification. If we give two images, it can able to easily detect the distinction between them with high accuracy. The CNN model will be trained with high volumes of both fake and real images, so that it will give high accuracy for the tampering detection of given image. After checking with metadata analysis and error level analysis, the results of the analysis phase go to the CNN model to identify the tamper detection.

The previous analysis results will be splitted and sending it to the train data and test data module. The input image is checked with already trained model on the convolution neural network (CNN). This convolution neural network has an input layer, and in the middle, it has multiple hidden layer for processing the input image and an output layer for identifying the modification sign on the digital image. And also, it uses all the input image for learning the model so that its accuracy of detection keeps on increasing when it get trained will more number of data of the image. The results from the test model will give the predicted value which show the given image is real image or fake image. With that, we can easily classify the fake image with the real image.

## **7 Expected Results**

In this paper, the objective is to identify the problem in the detection of fake images. In the existing system, the main problem is that they can be used to detect the specific tampering method like copy-move, splicing. We observed that fake images and their corresponding real image with the help of error level analysis algorithm. The main objective of this fake image detection is to make the effective model for classifying the fake image from the real image. Without prior knowledge of the original image, the system aims to verify the validity of the digital images. There are many ways for tampering the image because of the availability of powerful software made available in the Internet. But, here, we used error level analysis for detecting the compression ratio and also with the CNN model used for the classification of the tampered image. Expected outcome of this project is when we give new digital image to our system it is going to apply the ELA algorithm and CNN Algorithm to that given image then it will classify the given image is fake or real.

## 8 Conclusion

To achieve the identification of legitimacy of an image, we are using an algorithm called error level analysis algorithm which is used for finding the compression ratio of the given image. By analyzing the compression ratio and also with the help of convolution neural network, we can able to identify the fake image from real image with high accuracy.

## References

1. A. Swaminathan, M. Wu, Digital image forensics via intrinsic fingerprints
2. E. Kee, M.K. Johnson, H. Farid, Digital image authentication from JPEG header
3. A. Alani, Z. Al-Khanjari, Detection techniques of digital image forgery by using images metadata in digital investigation
4. J. Chen, X. Kang, Y. Liu, Z. Jane Wang, Median filtering forensics based on convolution neural network
5. Z. Lint, R. Wang, X. Tang, H.-Y. Shum, Detecting doctored images using camera response normality and consistency
6. R. Caldelli, R. Becarelli, I. Amerini, Image origin based on Social network provenance
7. M.A. Villan, J. Paul, K. Kuruvilla, E. P Elias, Fake image detection using machine learning
8. H.K. Tu, L.T. Thuong, H.V.U. Synh, Develop an algorithm for image forensics using feature comparison and sharpness estimation
9. A. Bharati, D. Moreira, J. Brogan, P. Hale, K.W. Bowyer, Beyond pixels: image provenance analysis leveraging metadata
10. P. Mullan, C. Riess, F. Freiling, Forensic source identification using JPEG image headers: the case of smartphones
11. M. Kirchner, R. Bohme, Hiding traces of resampling in digital images
12. J. Fridrich, Image watermarking for tamper detection, in *Proceeding IEEE International Conference Image Processing* (Chicago)
13. A. Popescu, H. Farid, Exposing digital forgeries by detecting traces of resampling
14. G. Cao, Y. Zhao, R. Ni, L. Yu, H. Tian, Forensic detection of median filtering in digital images
15. H.K. Tu, L.T. Thuong, H.V. Khao, N.C. Sy, A survey on image forgery detection techniques
16. A. Swaminathan, M. Wu, K.J.R. Liu, Component forensics of digital cameras: a non-intrusive approach
17. A. Swaminathan, M. Wu, K.J.R. Liu, Image tampering identification using blind deconvolution
18. N.T. Le, J. Wang, D.H. Le, C. Wang, T.N. Nguyen, Fingerprint enhancement based on tensor of wavelet subbands for classification. *IEEE Access* **8**, 6602–6615 (2020). <https://doi.org/10.1109/ACCESS.2020.2964035>
19. G. Rajmohan, C.V. Chinnappan, A.D. William, S.C. Balakrishnan, B.A. Muthu, G. Manogaran, Revamping land coverage analysis using aerial satellite image mapping. *Trans. Emerging Telecommun. Technol.* (2020). <https://doi.org/10.1002/ett.3927>
20. M. Stamm, K.J.R. Liu, Forensic detection of image manipulation using statistical intrinsic fingerprints
21. E. Kee, H. Farid, Exposing digital forgeries from 3-D lighting environments
22. J. Zheng, T. Zhu, Z. Li, W. Xing, J. Chang Ren, Exposing image forgery by detecting traces of feather operation
23. H. Farid, Exposing digital forgeries from jpeg ghosts
24. T. Bianchi, A. Piva, Detection of nonaligned double jpeg compression based on integer periodicity maps

25. A.E. Dirik, N. Memon, Image tamper detection based on demosaicing artifacts
26. E. Ardizzone, A. Bruno, G. Mazzola, Copy–move forgery detection by matching triangles of keypoints
27. J. Fridrich, D. Soukal, J. Lukás, Detection of copy move forgery in digital images
28. A.C. Popescu, H. Farid, Exposing digital forgeries by detecting traces of re-sampling
29. A. Krizhevsky, I. Sutskever, G. Hinton, Imagenet classification with deep convolutional neural networks
30. M.C. Stamm, K.J.R. Liu, Forensic estimation and reconstruction of a contrast enhancement mapping
31. A.C. Popescu, H. Farid, Exposing digital forgeries in color filter array interpolated images

# Comparative Study of Risk Assessment of COVID-19 Patients with Comorbidities



Satwika Kesana, Meghana Avadhanam, and T. Y. J. Naga Malleswari

**Abstract** COVID-19 coronavirus is now a widespread, vicious contagion that has costed more than a million lives in a span of less than a year, as of October 2020, according to the WHO. Since the termination of the proliferation of this bug is not occurring effortlessly, it is extremely important to at least tone down the mortality rate in mankind due to the coronavirus. Advanced technology such as Artificial Intelligence, is being utilized to provide Intelligent Support Systems for doctors to treat the COVID-19 infected patients more functionally. The target of this study is to understand, analyse, compare few specimens of the existing Artificially Intelligent COVID-19 Risk of Fatality Assessing Systems and suggest new, more efficient ways to assess the risk of fatality by comprehending the pros and cons of each of the specimen research papers. The objective of this study is to make it easier for future researchers to find the most accurate and efficient way to proceed with their research regarding COVID-19 risk assessment using Artificial Intelligence techniques. Research papers written by research authors from all around the world, regarding the risk assessment for patients of COVID-19 with comorbidities using AI techniques, are collected and inspected in depth. The procedures followed by these researchers to perform the analysis on datasets concerning populations from different parts of different countries are juxtaposed to recognize their resistances and vulnerabilities to COVID-19 coronavirus. Finally, the merits and demerits of each specimen research paper are analysed and put forth in a lucid, crisp manner to make it uncomplicated for subsequent researchers. Machine learning techniques such as classification and regression algorithms are used repeatedly on textual datasets. Although the resulting accuracy is good, the models are required to be more generalized to be widely used. This is because of the utilization of centred-datasets which

---

S. Kesana (✉) · M. Avadhanam · T. Y. J. Naga Malleswari  
Department of CSE, SRMIST, Kattankulathur, Chennai, India  
e-mail: [kk6420@srmist.edu.in](mailto:kk6420@srmist.edu.in)

M. Avadhanam  
e-mail: [aa6817@srmist.edu.in](mailto:aa6817@srmist.edu.in)

T. Y. J. Naga Malleswari  
e-mail: [nagamalt@srmist.edu.in](mailto:nagamalt@srmist.edu.in)

pertain to small regions. Therefore, we will attempt to use more generalized data and measure the accuracy with the deep learning technique—convolutional neural network, for our research.

**Keywords** COVID-19 · Coronavirus · Risk of fatality · Machine learning · Deep learning · Artificial intelligence · Convolutional neural network

## 1 Introduction

The COVID-19 coronavirus has shuddered the globe in all the possible roads showing devastating effects on public health, global economy, the food systems and the trade market. It is declared as a Medical Health Emergency by the WHO. Now, the entire world is fighting this global pandemic. There is an expeditious and a brisk rise in the number of coronavirus infected victims which is around 357,704 K as of October 2020. COVID-19 showed a disastrous effect on the global community hence putting great pressure on the healthcare professionals for the risk identification and reduction of the fatality or casualty rate among the patients of COVID-19. Identifying the susceptibility of a patient to COVID-19 coronavirus at an early stage is beneficial to provide effective clinical treatment, hence reducing the mortality rate. Lack of modern techniques in the risk identification sector is instigating more mutilation to the remediable and recoverable lives. To style an improved use of the prevailing resources and to attain enhanced outcomes, the entire medical history of a patient is gathered. A huge textual data and image data of various patients is collected to identify how much a patient is vulnerable to COVID-19. Several risk factors that are associated for developing severe COVID-19 have been identified which includes age, obesity, smoking history and comorbidities like diabetes, cardiovascular diseases, chronic obstructive lung (COPD), cancer, liver disease, renal or kidney diseases, hypertension etc. An individual is evaluated at a granular level, kept under intensive care and monitored thus saving the lives during the pandemic.

Since a scientifically justified vaccine has not been introduced into the global market yet, integration of artificial intelligence with the preventive medicine systems is in high demand. Non-medical therapeutic techniques such as artificially intelligent expert systems, machine learning techniques, data extracting and data mining techniques provide a tone of support in recognizing the mortality risk prediction of COVID-19. Machine learning methodologies like decision trees, naive Bayes algorithm, support vector machine (SVM), ensemble model like random forest and multinomial logistic regression have been used for training the dataset. Among these, an approach for predictive modelling like the decision tree algorithm and fine outlier detector like support vector machine are giving accurate and good results. The model needs to be more summarized, so we are endeavouring to utilize profound learning strategies like convolutional neural networks besides these machine learning approaches.

## 2 Related Work

To uplift the human race from this global pandemic, prior identification of the disease is important to give effective treatment to the patient. This calls for artificial intelligence, machine learning and deep learning techniques and prediction models to predict the mortality rate of COVID-19 patients with comorbidities.

In one study, based on the number and the type of comorbidities in a patient he/she is either admitted in ICU or kept on ventilator otherwise death. The most common and majorly impacting comorbidities are hypertension, cardiovascular diseases, diabetes, chronic obstructive pulmonary disease, chronic kidney diseases and immunodeficiency. Patients with two or more comorbidities are at a higher risk when compared to patients with one or no comorbidities [1]. A random forest model boosted by the AdaBoost algorithm using the F1 score as a primary metric is proposed in Iwendi et al. [2]. The pervasiveness of symptoms such as fever, fatigue, and dyspnea together with the comorbidities are extracted and risk analysis is done between severe and non-severe patients. Odds ratios and 95% confidence intervals are used in Yang et al. [3]. Machine learning algorithms like support vector machine, linear regression and multi-perceptron are used to identify predictive biomarkers that have a significant role in determining the risk factors in Zaki et al. [4]. Diabetes is found to be a crucial factor followed by heart diseases, hypertension and COPD [4]. A hybrid approach which resulted by fine-tuning the prediction from a baseline model as in 5. A mortality risk prediction model by an ensemble model which is derived from four other machine learning techniques like logistic regression, support vector machine, gradient boost decision tree and neural network algorithms which predicts the mortality rate of a patient in more than 20 days in advance is proposed in Gao et al. [6]. Blockchain in addition to artificial techniques was used in another paper to identify risk as in Nguyen et al. [7]. Partial derivative regression model and non-linear machine learning model is used in Kavadia et al. [8]. Additional risk factors like tachypnea, low SpO<sub>2</sub>, higher levels of IL-6, D-dimer concentration and troponin level are considered in a research paper which used Cox proportional hazard regression modelling as in Mikami et al. [9]. An integrated cycle of prevention, response and recovery are proposed using support vector machine classifier but this has a hurdle of overfitting and vague reporting as in Fakhruddin et al. [10]. In a previous study of categories of problems arisen due to COVID-19, it was recognized that there are broadly 13 types of such issues [11]. In another study, they concluded that the performance of predictive models can be improvised with the help of more in-detailed data belonging to patients [12]. In research regarding the estimation of uncertainty, they concluded that accuracy is highly correlated with uncertainty for patients infected with COVID-19 [13]. Image analysis to predict the rapid expansion and severity of COVID-19 in patients was done in [14, 15]. Pneumonia originated COVID-19 symptoms study was done by Huang et al. [16] for patients in Wuhan. Another research was done on patients with fever, cold or any other nearby symptoms and those patients who had travelled recently to China [17]. Patients affected with COVID-19 are known to have an increase in the speed of their pulse rate.

Researchers in Acharya et al. [18] built a deep learning model to classify the various kinds of heartbeats, which can also be used to understand whether or not a patient is infected with COVID-19. A study examined the quarantine routine and patient discharge process and mentioned the re-evaluation based on RT-PCR tests [19]. This study persevered to develop a fully automated deep learning system which classifies COVID-19 based on a specially made chest CT dataset [20, 21]. A typical CNN has nodes in one layer sparsely connected to nodes in the next layer which use the sliding scalar product, called convolution. To deal with the information, a completely associated feed forward neural network, as in NLP, when utilized for pictures, will not just arrange the pictures yet will naturally gain proficiency with the highlights too as in Naga Malleswari et al. [22].

### 3 Techniques Used

Several non-medical therapeutic techniques such as machine learning techniques, artificial intelligence and data mining techniques are used in various research and survey papers for identifying the risk factors of COVID-19, thereby training the dataset to predict the risk and fatality rate among patients of COVID-19. It is evident from the literature survey that the most commonly used artificial intelligence techniques include decision tree algorithm, SVM classifier, ensemble model like boosted random forest, univariate and multivariate logistic regression models, MNB classifier. In addition to these ML techniques, we are also using significant learning methodology such as convolutional neural networks (CNN) to get familiar with the imaging attributes in the convalescents of COVID-19 [23].

#### 3.1 Decision Tree Algorithm

This is a supervised learning technique that uses CART algorithm, which stands for classification and regression tree algorithm. This is the most used algorithm in the medical decision making because it provides high classification accuracy with the collected data and it is a most reliable technique that can be used for medical diagnosis.

#### 3.2 Support Vector Machine

A managed artificial calculator which discovers a hyperplane that differentiates the classes to a large extent. This is mostly used for classification problems though it can be used for both classification and regression type problems. SVM produces the best results because it works well when there are too many features to handle (in

our study, there are too many clinical parameters). Using SVM gives us more robust results.

### ***3.3 Linear Regression***

This algorithm is used when a dependent variable is to be predicted based on an independent variable. This technique is performed after correlation. When multiple explanatory variables (independent variables) are used then it is called multiple linear regression. Mostly used technique in Epidemiology.

### ***3.4 Random Forest Classifier***

This is an ensemble ML model which is similar to decision trees. The ensemble model used is bagging and the individual model used is the decision tree. The bootstrap strategy is used for training this algorithm. This can be effectively used when the dataset is highly varied but under low bias. These are a way of averaging multiple decision trees. Random forest searches for the best feature (clinical parameter) among all the collected parameters hence making the model more random than just growing the trees.

### ***3.5 Boosted Random Forest***

This consists of two parts: one is AdaBoost and the other one is Random forest classifier. This algorithm gives the average of all the results obtained from multiple decision trees and hence gives an accurate final result. This is a more generalized method when compared to other traditional ml algorithms. To tune the hyperparameters, a grid search is used to improve the performance of the model.

### ***3.6 Logistic Regression Methods***

It is a supervised ML algorithm to predict the probability of a target variable. This calculates the class membership probability. This is an extensively used algorithm to identify the risk factors for various diseases and for predicting the mortality from a particular ailment. This is a great tool used in medical institutions to identify those patients who are more vulnerable to a particular disease and make necessary arrangements to treat them.

### ***3.7 Multinomial Naive Bayes Algorithm***

This algorithm is based on Bayes theorem which assumes the features in the dataset to be mutually independent. This is used to calculate the conditional probability of an event. Since it is highly scalable it has found its application in disease prediction-based on clinical parameters. It does not require a big training data. This type of algorithm takes a small sample dataset to predict the model. Medical pros utilize this naive Bayes to demonstrate if a patient is at high danger from specific afflictions such as cancer, diabetes, blood pressure.

### ***3.8 Convolutional Neural Network***

This is a deep learning algorithm which makes use of perceptron to analyse the data. This derives a model on accurate prediction on various factors or issues. The upside of utilizing this procedure is that it consequently distinguishes the significant highlights with no human intercession or supervision. It is computationally more efficient than traditional ml techniques because it is best to use where we have to predict the images (CT scans and other medical tests).

## **4 Literature Survey**

The table data represents the pros and cons of some previous work that has been taken into inspection. The data has been prioritised, not by order of importance but as follows (Table 1).

Papers 1, 3, 4 and 10 are styled using a similar strategy of using the most feasible machine learning classification principles on different datasets to achieve a fairly precise predictive model. Moreover, these papers use epidemiological as well as clinical data to make intelligent decisions. Three major prognostic factors (biomarkers) which gave 90% accuracy for foretelling the risk of fatality are used in paper [24]. Ensemble models are used in combination with the traditional algorithms of machine learning in the papers [25, 26]. Some papers are modelled with the help of demographic data combined with clinical attributes [26–28]. Paper [29] makes use of images of CT scans of patients to detect and predict anomalies that affect malignantly with respect to the comorbidities found in COVID-19 patients.

Paper [30] is built on an artificial neural network with tenfold cross-validation. Risk prediction models inherit primarily the same underlying concept but produce so many variations with diverse algorithms, datasets and attributes. N-fold validation techniques are made use of in paper [31]. A clear distinction is made between the patients with the varied level concentration of D-dimer as in paper [32]. There is future scope for the results to be made more generalized for the world population in

**Table 1** Collation of research papers deploying AI techniques for COVID-19 assessment

S No.	Algorithm used	Merits	Demerits	Accuracy (in %)
1	Supervised XGBoost classifier [1]	Lays out precarious factors and predicts the risk of death, thus the work being two folded Highlighting features include lymphocytes, LDH and hs-CRP Circumvent Overfitting	Exclusively data-directed project A retrospective study which is self-aborted This is a single centred and small sample-study	> 90
2	Logistic regression and MNB, stochastic gradient boosting algorithms [2]	Radiographic images of the chest are taken to begin with COVID-Net, and finally, 80% accuracy is obtained using this deep CNN model Ticketed patients into 4 classes such as SARS, ARDS, COVID, Both (COVID and SARS)	Feature engineering can be done enormously to yield better accuracy The size of the data set can be increased than the one that is studied	96.2
3	Ensemble model using 4 ML methods-logistic regression, gradient boosted decision tree, SVM and neural network [3]	Prognosticates death of a patient up to 20 days beforehand	A broader section of patients from different geographical locations can also be taken in the dataset	95.5
4	Decision tree [4]	Special purpose python libraries are imported to implement algorithms Five-fold cross-validation approach is used	Accuracy is solely taken as the deciding factor for model performance	99.85
5	Univariable and Multivariable logistic regression methods [5]	The algorithm is trained on a multicontinental dataset, patients of COVID-19 are detected based on their CT scan	The model was presented for two groups with a prevalence of less than 200 and ranging from 200 to 1000, and the model errors are 9.42 and 17.08	90.8

(continued)

**Table 1** (continued)

S No.	Algorithm used	Merits	Demerits	Accuracy (in %)
6	Multivariate unconditional logistic regression method [6]	The analysis showed that gender, age, coronary heart disease, hypertension and concurrent myocardial damage and were found to be the major factors of risk in patients with severe and critical COVID-19 during hospitalization	The data set chosen consists of only 100 patients. So, this is a small sample-study Require a larger dataset for accurate results	83.3
7	Postprocessing multicalibration algorithm [7]	The patient is evaluated and diagnosed at a personalized and a granular level The model is prejudiced and positioned hence deployed in health institutions for the betterment in testing and treatment	Data is poorly reported in some regions There is a likelihood for sky-high bias	94.3
8	Deep ANN [8]	Oxygen saturation, count of WBC and APACHE-II risk prediction score are taken as parameters. AUC ROC value obtained is 0.92%	A single—centred study The dataset taken into account for this study has only a small number of patients which is a challenge thrown to compute the statistics involved	92.0
9	Linear regression analysis and ROC curve Analysis [9]	A correlation between the D-dimer concentration and the operation of organ systems and some inflammatory factors is established in this study Patients are categorized as the ones with normal D-dimer concentration group and the patients with upraised levels of D-dimer	Since this study uses linear regression, the data assumes a linear relationship between the patients and the features which are sometimes problematic	95

(continued)

**Table 1** (continued)

S No.	Algorithm used	Merits	Demerits	Accuracy (in %)
10	Logistic regression modelling [10]	Data transmission devices and some mobile apps are used to collect the health record of patients. The discharged and the suspected patients are followed up collecting their samples of blood and urine even after a patient is fully-recovered	The input factors from patient required for this model are not easily available. Hence, the results are not generalized for the world population	-

the paper [33]. Variation in datasets and algorithms produce a variety of accuracies and the most optimal among them is chosen. The trend seems to be using these models as decision support systems for the treatment of COVID-19 patients with comorbidities.

## 5 Principal Findings

The above-displayed stats show us varied data that can lead us to interesting information regarding the research papers analysed in this survey. This survey is based on research papers that were written by research authors from all around the world and hence this is a crisp inference of the combination of the above papers.

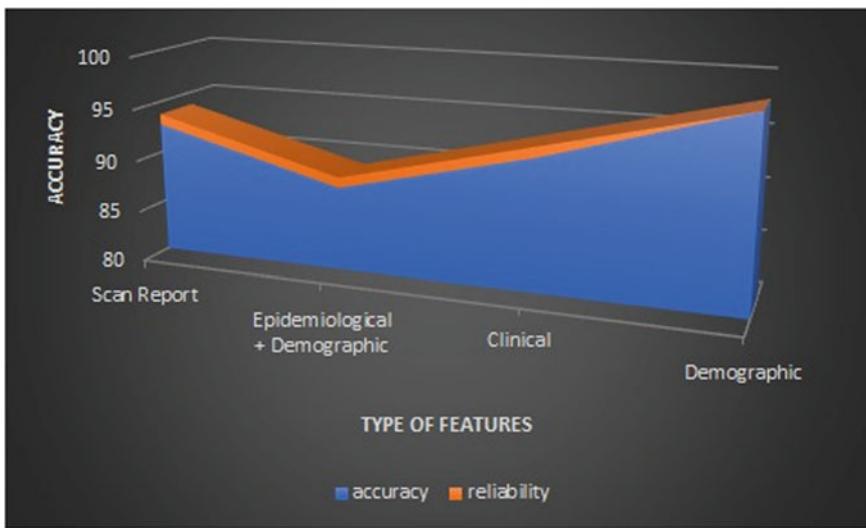
The papers use a wide range of attributes to train their artificially intelligent model.

Broadly, these attributes/features can be sorted into two categories: epidemiological and demographic.

Epidemiological features are those features that are medical in nature, such as control of diseases, observation of symptoms, blood sample reports, chemical levels in the blood, urine samples, more patient-details. For example, one of the above-mentioned papers makes use of a parameter called “concurrent myocardial damage”, which is an epidemiological feature.

Demographic attributes are the parameters that relate physically to the population or a section of the population. Some examples of demographic features are age, gender, ethnicity, etc. Usually, demographic structures are a critical factor in discovering patterns that can estimate the spread of the contagion like COVID-19 coronavirus.

Clinical features are those signs and symptoms which are exhibited by a patient when they are under surveillance of a doctor/hospital. For instance, temperature changes, fever, pain, cough, etc.



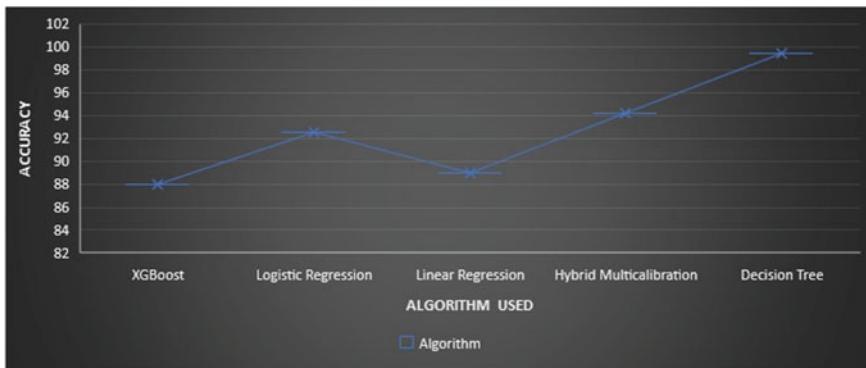
**Fig. 1** Graph representing the accuracies and reliabilities given by AI models verses different types of features

From Fig. 1, it can be understood that the models which utilise attributes which are demographic by nature, such as percentage of the population infected by corona, age, gender and other such population-related features, produce higher accuracy as compared to the epidemiological attributes. The shortcoming of this statement is that although demographic data produces a higher accuracy, it cannot be blindly stated that demographic data alone is sufficient to assess risk and predict mortality or survival.

Epidemiological features, including clinical features, are vital for assessing the risk of a patient to COVID-19 and performing prognosis for their survival. For example, an infected patient's vulnerability to COVID-19 coronavirus cannot be solely determined by his/her age and hypertension parameters. An in depth understanding of his/her health report is imperative if we want globally accurate results.

Hence, it is obligatory to take into account, patients' epidemiological features.

Figure 2 represent the various methodologies used by researchers and their corresponding efficacies. From Table 2, it is seen that the utilisation of machine learning classifier algorithm—logistic regression is the maximum. This is because logistic regression is the simplest classifier algorithm that can be used for complex problems. It can also interpret the relations among dependent and independent variables. On the other hand, it is clearly visible that the decision tree algorithm is used a nominal number of times although it provides an accuracy greater than any of the other classifier algorithms. This is because decision trees carry discrepancies such as overfitting and underfitting. Such inconsistencies cause decision trees to be used a minimal number of times.



**Fig. 2** Graph representing the average accuracies categorized by the algorithm used in different AI models

**Table 2** The skeletal structure of the artificially intelligent models used in the above-mentioned papers

No. of features	Dataset size	Accuracy	Type of features	Algorithm used
3	375	90	Epidemiological + clinical	XGBoost
3	4179	94.3	Epidemiological + clinical	Hybrid multicalibration model
5	1505	99.85	Demographic	Decision tree
5	1280	92.8	Scan report	Logistic regression
6	105	80	Epidemiological + demographic	Logistic regression
7	72	88.9	Epidemiological + clinical	Linear regression
14	2160	96.21	Epidemiological + demographic	Logistic regression
36	2243	97	Epidemiological + clinical	Logistic regression
40	212	96.2	Demographic	Logistic regression

- Overfitting—Fits noise along with data. When there are too many data points, the model gets overfit and hence memorizes instead of learning.
- Underfitting—occurs when the data is too simple to fit a complex model. Leads to assumptions and hence incorrect predictions.

## 6 Conclusion and Future Scope

Artificial intelligence techniques are widely being used in recent times to predict the risk and fatality rate in patients of COVID-19 to reduce the number of deaths and increase efficient treatment procedures at the right time. A researcher who will pursue the prediction and risk assessment of mortality in COVID-19 patients with comorbidities will be required to find a dataset, algorithm(s), parameters and feasible accuracy in order to gain new research in this area. This comparative study will be useful for future researchers in the following mentioned ways. In a broad sense, it can be stated that a majority of the research papers employed **logistic regression algorithm** to assess the risk of fatality of COVID-19 patients with comorbidities and accomplish a viable accuracy. Despite the privileges manifested by all kinds of textual datasets, many of these models perform precisely for only small-scale localities. On the other hand, the experiments that make use of **epidemiological plus demographic datasets** considering the chemical level of substances in blood samples provide great precision and more importantly, greater reliability. Overall, manifold papers were analysed and understood. Their efficacies and parameters were compared and some useful inferences were drawn regarding their performances. Therefore, it can be conveyed that COVID-19 patients can be positively controlled and be monitored with the aid of artificially intelligent support, surveillance and treatment.

## References

1. W.-J. Guan, W.-H. Liang, Y. Zhao, H.-R. Liang, Z.-S. Chen, Y.-M. Li, X.-Q. Liu, R.-C. Chen, C.-L. Tang, T. Wang, C.-Q. Ou, L. Li, Comorbidity and its impact on 1590 patients with Covid-19 in China: a nationwide analysis. *Euro. Respir. J.* (2020)
2. C. Iwendi, A.K. Bashir, A. Peshkar, R. Sujatha, J.M. Chatterjee, S. Pasupuleti, R. Mishra, S. Pillai, O. Jo, COVID-19 patient health prediction using boosted random forest algorithm. *Front. Public Health* (2020) <https://doi.org/10.3389/fpubh.2020.00357>
3. J. Yang, Y. Zheng, X. Gou, K. Pu, Z. Chen, Q. Guo, R. Ji, Prevalence of comorbidities and its effects in patients infected with SARS-CoV-2: a systematic review and meta-analysis. *Int. J. Infect. Dis.* **94**, 91–95 (May 2020)
4. N.M. Zaki, E.A. Mohamed, S.W. Ibrahim, G. Khan, The influence of comorbidity on the severity of COVID-19 disease: systematic review and analysis (2020) <https://doi.org/10.1101/2020.06.18.20134478>
5. N. Barda, D. Riesel, A. Akriiv et al., Developing a COVID-19 mortality risk prediction model when individual-level data are not available. *Nat. Commun.* **11**, 4439 (2020). <https://doi.org/10.1038/s41467-020-18297-9>
6. Y. Gao, G. Cai, W. Fang et al., Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. *Nat. Commun.* **11**, 5033 (2020). <https://doi.org/10.1038/s41467-020-18684-2>
7. D.C. Nguyen, M. Dinh, P.N. Pathirana, Seneviratne: Blockchain and AI-Based Solutions to Combat Coronavirus (COVID-19)-like Epidemics: a survey (2020). <https://doi.org/10.36227/techrxiv.12121962>

8. D.P. Kavadia, R. Patanb, M. Ramachandran, A.H.Gandomi, Partial derivative nonlinear global pandemic machine learning prediction of COVID 19 (2020) <https://doi.org/10.1016/j.chaos.2020.110056>
9. T. Mikami, H. Miyashita, T. Yamada, M. Harrington, Retrospective cohort study on risk factors in patients with COVID-19 In New York City (2020) <https://doi.org/10.1002/jmv.26337>
10. B.S.H.M. Fakhruddin, K. Blandhard, D. Ragupathy, Are we there yet? The transition from response to recovery for the COVID-19 pandemic. *Prog. Dis. Sci.* **7**, 100102 (2020)
11. T.T. Nguyen, Artificial intelligence in the battle against coronavirus (COVID-19): a survey and future research directions (2020) <https://doi.org/10.13140/RG.2.2.36491.238461>
12. I. Al-, M. Alshahrani, T. Almutairi, Building predictive models for MERS-CoV infections using data mining techniques. *J. Inf. Public Health* **9**(6), 744–748 (2016)
13. B. Ghoshal, A. Tucker, Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection. [arXiv:2003.10769](https://arxiv.org/abs/2003.10769) (2020)
14. Y. Zhang, P. Geng, C.B. Sivaparthipan, B.A. Muthu, Big data and artificial intelligence based early risk warning system of fire hazard for smart cities. *Sustain Energy Technol. Assessments* **45**, 100986 (2021)
15. D.V. Pham, G.L. Nguyen, T.N. Nguyen, C.V. Pham, A.V. Nguyen, Multi-topic misinformation blocking with budget constraint on online social networks. *IEEE Access* **8**, 78879–78889 (2020). <https://doi.org/10.1109/ACCESS.2020.2989140>
16. C. Huang, Y. Wang, X. Li, B. Ca, Clinical features of patients infected with 2019 novel coronavirus in Wuhan China. *Lancet* **395**(10223), P497-506 (2020)
17. W. Kong, P.P. Agarwal, Chest imaging appearance of COVID-19 infection. *Radio. Cardio. Imag.* **2**(1) <https://doi.org/10.1148/rccm.2020200028>
18. U.R. Acharya, S.L. Oh, Y. Hagiwara, J.H. Tan, M. Adam, A. Gertych, R. San Tan, A deep convolutional neural network model to classify heartbeats **89**, 389–396 (2017) <https://doi.org/10.1016/j.combiomed.2017.08.022>
19. L. Lan, D. Xu, G. Ye, C. Xia, S. Wang, Y. Li, H. Xu, Positive RT-PCR test results in patients recovered from COVID-19. *JAMA* **323**(15), 1502–1503 (2020). <https://doi.org/10.1001/jama.2020.2783>
20. A. Narin, C. Kaya, Z. Pamuk, Automatic detection of coronavirus disease (COVID-19) using X-ray images and deep convolutional neural networks. [arXiv:2003.10849](https://arxiv.org/abs/2003.10849) (2020)
21. X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. Chestx-ray8: hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Honolulu, HI, 2017), pp. 3462–3471
22. T.Y.J. Naga, T. Maheswarareddy, B. Kushal, CNN for image processing to detect weeds using IOT. *Int. J. Psychosoc. Rehabil.* **24**(8), 1080–1087 (2020)
23. T.Y.J. Naga, C.L. Dondapati, A. Ghosh, Classification of eye disorders based on deep convolutional neural network. *Int. J. Innov. Technol. Explor. Eng. (IJITEE)* **9**(6), 1388–1393 (2020)
24. Li. Yan, H.-T. Zhang, J. Goncalves, An interpretable mortality prediction model for COVID-19 patients. *Nat. Mach. Intell.* **2**, 283–288 (2020)
25. Y. Gao, G.-Y. Cai, Q.-L. Gao, Machine learning based early warning system enables accurate mortality risk prediction for COVID-19. *Nat. Commun.* **11**, 5033 (2020). <https://doi.org/10.1038/s41467-020-18684-2>
26. A.M.U.D. Khanday, S.T. Rabani, Q.R. Khan, N. Rouf, Machine Learning based approaches for detecting COVID-19 using Clinical text data. *Int. J. Inf. Technol.* **12**, 731–739 (2020)
27. Z. Wei, Z. Bing, J. Jiu, Y. Xue, Logistic regression analysis of death risk factors of patients with severe and critical coronavirus disease 2019 and their predictive value (2020) <https://doi.org/10.3760/cma.j.cn121430-20200507-00364>
28. N. Barda, D. Riesel, A. Akriv, J. Levy, U. Finkel, G. Yona, D. Greenfeld, S. Sheiba, J. Somer, E. Bachmat, G.N. Rothblum, U. Shalit, D. Netzer, R. Balicer, N. Dagan, Developing a COVID-19 mortality risk prediction model when individual-level data are not available. *Nat. Commun.* **11**, 4439 (2020)

29. F. Zhou, Y. Ting, D. Ronghui, G. Fan, Y. Liu, Z. Liu, Artificial intelligence for the detection of COVID-19 pneumonia on chest CT using multinational datasets. *Nat. Commun.* **11**, 4080 (2020)
30. D. Assaf, Y. Gutman, Y. Neuman, G. Segal, S. Amit, S. Gefen-Halevi, N. Shilo, A. Epstein, R. Mor, A. Biber, G. Rahav, I. Levy, A. Tirosh, Utilization of machine-learning models to accurately predict the risk for critical COVID-19. *Nat. Commun.* **11**, 5033 (2020)
31. L.J. Muhammad, M.M. Islam, S.S. Usman, S.I. Ayon, Predictive Data Mining models for novel coronavirus infected patient's recovery. *SN Comput. Sci.* **1**, 206 (2020)
32. Y. Xu, Y. Qian, Q. Gu, J. Tang, Relationship between D-dimer concentration and inflammatory factors or organ function in patients with coronavirus disease 2019 **32**(5), 559–563 (2020). <https://doi.org/10.3760/cma.j.cn121430-20200414-00518>
33. Y. Liu, Z. Wang, R. Jingjing, Y. Tian, M. Zhou, T. Zhou, K. Ye, Y. Zhao, Y. Qiu, J. Li, A COVID-19 risk assessment decision support system for general practitioners: design and development study. *J. Med. Internet Res.* **22**(6), e19786 (2020)
34. O. Gozes, M. Frid-Adar, H. Greenspan, P.D. Browning, H. Zhang, W. Ji, A. Bernheim, and Siegel, Rapid AI development cycle for the Coronavirus (COVID-19) pandemic. [arXiv:2003.05037](https://arxiv.org/abs/2003.05037) (2020)

# Neural Network for 3-D Prediction of Breast Cancer Using Lossless Compression of Medical Images



P. Renukadevi and M. Syed Mohamed

**Abstract** Neural network (NN) technology plays a vital role to predict the breast cancer for women. All existing system has to find the breast cancer with tissue in the breast of the women. In our proposed system, it is mainly focused on magnetic resonance imaging (MRI) scanning image of breast and finds the tissue size using 3-D image lossless compression in NN. The main feature of this paper is the use of 3-D predictor to classify the images and using NN technique to find the accuracy level of breast cancer.

**Keywords** 3-D predictor · Lossless compression · MRI image · Neural network

## 1 Introduction

Breast cancer prediction is vital role-playing in machine learning technique. So that we include lossless compression in medical image to predict the breast cancer in machine learning technique. It is the most common cancer for women's in India, and account ratio is 14% of cancers in women. It will affect any age bend but early-stage affect's thirty age women and peak age is 50–64 age [1, 2]. Initial stage to found it is curable but peak stage crossed sudden death happen. So, it is more dangerous disease for women's; that is, reason mainly we are focused to predict the cancer level through MRI image to scan the breast of women using lossless compression of medical image and NN method.

This paper is organized as follows. Section 2 reviews the existing algorithm in lossless compression and ANN technique. Section 3 detail explains about the proposed system to find the accuracy level of breast cancer. Section 4 shows experimental result of proposed system.

---

P. Renukadevi (✉) · M. Syed Mohamed

Department of Information Technology, Sri Ram Nallamani Yadava College of Arts and Science (Affiliated to Manonmaniam Sundaranar University, Tirunelveli), Tenkasi, India

## 2 Algorithm for Lossless Compression and Neural Network of Breast Cancer

In “Artificial Neural Network for Prediction of Breast Cancer,” Singhal and Pareek [3] to predict a breast cancer is using back-propagation algorithm. In the proposed work, feed-forward back-propagation algorithm is to provide the best classifiers to predict the accuracy level of breast cancer.

“Lossless Compression of Medical Images Using 3-D Predictors,” Luís et al. [4] main feature of the work is the use of 3-D predictor and 3-D block octree for partitioning and classification of the image and increases the ratio of the image quality to analysis.

“Artificial Neural Networks in Image Processing for early Detection of Breast Cancer,” Mehdy et al. [5–7] this paper is discussed the all NN methods and how to improve the accuracy and efficient to predict the breast cancer in medical imaging field.

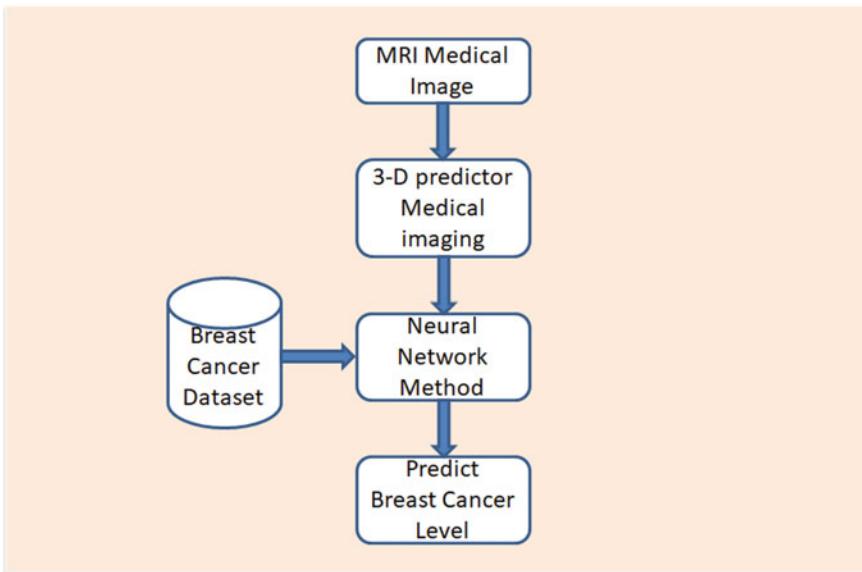
“Prediction of Breast Cancer Using Support Vector Machine and K-Nearest Neighbors,” Islam et al. [8] the proposed model provides more efficient and accuracy provides to predict the breast cancer and it is very helpful to find the breast cancer for medical staffs’ also.

“A Study on Prediction of Breast Cancer Recurrence Using Data Mining Techniques,” Ojha and Goel [9, 10] in this paper proposed with method is efficient and provides accuracy in breast cancer prediction. So, they are concluded like classification method is very efficient and accuracy to find the breast cancer prediction.

## 3 Proposed System

In this proposed system, we are combined with major technology: (1) Lossless compression of medical image and (2) neural network in classification method. We have five modules in our proposed system:

1. MRI Medical Image.
2. 3-D Predictor Medical Image.
3. Neural Network Method.
4. Breast Cancer Dataset.
5. Predict the Breast Cancer Level.



**Fig. 1** Architecture of proposed system

### **3.1 MRI Medical Image**

MRI medical image is a method to scan the human body, and in our proposed system, we are mainly focusing on only breast cancer patients. Who has taken the scanning the breast cancer is the input of our proposed system.

### **3.2 3-D Predictor Medical Imaging**

In this module, we are referred [4] in particular focusing on how to divide the image into blocks and pixels. Based on this module, only our proposed system provides more efficient and accuracy as compared to existing algorithms. Usage of this module is to perform and motivate into partitioning scheme to provide block sizes from  $32 \times 32 \times 32$  to  $2 \times 2 \times 2$ , and also, this module provides better image pixels to find the accuracy and improves the efficient of proposed system (Fig. 1).

### **3.3 Neural Network**

Neural network is one of the best methods to find the accuracy and efficient of the model. In our proposed system, we have predicted the breast cancer level of human

begins. In this module, we have two inputs: (1) Dataset of breast cancer and (2) 3-D medical image data. We are using support vector machine (SVM) to predict the accuracy level and improve the efficient use in medical staff.

### **3.4 Breast Cancer Dataset Description**

In this, we have six attributes to find the accuracy level of breast cancer for human begins. Attribute names are Tumor Size, Inv-Nodes, Node-Caps, Deg-malig, Irradiate, and Class. In this, six attributes each and every attributes is playing vital role to predict the accuracy level of breast cancer for human begins.

#### **3.4.1 3-D Medical Image Data**

In this module, basic image is compressed and reduced the size into  $2 \times 2 \times 2$ . Based on the reduced image, we are finding the level of breast cancer and accuracy of the breast cancer of human begins.

### **3.5 Support Vector Machine Algorithm**

SVM is one of the classifier algorithms that can be defined by separating hyperplane. It has n-dimensional space, and hyperplane was notion us  $(n-1)$  dimension with flat subspace and no need to pass origin dimension. The hyperplane is not visualized in higher dimension but the notion of an  $(n-1)$  dimensional flat subspace still applies [8]. We are using this technique to improve the efficiency of proposed system. Because in Medical Image data have only reduced pixel size and dataset have some trained data is available.

$$\alpha_0 + \alpha_1 b_1 + \alpha_2 b_2 + \dots + \alpha_x b_x = 0 \quad (1)$$

where  $\alpha_0, \alpha_1, \alpha_2, \dots, \alpha_x$  are hypothetical values and  $b_1, b_2, \dots, b_x$  are data points in sample space with “x” dimension.

## **4 Experimental Results**

We have proposed the SVM-based 3-D predictor lossless compression technique. We have trained data with SVM method and test data is also using SVM method to find the accuracy level of breast cancer based on 3-D predictor value [3]. In [3],

value is “3.495” is 16-bit depth content compression value, this value and our test data we are supplied to SVM to predict the level of accuracy.

We have some performance metric indices like True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN).

- True Positive (TP) → Correctly Identified
- True Negative (TN) → Incorrectly Identified
- False Positive (FP) → Correctly Rejected
- False Negative (FN) → Incorrectly Rejected

We have measuring the performance of proposed system using below formulas:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (2)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

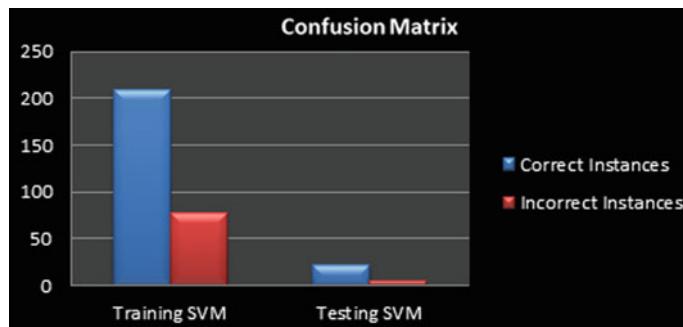
$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{P} + \text{N}} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (4)$$

$$\text{F - Measure} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

$$\text{MCC} = \frac{(\text{TP} \times \text{TN}) - (\text{FP} \times \text{FN})}{\sqrt{(\text{TP} + \text{FP})(\text{TP} + \text{FN})(\text{TN} + \text{FP})(\text{TN} + \text{FN})}} \quad (6)$$

Total trained data instance is 286 in that correctly identified instances 209 and remaining are incorrect instance. Based on this, we are generating the confusion matrix with training and testing data (Fig. 2).

We have using Weka tool to visualize the recurrence-events and no-recurrence-events for the attribute-based.

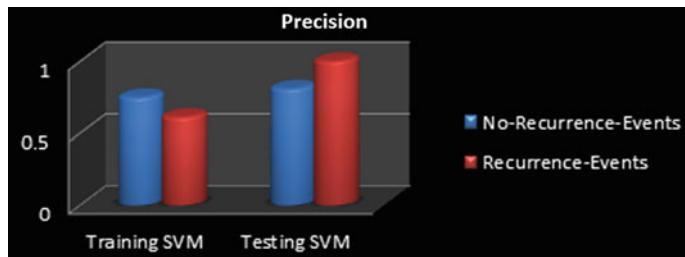


**Fig. 2** Confusion matrix for proposed system

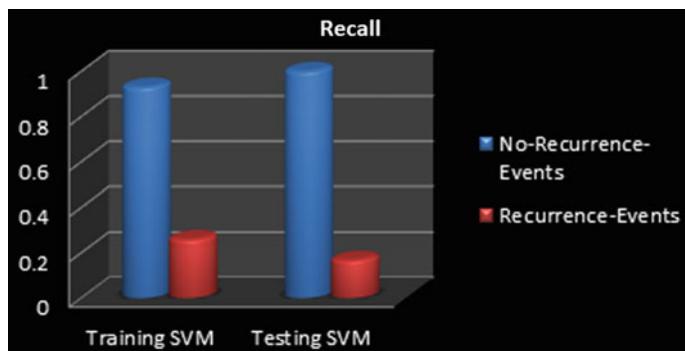
Above all, visualize verified with both test data and training data with SVM technique using neural network.

## 5 Conclusion

In this proposed system, we have implemented 3-D predict algorithm to reduce the pixel size of original image because of original image takes long-time process to find the accuracy level of breast cancer. So that we implemented 3-D predict algorithm with neural network with SVM technique to classify and better accuracy to predict the breast cancer level. We have taken the wisconsin breast cancer dataset to test our proposed system and provide the better accuracy level. We have proposed better accuracy level in precision and recall method that represents in Figs. 3 and 4 with testing data. Finally, we conclude that we have taken particularly in one domain in medical field, and in the future, we can apply all the field and provide better accuracy level (Figs. 5, 6, 7, 8, and 9).



**Fig. 3** Architecture of proposed system



**Fig. 4** Recall of proposed system

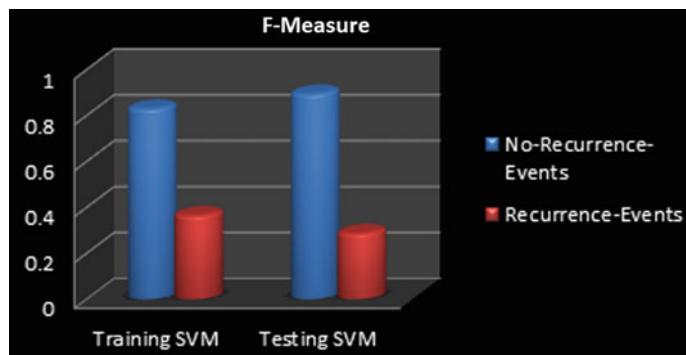


Fig. 5 F-Measure of proposed system

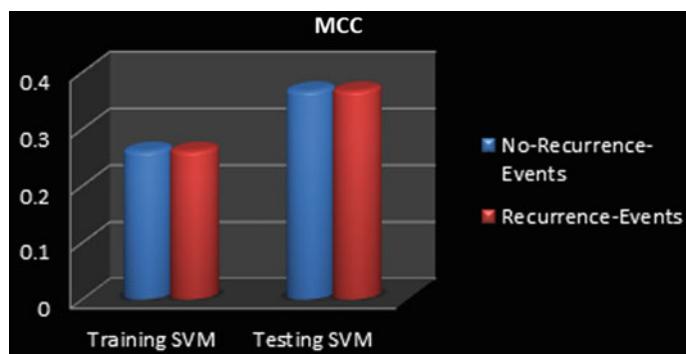


Fig. 6 MCC of proposed system

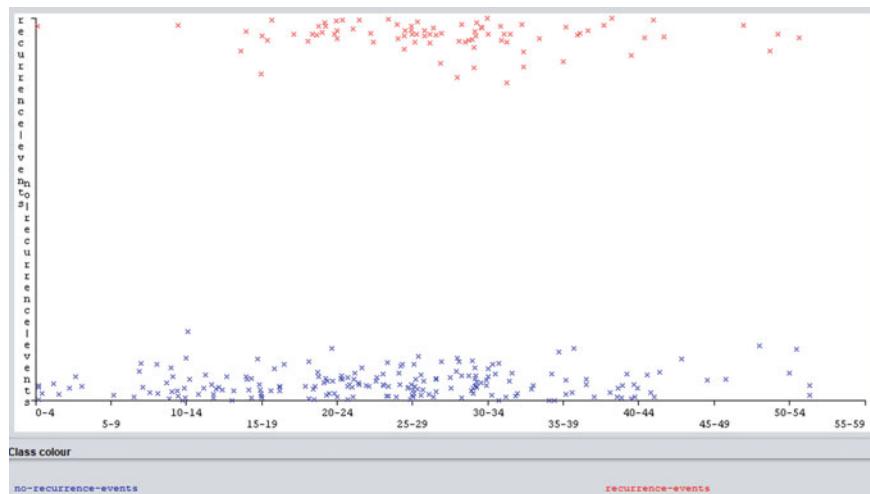
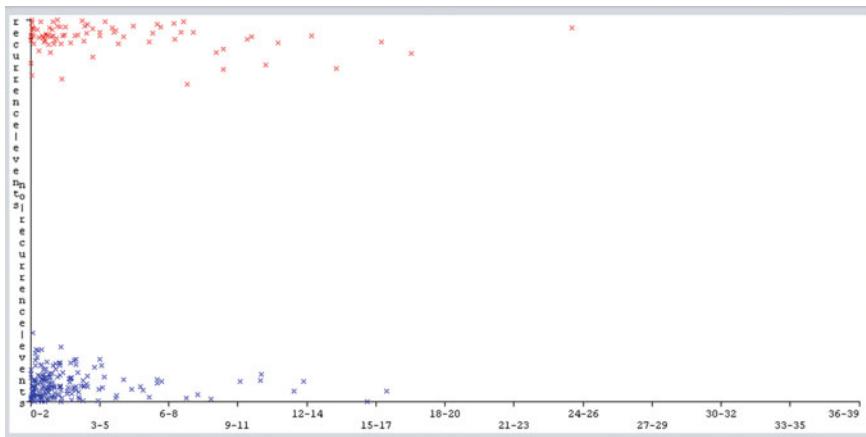
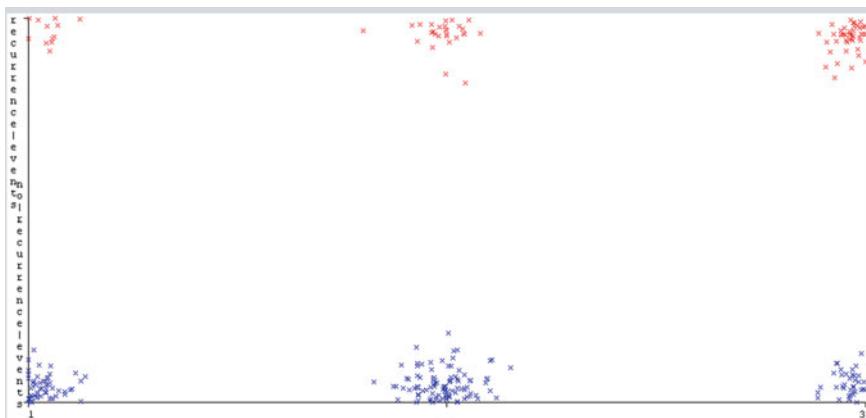


Fig. 7 X-axis for tumor size and Y-axis for event class



**Fig. 8** X-axis for inv-nodes and Y-axis for event class



**Fig. 9** X-axis Deg-Malig and Y-axis event class

## References

1. J. Ferlay, I. Soerjomataram, M. Ervik, R. Dikshit, S. Eser, C. Mathers, F. Bray, LOBOCAN 2012 v1.0, in *Cancer Incidence and Mortality Worldwide: IARC CancerBase* vol. 11. (International Agency for Research on Cancer, Lyon, France, 2013)
2. F. Bray, J.S. Ren, E. Masuyer, J. Ferlay, Global estimates of cancer prevalence for 27 sites in the adult population in 2008. *Int. J. Cancer* **132**(5), 1133–1145 (2013)
3. P. Singhal, S. Pareek, Artificial neural network for prediction of breast cancer, in *IEEE Xplore* Part Number: CFP18OZV-ART; ISBN:978-1-5386-1442-6
4. L.F.R. Luís, N.M.M. Rodrigues, L.A. Da Silva Cruz, S.M.M. De Faria, Lossless compression of medical images using 3-D predictors *IEEE Trans. Med. Imag.* **36**(11) (2017)

5. G.R. Nitta, T. Sravani, S. Nitta, B. Muthu, Dominant gray level based K-means algorithm for MRI images. *Heal. Technol.* **10**(1), 281–287 (2019). <https://doi.org/10.1007/s12553-018-00293-1>
6. N.T. Le, J.-W. Wang, C.-C. Wang, T.N. Nguyen, Automatic defect inspection for coated eyeglass based on symmetrized energy analysis of color channels. *Symmetry* **11**, 1518 (2019). <https://doi.org/10.3390/sym11121518>
7. M.M. Mehdy, P.Y. Ng, E.F. Shair, N.I. Md Saleh, C. Gomes, Artificial neural networks in image processing for early detection of breast cancer. *Hindawi Comput. Math. Meth. Med.* **2610628**, 15 (2017). <https://doi.org/10.1155/2017/2610628>
8. M.M. Islam, H. Iqbal, M.R. Haque, M.K. Hasan, Prediction of breast cancer using support vector machine and K-nearest neighbours, in *2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)*. (Dhaka, Bangladesh, 2017)
9. U. Ojha, S. Goel, Study on prediction of breast cancer recurrence using data mining techniques (IEEE, 2017). 978-1-5090-3519-9/17/\$31.00\_c
10. G. James, D. Witten, T. Hastie, R. Tibshirani, An introduction to statistical learning, 1st ed., (2013)
11. H.B. Burke, P.H. Goodman, D.B. Rosen, D.E. Henson, J.N. Weinstein, F.E. Harrel, D.G. Ostwick, Artificial neural networks improve the accuracy of cancer survival prediction. *Cancer* **79**(4), 857–862 (1997)

# An Investigation of Paralysis Attack Using Machine Learning Approach



S. Surya and S. Ramamoorthy

**Abstract** Many people in today's world suffer from paralysis and most of the paralytic patients are dependent on caretakers. As a result, many survivors have lost some of their abilities that are required to perform daily tasks. Paralysis is a condition in which there is impairment of one or more muscles in the body. In order to assist these patients, fingers of the hand play a major role. The main aim is to achieve light weight; comfortable gloves which can allow finger movement that has lost strength or nerve control. The soft hand glove could be used as an assistance to help the patient grip, pick up and lift objects. Using this glove, the patient gradually starts to restore the functionality of his hand. In our literature survey found some papers that use various types of methods using machine learning algorithms to assist paralyzed patients to address the challenges toward its research directions.

**Keywords** Paralysis · Assistive device · Glove · Machine learning algorithm · Objects

## 1 Introduction

Most of the severe strokes are acting disability conditions up to 80% of survivors are able to suffer from upper extremity disorders to have a serious effect on the ability of a person who performs everyday tasks and affects their quality of life. Attaining and grabbing objects is a part of many everyday tasks involving the use of the upper extremity. Deficits in approaching and gripping are normal after stroke. In this paper discussed the stroke, types of stroke, paralysis attack and machine learning algorithm.

---

S. Surya (✉)

Research Scholar, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai, India

S. Ramamoorthy

Associate professor, Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai, India

e-mail: [ramamoos@srmist.edu.in](mailto:ramamoos@srmist.edu.in)

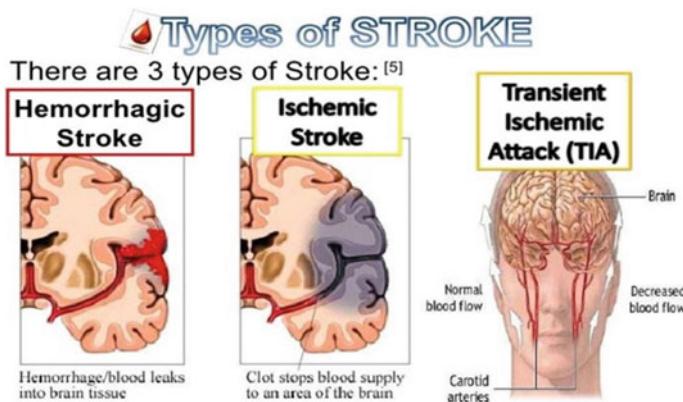
## 1.1 Stroke

A stroke results in a lack of vascular system which triggers the corresponding blood vessels to be cut off through the supply of nutrients and oxygen. If tissue is shut off from its supply of oxygen for more than 3 to 5 min it begins to die. A stroke would cause loss of balance, sudden chest pain, and speech incapacity, loss of memory and limits of thought, daze-like condition, or death. Stroke affects the people among all age groups. Strokes may occur as hemorrhagic strokes, ischemic strokes or TIAs (Fig. 1).

**Hemorrhagic stroke:** Hemorrhagic stroke occurs whenever a damaged blood vessel breaks up inside the brain. Sudden blood vessel leakage or damage happens. Blood pressure that exits from the blood vessel may also cause damage to the brain tissue around it. Hemorrhagic stroke seems to be the most serious form of brain attack.

**Ischemic stroke:** Ischemic stroke occurs when a blood vessel produces a clot within the brain that decreases blood flow to the brain. A blockage in the brain that occurs in a blood stream is called a “bleeding.” A blockage forming in another part of the body, such as the lining of the neck or heart, and passing into the brain is called an “embolus.” Blood clots commonly occur from a disorder called “atherosclerosis,” the build-up of fatty plaques within the walls of the vessel.

**Transient ischemic attack:** It happens the flow of blood control is cut off for a short period of time to some part of the brain, usually 15 min or less. Although transient ischemic attack is a painless occurrence, it is a significant warning sign that a stroke may follow. Treat a TIA as seriously as a stroke.



**Fig. 1** Types of stroke

## 1.2 *Paralysis Attack*

Paralysis is defined as the complete impaired muscle function in any part of the body. It occurs when the transmission of signals between the muscles and the brain seems to have a problem. Because of paralysis patients are unable to move some of their body parts and for their need or support it is also very difficult for them to speak to another person. A healthy workforce is vital to performing daily tasks. A certain loss in the hand's ability to complete these activities could even decrease significantly one's independence, even sometimes resulting in work, social, and family interaction restrictions. Many injury problems, chronic diseases, and defects, including brain injury, cerebral nerve injury, stroke, and arthritis, may lead to decreased function in the hand. As both a result from aging (e.g., stroke) and medical illnesses, hand function also decreases. It results in a diminished capacity to grasp objects and control them through a variety of sizes, shapes, and weights. Consequently, people with reduced grip strength in the hand function can experience reduced levels of performance, reduced independence in basic daily life activities and decreased quality of life.

Assistive technologies have the ability to enhance hand function and freedom on a regular basis activities and without anyone help. There are several different devices available for helping with or enhancing hand work. Some of the existing devices, however, are only used in medical care or clinics because these devices are very costly, not easy to be using and are too complex to use for practical activities. Using the soft hand glove to support people with reduced hand function will add extra power. The glove is compact which can be used to help the grip over a large variety of gross motor skills. Using glove the rehabilitation process can achieve some of the following functional tasks without human interaction (Table 1).

**Table 1** Functional tasks

S. No.	Task	Description
1	To drink	The individual grabs and opens a bottle of water, fills a glass with some water, closes the water bottle, takes and returns the glass of water to its original place
2	Eating	To prepare pieces of cucumber, the person takes a knife, cucumber, and plate. After slice pieces of cucumber the person returns the blade, cucumber, and plate to start place
3	Cleaning households	The person takes a cloth, wrings the cloth, and cleans the table
4	Reading (and writing)	The person keeps a document in the most affected hand and returns the document to its initial location like reading and writing purposes
5	Dress-up	The person takes off the coat hanger from the jacket, puts on the suit, closes the zip, opens the suit, and returns back to the rack
6	Door	The person removes, closes or opens the door, and returns the key to the initial position

### 1.3 Paralysis Attack: Attack

Patients who may have affected their hand movements due to injuries or nerve-related disorders, such as paralysis and muscular dystrophy, will have an opportunity to regain their hand movements using a new lightweight and advanced treatment device called glove. Using ML algorithms has been implemented to increase recognition performance. Supervised learning algorithms require labeling of the qualified data set, while unsupervised learning algorithms should not need fine-tuning or labeling of data. The machine learning algorithm can predict what that person is attempting to do, and instruct the soft hand device to assist appropriately. The patient wishes to grip the product, soft actuators can be enabled to provide the user's fingers with an approximate rate of adaptive force, the machine learning algorithm to operate and an actuation module to help us move the hand. The current device is a model and we would like to miniaturise it so that it can be conveniently carried by a patient.

## 2 Related Work

### 2.1 Detection of the Intention to Grip

Van Ommeren et al. [1] developed high rate of accuracy percentage is reduced from 96.8 to 83.3%, respectively, transferred these particular—user and group of users worked together. The hand and wrist movement of 10 patients with strokes was measured using SVM classifier and inertial sensor during reach and grasp movement patterns. This strategy is used to actuate grip-supporting devices to help patients with strokes in basic daily living activities. From 300 to 750 ms, subsequent grip was observed.

### 2.2 To Detect the Intention Grasping Particular Stages

De varies et al. [2] clarified the stages for user group accuracy from solo user to group of user classificatory (98.2 and 91.4%, respectively). Six kernels have been examined: initial stage of the Linear kernel secondary stage of Quadratic kernel third stage of Cubic kernel, fourth stage of Fine Gaussian kernel, fifth stage of Coarse Gaussian kernel acted during the progress. The Cubic kernel system was used to perform in single user technology and Fine Gaussian kernel technology carried out for combined group technology. Slightly quicker grasp detection ranges from 600 to 1200 ms. Two IMUs can be used to handle grip-supporting tools to assist hand work during regular operations.

### ***2.3 Learning Continuous Grasp Stability***

Schill and Laaksonen et al. [3] this article concentrated on trying to recognize labeled data stability, using a well-known classifier, support vector machines, and built using the ARMAR-IIIb humanoid hand sensor comprehension data collected. Training data should use all available information to collect from the robotic hand, using both tactile sensor feedback and hand finger configuration in this. No need to know the hand's kinematic specification to determine true touch positions while analytically solving the stability of the grasp. Better and faster decisions on stable grasp. ARMAR-IIIb is a versatile anthropomorphic pneumatic-operating hand.

### ***2.4 Comparing Recognition Methods***

Leon et al. [4] have developed and tested various methods of identifying 4 different grip positions performed when wearing the SCRIPT orthose. Considering the support vector machine (SVM), this approach achieves a success rate of more than 90% with minimal computation power.

### ***2.5 Grasp Controls for Artificial Neural Network and Process***

Gandolla et al. [5] introduced the new technology of an electromyography (EMG) compression process and device constructed on the hand robotic assistive technology. Hand grip operation (i.e., pinch, gripping, grabbing object) was predicted using EMG signals on the surface, recorded at EMG process and used from 10 bipolar a ring arrangement to 2–3 cm gap from the whole concern for neural segments. Instead of three flow controls and EMG technologies are performed in electrical passes. EMG electrical controls are transformed in many segments and auto process identification results are gathered and stored as signals.

### ***2.6 Wearable Soft Robotic Glove***

Van Ommeren et al. [6] as evaluated with the Jebsen-Taylor Hand Function Test. This type of test function achieved a good likelihood of accepting the glove. This glove was evaluated for functional task performance during daily life activities.

## 2.7 Distance Metrics Analysis for EEG Signal

Choong Wen Yean et al. [7] this paper created by K-Nearest Neighbor (KNN) technical and pass metric systems classified in detail. Methods and steps are performed and stored as signals. Comprising input transmitted signals to output signals and stored the basic performances and varied ranges. For emotional electroencephalogram (EEG) classification among stroke and regular people. Alpha, beta, and gamma ranges are transmitted the signal process into analogical systems. The distance analogical metrics are transferred into output signals. Emotional EEG data performed the signal process from varied levels and transmits. The output controls and transmitting powers are changed together in linear process segments and used in automation controls. Euclidean executes by providing reasonable precision, where even the accuracy in all three frequency bands was average [8, 9].

## 2.8 Using ML Approaches to Predict Functional Upper Extremity

Ibrahim Alumbark et al. (2018) this article presents knowledge to enhance the understanding of Upper Extremity's complex relationships between laboratory motor function and home use. The collection of biomechanical factors can be used at home and in production to estimate the amount of parallel arm usage. The result shows that the trunk compensation features received a fine, 84.4% accuracy forecast of Upper Extremity usage at home. And clinical features such as Fugl-Meyer and ARAT ratings can also be used to predict arm usage, depending on biomechanical features. By integrating clinical and biomechanical features with machine learning methods (e.g., random forest algorithm with PCA), Upper Extremity home use is effectively measured with 93.3% accuracy.

## 2.9 Deep Learning Framework for Movement of Arm

Madhuri Panwar et al. [10] a Rehab-Net deep learning system for effectively identifying the three movements of the upper limbs, including forearm lengthening, flexors and motion all through daily activities and without using feature selection [11]. Rehab-Net was able to carry out automated function extracting data (collected from the wrist) on preprocessing stage acceleration and identify the three movements of patients with strokes under sub-naturalistic and naturalistic conditions. This paper obtained an average output around the range of 97.89% from the semi-naturalistic results and exceeding data are transferred into 88.7% for natural algorithms and machining parameters are used. Machining and vector parameters and initial stages

are transferred rate of machining process results recognition takes 48.89, 44.14, and 27.64%.

## **2.10 Brain Stroke Detection**

Gaidhani et al. [12] the aim of the investigation was to identify the machining languages and transfer the CNN brain strokes techniques. Deep learning approaches are carried out MRI scan performances and machining stages. Two different types of fully convolutional neural networks, LeNet and SegNet, signals are used as contact signals and through normal and anomalous images to differentiate abnormal areas. This form of classification model achieves 96–97% accuracy and 85–87% accuracy in the segmentation model.

## **2.11 Automatic Detection of Compensatory Movement Patterns**

In this paper, Sijicai et al. [13] present a compensatory movement pattern recognition device that uses a pressure distribution mattress. Pressure distribution data across all movements was collected and analyzed to generate set of the attributes (average sensor values, lateral pressure center, longitudinal pressure center, left-to-right flow rates, and the front-to-back flow rates) reflecting the information within each occur automatically. To recognize compensatory behavior patterns which indicate strong performance and accuracy, four machine deep learning stages were executed. Both classifiers of k-Nearest Neighbor (kNN) signals and support vector machine (SVM) code have attained large signals into high classification accuracy ( $F1$  Score = 0.934). Identifying reparation through all the stages are managed. The SVM basic techniques are showed in better classification performances and forward methods are taken from deep signals. ( $F1$  Score = 0.933), and scapular elevation results are ( $F1$  Score = 0.881) rotated with deep parameters ( $F1$  Score = 0.854), and signal transferred patterns are classified in learning techniques and output controls.

## **2.12 Learning from ‘Humans How to Grasp**

Santina et al. [14] through this paper, the following goals have been accomplished, such as creating a new deep convolutional neural network that analyzes the moving object and analyses any behavior a person takes to understand the target element,

**Table 2** Classification of assistive devices for paralysis attack

S. No.	Type of paralysis	Assistive device	Type of sensor/data/function	Classifiers	Benefits
1	Ischemic stroke	–	–	LR, DT, SVM, RF, xgboost	Predict the function outcome
2	Stroke		Calculate distance metrics	KNN, EEG (DFA)	High classification rate
3	Localized paralysis (monoplegia)	Robotic hand (arm)	Tactile sensing	SVM	Learning continuous grasp stability
4	Localized paralysis (monoplegia)	Robotic hand (palm & finger)	Tactile feedback & hand kinematics data	SVM	Blind robotic grip stability estimation
5	Localized paralysis (monoplegia)	Soft robotic glove	JebSEN-taylor hand function test	Unsupervised	Support grip strength
6	Chronic upper limb	Robotic hand		ANN EMG based controller	Pinching, grasp an object
7	Stroke or chronic diseases	Soft robotic glove	Tendon-driven mechanism (flexion force)		Supports grip and hand opening
8	Stroke (upper limb)	SEM glove	Inertial sensing, force sensor	SVM	Detect grasp intention

propose and create a massive alternate to methodologies. The next goal is to incorporate them in an interactive robotic system and thoroughly evaluate the proposed 111 independent grip design obtaining an average success rate of 81.1% (Table 2).

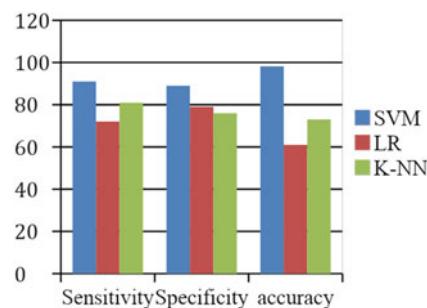
### 3 Comparative Analysis for Different Methods for Glove

Different types of wearable gloves are designed for paralysis attack patients based on machine learning algorithms. The comparative analyses are tabulated in Table 3.

From the above comparison chart of machine learning algorithms the accuracy classification of the given dataset is represented by the overall percentage of data records that are better classified by the classifier techniques. The Specificity and Sensitivity are substitutes to the measure of accuracy that are used to evaluate the classifier's performance. By equating k-nearest process algorithm (k-NN) and Logistic Regression the flow rate of support vector machine (SVM) provides as high accuracy rate (Fig. 2; Tables 4 and 5).

**Table 3** Comparative analysis for different methods for wearable glove

Authors	Methodologies	Recognition accuracy		Sensor
		Single user	Multi user	
Van Ommeren et al. [1]	SVM	96.8%	83.3%	Inertial sensing
de Varies et al. [2]	SVM	98.2%	91.4%	Minimal inertial sensing
Haiming Huang	SVM LR KMC	> 85%		Optical sensor, Pressure sensor, Double sensor
Santina et al. [3]	Deep neural network	81.1%		IM units
Leon et al. [4]	SVM	> 90%		Blending and leaf spring (flex sensor)
Tavakolan et al. [11]	sEMG signals,SVM	Seniors—90.6% and young volunteers-97.6% average accuracy		Torque sensor
Puthenveettil et al. [15]	LDA, CyberGlove	Accuracy improved (Therapy)		Upper extremity rehabilitation-accuracy improved

**Fig. 2** Comparison of machine learning algorithm of stroke**Table 4** Complexity of assistive devices for paralysis attack

Authors	Assistive devices	Complexity
Van Ommeren et al. [1]	SEM glove	No more than two IMU sensor
Gandolla et al. [5]	Hand robotic device	Using a 200 ms window, EMG classifiers failed when tested on neurological patients

## 4 Conclusion

This paper discusses various methodologies for the grasp, release and the lifting objects for assisting paralyzed patients. Interactive objects can be very supportive

**Table 5** Abbreviations and acronyms

ML	Machine learning
ADL	Activities of daily living
TIA	Transient ischemic attack
SVM	Support vector machine
IMU	Inertial measurement unit
kNN	k-Nearest neighbor
EMG	Electromyography
EEG	Electroencephalogram
LDA	Linear discriminant analysis

when it comes to stroke rehabilitation. The experimental results of the machine learning algorithm methods also discussed in this paper. Therefore, various approaches are needed to support these people and it is our responsibility as future engineers to create new technology to help disabled patients. This innovative device is an improvement from traditional methods of manual healing, because it has sensors that interpret muscle signals and conform to the human hand's normal movements, reducing the risk of pain and injury. This glove is also lightweight and waterproof, meaning rehabilitation activities could be performed with greater ease and comfort by patients who recover at home or are bedridden.

## References

1. A.L. van Ommeren, L. Sawaryn, B. Prange-Lasonder, G.B. Buurke, J.H. Rietman, J.S. Velt, AinkPH.detection of the intention to grasp during reaching in stroke using inertial. *IEEE Trans. Neural Syst. Rehabil. Eng.* **27**(10), 2128–2134 (2019)
2. T.J.C. deVries, A.L. van Ommeren, G.B. Prange-Lasonder, J.S. Rietman, P.H. Veltink, Detection of the intention to grasp during reach movements. *J. Rehabil. Assistive Technol. Eng.* **5**, 2055668317752850 (2018)
3. C. Della, V. Arapi, G. Averta, F. Damiani, G. Fiore, A. Settimi, M.G. Catalano, D. Bacciu, A. Bicchi, M. Bianchi, Learning from humans how to grasp: a data-driven architecture for autonomous grasping with anthropomorphic soft hands. *IEEE Robot. Auto. Lett.* **4**(2), 1533–1540 (2019)
4. B. Leon, A. Basteris, F. Amirabdollahian, Comparing recognition methods to identify different types of grasp for hand rehabilitation, in *Proceedings of the 7th International Conference on Advances in Computer-Human Interactions (ACHI '14)* (Barcelona, Spain, 2014), pp. 109–114
5. M. Gandolla et al., Artificial neural network EMG classifier for functional hand grasp movements prediction. *J. Int. Med. Res.* **45**(6), 1831–1847 (2017)
6. A.L. van Ommeren et al., The effect of prolonged use of a wearable soft-robotic glove post stroke—a proof-of-principle, in *Proceeding 7th IEEE International Conference Biomedicine Robotics Biomechatronics (Biorob)* (Enschede, The Netherlands, 2018), pp. 445–449
7. D.V. Pham, G.L. Nguyen, T.N. Nguyen, C.V. Pham, A.V. Nguyen, Multi-topic misinformation blocking with budget constraint on online social networks. *IEEE Access* **8**, 78879–78889 (2020). <https://doi.org/10.1109/ACCESS.2020.2989140>

8. C. Siqi, G. Li, S. Huang, H. Zheng, L. Xie, Automatic detection of compensatory movement patterns by a pressure distribution mattress using machine learning methods: a pilot study. *IEEE Access* **7**, 80300–80309 (2019)
9. C.W. Yean, W. Khairunizam, M.I. Omar, M. Murugappan, B.S. Zheng, S.A. Bakar, Z.M. Razlan, Z. Ibrahim, Analysis of the distance metrics of KNN classifier for EEG signal in stroke patients, in *2018 International Conference on Computational Approach in Smart Systems Design and Applications (ICASSDA)* (IEEE, 2018), pp. 1–4
10. P. Madhuri, D. Biswas, H. Bajaj, M. Jöbges, R. Turk, K. Maharatna, A. Acharyya, Rehab-net: deep learning framework for arm movement classification using wearable sensors for stroke rehabilitation. *IEEE Trans. Biomed. Eng.* **66**(11), 3026–3037 (2019)
11. M. Tavakolan, Z.G. Xiao, C. Menon, A preliminary investigation assessing the viability of classifying hand postures in seniors. *BioMed. Eng. Online* **10**(79) (2011)
12. B.R. Gaidhani, R.R. Rajamenakshi, S. Sonavane, Brain stroke detection using convolutional neural network and deep learning models, in *2019 2nd International Conference on Intelligent Communication and Computational Techniques (ICCT)* (Jaipur, India, 2019), pp. 242–249. <https://doi.org/10.1109/ICCT46177.2019.8969052>
13. S. Ramesh, C. Yaashwanth, B.A. Muthukrishnan, Machine learning approach for secure communication in wireless video sensor networks against denial-of-service attacks. *Int. J. Commun. Syst.* **33**(12) (2019). <https://doi.org/10.1002/dac.4073>
14. C. Della, V. Arapi, G. Averta, F. Damiani, G. Fiore, A. Settimi, M.G. Catalano, D. Bacciu, A. Bicchi, M. Bianchi, Learning from humans how to grasp: a data-driven architecture for autonomous grasping with anthropomorphic soft hands. *IEEE Robot. Automation Lett.* **4**(2), 1533–1540 (2019)
15. S. Puthenveettil, G. Fluet, Q. Qinyin, S. Adamovich, Classification of hand preshaping in persons with stroke using Linear Discriminant Analysis, in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC '12)* (2012), pp. 4563–4566

# A Survey on Feature Selection and Classification Techniques for EEG Signal Processing



K. Saranya, M. Paulraj, and M. Brindha

**Abstract** The electroencephalogram is a test that is used to keep track on the brain activity. These signals are generally used in clinical areas to identify various brain activities that happen during specific tasks and to design brain–machine interfaces to help in prosthesis, orthosis, exoskeletons, etc. One of the tedious tasks in designing a brain–machine interface application is based on processing of EEG signals acquainted from real-time environment. The complexity arises due to the fact that the signals are noisy, non-stationary, and high-dimensional in nature. So, building a robust BMI is based on the efficient processing of these signals. Optimal selection of features from the signals and the classifiers used plays a vital role in building efficient devices. This paper concentrates on surveying the recent feature selection, feature extraction, and classification algorithms used in various applications for the development of BMI.

**Keywords** EEG · Prosthesis · Orthosis · Exoskeletons

## 1 Introduction

Brain–machine interface is an emerging field having diverse applications from rehabilitation to human augmentation. A brain–machine interface is a system that directly communicates with external devices through brain signals. This interface is developed mainly using three steps namely signal collection, signal processing, and interpretation and finally outputting the commands to external devices to act accordingly.

---

K. Saranya (✉) · M. Paulraj

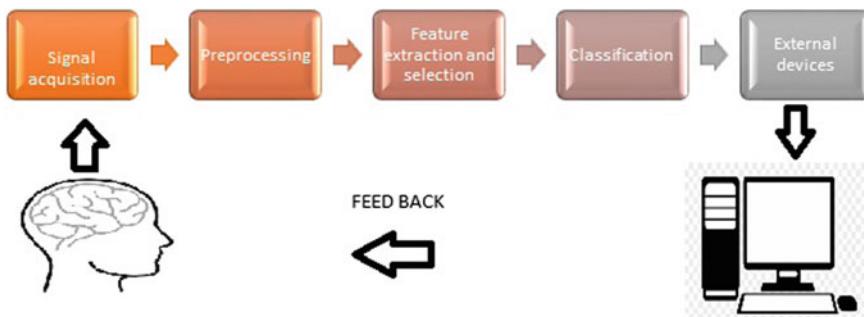
Department of Computer Science and Engineering, Sri Ramakrishna Institute of Technology, Coimbatore, Tamil Nadu, India

M. Paulraj

e-mail: [principal@srit.org](mailto:principal@srit.org)

M. Brindha

Department of Computer Science and Engineering, Chennai Institute of Technology, Chennai, Tamil Nadu, India



**Fig. 1** EEG signal processing steps

The brain signals are measured using three ways namely non-invasive, semi-invasive, and invasive method.

In non-invasive method, the electrical signals are measured by implanting electrodes in specific areas of the brain. This technique detects signals with high accuracy via surgical options, but only trained professionals could carry out this surgery. In semi-invasive method, the electrodes are kept on the surface of the brain to measure the electric impulses that arise from the brain. Electrocorticography (ECog) is used in this type of method. The non-invasive technique is generally carried out by electroencephalography (EEG) in which the electrodes are kept on the scalp to monitor and measure brain signals. This method is cost effective and gives excellent time resolution but spatial resolution is low which can be corrected by various filters. Recently, EEG is gaining more attention in brain–machine interface research activities. So, EEG signals are focused in this survey due to its wide area of usage.

Generally, a BMI is developed through a sequential set of process starting from the data acquisition phase and ending in a useful device having the control signals for a specific purpose. Figure 1 depicts the process that is carried out throughout the development of a brain–machine interface. EEG signal processing is the measure of electrical activity of the brain. Measuring and analyzing these signals can lead to develop various applications like epilepsy detection, sleep disorder detection, alcoholic detection, emotion detection, motor activity classification, visual classification, etc.

## 2 Methodology

The raw EEG signals acquired has to be preprocessed since the signals are combined with noise data (artifacts) such as eye blinks, muscle movements, and heartbeat. There are many techniques to remove the noise from data such as filters can be applied to the data. High pass filters and low pass filters are used to set up a cut-off frequency range between 1 and 90 Hz since in EEG frequencies above it are not

taken into account. After removing the artifacts from the data, the recordings are cut into epochs which will be helpful in selecting features.

The feature extraction and selection step is used to analyze and select specific information from the signals. It is highly complex to recognize the patterns in the acquired signals through naked eye. Hence, algorithms like band power (BP) and frequency representation (FR) are applied on the preprocessed data to recognize the patterns according to the human intent. With the computed feature vectors, classifiers are trained to learn the specific tasks for which the BMI is designed for. This is associated with the commands to the external device. Figure 1 shows the methodological process involved in the EEG signal processing phase. This paper deals with the various feature selection, feature extraction, and classification techniques used in the EEG signal processing.

### 3 Related Works

Jomar et al. [1] developed a brain–machine interface for visual imagery recognition using deep learning methods for feature selection and extraction process thereby reducing the external interference in the selection process. Further for classifying the visual imagery tasks like imagining squares, circles, and triangles, various deep learning models like convolution neural network (CNN), recurrent neural network (RNN), and deep belief networks were used. The results show that deep learning-based brain–machine interface outperformed in performance when compared to conventional classifiers. Among the deep learning classifiers, CNN-LSTM achieved highest performance of about 90.51% precision, and among conventional classifiers, SVM had the highest performance of about 83.7% precision for visual imagery recognition.

Xu et al. [2] focused on using orthogonal regression for feature selection process for the application of emotion recognition. This method preserves more discriminative feature sub-space than classical feature selection methods. The experiments from the recorded dataset have been conducted with the feature selection using orthogonal regression (FSOR) algorithm and compared against relief F and information gain which are two popular feature selection techniques. The classification process was carried out from the trained dataset using linear support vector machine, and the accuracy obtained was about 76.4% having 390 features that this outperforms the other two feature extraction method.

Rammy et al. [3] concentrated on spatiotemporal-based feature extraction using common spatial pattern (CSP) and fast Fourier transform (FFT) filters to obtain the spatiotemporal discriminative features from the motor imagery (MI) tasks. The author also paid main attention on the preprocessing technique to obtain optimal parameters. The experiment was carried out using BCI Competition IV dataset 2a which contains the motor imagery signals. The results were compared against long short-term memory (LSTM) with LDA, SVM, and CNN classifiers. From the comparison

made, LSTM outperforms other classifiers with a kappa value of 0.64 and standard deviation of about 0.109. The author concluded that while using recurrent deep learning technique for classification, the result is better.

Shiam et al. [4] used unsupervised feature selection for the same motor imagery (MI) tasks using BCI Competition III (IV A) dataset. The data is preprocessed using fourth-order Butterworth bandpass filter by selecting four frequency band between 8 and 35 Hz where the motor imagery signals are identified. Common spatial filtering (CSP) technique is used to extract the features where the variance of one class is maximized while the variance of other class is minimized. The extracted features use the proposed unsupervised discriminative feature selection (UDFS) technique to select the dominant features. This selected data are trained using SVM classifier. Results show that the UDFS has achieved a classification accuracy of 89.87% when using this UDFS selection method. This result is considerably high when compared with recently developed algorithm.

Tavares et al. [5] investigated the automatic classification of patients having Alzheimer's disease (AD) and healthy patients (HP) by reducing the EEG features and channels. The dataset was taken from 19 AD patients and 17 healthy subjects. Here, the feature is extracted using power spectrum density (PSD). After this phase, eight different classifiers including regression, SVM and tree were used to classify between the AD and HS people. The results show that SVM classifier outperforms other classifiers in all criteria (with feature selection, without feature selection and in number of features and channels).

Bayatfar et al. [6] proposed a feature extraction method based on time domain to extract the features from the DREAMS Apnea database to get the relevant features for diagnosing sleep apnea syndrome (SAS). Irrelevant features are removed using minimal redundancy maximal relevance (mRMR) algorithm. Finally, random undersampling boosting algorithm is used as learning method in classification process. The experiment was carried against the novel k-fold cross-validation technique and found out that the proposed algorithm works efficiently with 89.9% of classification accuracy and outperforms the existing methods.

Nandy et al. [7] conducted experiment on publicly available CHB-MIT database for automatic seizure detection from non-seizure EEG signals. Initially, the signals were preprocessed using bandpass filter, and the features are extracted from time, spectral, and wavelet domain. Entropy-based features were also extracted. Multi-objective evolutionary algorithm (MOEA) was used to select the features. Finally, support vector machine was used for the classification process. The experiment was conducted against LDA and QLDA and found that SVM shows highest accuracy of about 97.05% than the other two algorithms.

Bhatti et al. [8] selected motor imagery tasks for BCI development. He selected two datasets namely open BCI dataset and emotive Epoc EEG signals. Initially, a filter bank using various frequency cut-offs is used to split the EEG signals into sub-bands. Then, common spatial pattern (CSP) and linear discriminant analysis (LDA) is used to extract the features from the filtered signals. Sequential backward floating selection (SBFS) is used to select the relevant features from extracted features. Finally, radial basis function neural network (RBFNN) is applied as a learning method for

classification. The results shows that the proposed classification algorithm achieves an accuracy of 93.05% and 85% for both datasets and depicts that proposed algorithm outperforms other techniques.

Koch et al. [9] concentrated on the development of an automated classification of Parkinson's disease (PD) patients. Initially, the data was recorded using 125 patients using 21 channels. The recorded EEG signals were preprocessed using bandpass filter of five epochs. The features are extracted based on time series using fast Fourier transform (FFT). A new feature selection algorithm called Boruta is used for selecting the most important features using random forests technique. Again the RF is used for classification process. Finally, the hyper-parameters of the classified model are optimized using Bayesian technique. The result shows that the new optimized automated model built produces an accuracy of about 91% in detecting the PD patients.

Maswanganyi et al. [10] used a publicly available BCI competition IV motor imagery movement dataset. Independent component analysis [ICA] is used for preprocessing technique which removes the unwanted artifacts like physiological movements. The preprocessed signals use wavelet packet transform (WPT) technique to extract the relevant features. From the extracted features, the best subsets are selected using DEFS algorithm which is differential evolution-based channel and feature selection technique. Finally, Naïve Bayes (NB) algorithm is used for the classification process which classifies four different tasks like imagining left hand, right hand, foot, and tongue. Result shows that NB algorithm achieves a classification accuracy of 73.06% and outperforms K-NN algorithm.

Satyam Kumar et al. [11] worked on motor imagery classification using same BCI competition IV dataset using phase difference sequence. Here, bandpass filter is used as preprocessing technique. The feature is extracted using CSP, phase-locking value (PLV), instantaneous phase difference (IPD-CSP) techniques. After this, Lasso algorithm is used for feature selection process. Finally, the selected features are trained and classified using SVM algorithm. Result shows that the proposed IDP-CSP method outperforms existing method.

Pane et al. [12] concentrated on channel selection of EEG for emotion recognition. The experiment was made using publicly available SEED dataset. The data was preprocessed using bandpass filter and short-term Fourier transform (STFT) to remove the artifacts. Differential entropy feature is used for feature extraction here. Then, the channel is selected using stepwise discriminant analysis (SDA) where the most relevant channel alone are taken into account. Finally, the selected features are trained and classified using linear discriminant analysis (LDA). When compared to other channel selection methods when the proposed method is used the classification accuracy obtained was 99.85% when 15 channels were used.

Mzurikwao et al. [13] also focused on channel selection for motor imaginary tasks. Here, the data is taken from four subjects with their consent using 64 channels. It has been considered that when more channels are used the spatial information will be high. The recorded EEG signal is preprocessed using normalization technique. The convolution neural network is used as both feature extraction technique and training model. The accuracy obtained after the use of CNN technique was about 99.77% with 0.1 window size. The result was compared against various classification

techniques like LDA, ANN, and KNN. It has been concluded that the proposed method outperforms the other techniques.

Garro et al. [14] used artificial bee colony algorithm (ABC) for effective channel selection. The author feels that increasing the channels to obtain more spatial information increases the complexity in computation. As an alternate to this problem, selecting most relevant channel that produces significant information is carried out in this paper using ABC algorithm and fractal dimension algorithm. Dataset IVa from BCI competition III was taken for the experiment. It contains motor imagery tasks. The data was preprocessed using fourth-order Butterworth Filter. The fractal dimension method is applied for feature extraction method. Finally, the extracted features are trained using Fisher classifier. The result shows that the proposed methodology provides better accuracy with selected number of channels which is 15 using ABC algorithm and the percentage of accuracy was 86.7% than using large number of channels.

Indah Agustien Siradjuddin et al. and others [15–17] used EEG dataset taken from UCI machine learning repository for alcoholic detection. The detection of alcohol consumed by the subjects using urine or breath test is time limited; hence, the author has chosen to detect using EEG signals. The selected dataset used 64 channels from ten subjects; these signals were preprocessed using independent component analysis technique (ICA). Secondly, the features are extracted using discrete wavelet transform. The extracted features are selected using genetic algorithm (GA). Selected features are trained using back propagation neural network. The results are compared with feature selected dataset and without feature selection. The outcome depicts that while using selected features using GA for training accuracy of the detection was about 79.38% where without feature selection the accuracy was about 75.5%.

Giannakaki et al. [18] focused on classifying various emotion states from the EEG signals. The dataset was taken from ENTERFACE workshop 2006 where three emotions were detected. The dataset was preprocessed by applying bandpass filter. Fifteen channels were taken to reduce the computational complexity. The obtained signals are then used for extracting features and selecting features using various machine learning techniques. Finally, the selected features are trained and classified. It has been observed that random forest classifier attains the highest accuracy of 75.59% while classifying the emotional states.

Lahani et al. [19] conducted experiment on DEAP (Dataset for Emotion Analysis using Physiological Signals) dataset to analyze two emotional states of brain which is happy and sad. The data use fast Fourier transform (FFT) for preprocessing and frequency cepstral coefficient (FCC) for extracting the features. This technique is compared against Kernel density estimation (KDE) technique. These extracted signals are then classified using K-nearest neighbor (KNN) classifier. The results show that the accuracy obtained, while using FCC technique outperforms KDE with an accuracy of 90%.

Alharbi et al. [20] concentrated on classifying on single trial using event-related potential (ERP). So, the author classified the EEG signals obtained by seeing RGB colors using various methods. Here, the signals are decomposed using empirical mode decomposition (EMD), and the features are extracted using various methods

namely event-related spectral perturbations, target mean, autoregressive, and EMD residual to find the best method to classify colors dataset. A proposed feature selection algorithm is used to the extracted dataset, and a comparative study is made out of it. The results show that the proposed feature selection algorithm with SVM classifier achieves higher accuracy than other methods. It is also found out that execution time while using color stimulus is very much less when compared to other ERP's.

Datta et al. [21] obtained the visual imagery hand movement dataset taken from 8 subjects using 14 channels. The acquired signals are extracted using wavelet transform. Then, the best features are selected using principal component analysis. The features extracted from different regions of the brain are mapped together using regression analysis on artificial neural network. Finally, the selected features are trained using back propagation learning model to predict the movements. After various comparative studies, the results show that Levenberg–Marquardt optimization technique performed best in terms of mean squared error.

Rivero et al. [22] analyzed publicly available epilepsy dataset and tried to classify between seizure activity and without seizure activity. The signals are preprocessed using fast Fourier transform (FFT). After that, the features are extracted using genetic algorithm. Finally, the extracted features are classified using KNN algorithms. The results show that while using the GA algorithm along with KNN classifier the classification accuracy obtained was about 98.73% while SPWD time frequency analysis along with ANN produced an accuracy of about 97.7%. It has been inferred that when using GA the accuracy was far better in feature selection process.

Hassan et al. [23] used the dataset taken from Dataset III (2003), Dataset IIIb, and Dataset bci7 to classify the motor imagination tasks. Initially, the signals are preprocessed using FFT to obtain time and frequency domain information. After extracting the features using time and frequency domain information, feature subset selection (FSS) is used to select the relevant features. Finally, a multilayer perceptron back propagation neural network model is used to classify the selected features. The classification accuracy reached about 100%, while using Dataset III and IIIb and 97–99% when bci7 dataset was used.

## 4 Conclusion

Brain–machine interface is a fast booming area of research and has a wide variety of scope for the society. It helps in people having physical inability. By surveying above papers, it has been concluded that proper channel selection, band selection, and feature selection algorithms play an important role in development of effective BMI. From Table 1, it has been concluded that when trying to develop a BCI for motor imagery tasks, the frequency band between 8 and 30 Hz of full band is taken into account for preprocessing, and the bands selected were alpha/mu and beta bands for motor imagery tasks. It has also been inferred that while applying deep learning techniques for feature extraction and selection, the classification accuracy achieved was higher. From Table 2, it has been inferred that for emotion recognition, a full

**Table 1** Comparisons for motor imagery tasks

Preprocessing	Feature extraction	Feature selection	Classification	Accuracy
Temporal Segmentation, [3] 22 channels	CSP-FFT	CSP-FFT	Convolution neural network-long short-term memory network (CNN-LSTM)	0.64 KAPPA
Fourth-order Butterworth bandpass filter [4] 118 channels	Spatial filtering	Unsupervised discriminative feature selection (UDFS)	Support vector machine (SVM)	89.87
Filter bank [8] 118 channels	LDA and CSP	Sequential backward floating selection	Radial basis function neural network (RBFNN)	93.05
Independent component analysis (ICA) [10] 22 channels	Wavelet packet transform (WPT)	Differential evolution-based feature selection (DEFS)	Naïve Bayes (NB)	73.06
Bandpass filter [11] 4 channels	Instantaneous phase difference (IPD)	IPD-CSP	Linear support vector machine (LSVM)	73.77
Min-max normalization [13] 64 channels	CNN	CNN	Convolution neural network (CNN)	99.7
Fast Fourier transform (FFT) [21] three channels	Time and frequency domain	Information feature subset election (FSS)	Multilayer perceptron back propagation neural network	97.77

**Table 2** Comparison for emotion classification

Preprocessing	Feature extraction	Feature selection	Classification	Accuracy
Independent component analysis (ICA) [2] 19 channels	Autoregressive model	Feature select ion with autoregression (FSOR)	Linear support vector machine	76.4%
Bandpass filter [12] 62 channels	Differential entropy	Stepwise discriminant analysis (SDA)	Linear discriminant analysis (LDA)	99.85
Bandpass filter [16] 15 channels	Machine learning techniques	Machine learning techniques	Random forest	75.12
Bandpass filter [17] 32 channels	Frequency cepstral coefficient (FCC)		K-nearest neighbor (KNN)	90%

band of range 1–50 Hz was taken into account and the features were extracted from alpha, beta, theta, gamma, and delta bands. It has been inferred that usage of SDA and LDA works best in classifying emotions and achieves an accuracy of about 99.85%. From Table 3, it has been inferred that for visual imagery of shapes, gamma band was chosen to achieve an accuracy of 89.44%. Delta and theta bands were chosen when color stimulus was used. SVM were most popularly chosen to achieve a high accuracy rate. From Table 4, it has been inferred that for seizure detection a full band of 0.5–30 Hz is taken into consideration and while using evolutionary algorithms the classification accuracy is higher than other techniques.

**Table 3** Comparison for visual imagery tasks

Preprocessing	Feature extraction	Feature selection	Classification	Accuracy
Built in fifth order sync filter [1] 14 channels	CNN-LSTM	CNN-LSTM	Support vector machine (SVM)	89.44%
Filter [18] 4 channels	Target mean and EMD residual	Recursive feature elimination algorithm (RFEA)	SVM	97.94
Spatial filtering by common average referencing (CAR) [19] 14 channels	Wavelet approximate	Principal component analysis (PCA)	Backpropagation artificial neural network (BPANN)	–

**Table 4** Comparison for neural disease prediction

Preprocessing	Feature extraction	Feature selection	Classification	Accuracy
Independent component analysis [5] 32 channels	Power spectral density (PSD)	Backward wrapper method	Support vector machine (SVM)	95.6
Bandpass filter [6] 1 channel	Time domain based extraction	Minimal redundancy maximal relevance (mRMR)	Random undersampling boosting (RUSBoost)	88
Bandpass filter [7] 4 channels	Adaptive synthetic algorithm (ADASYN)	Multiobjective evolutionary algorithm (MOEA)	SVM	97.05
Visual Inspection [9] 21 channels	Automatic feature extraction using FFT	Boruta algorithm	Random forest (RF)	91
Butterworth bandpass filter [14] 118 channels	Fractal dimension method	Ant bee colony (ABC) optimization technique	Fisher	86.98

(continued)

**Table 4** (continued)

Preprocessing	Feature extraction	Feature selection	Classification	Accuracy
Independent component analysis (ICA) [15] 64 channels	Discrete wavelet transform (DWT)	Genetic algorithm	Backpropagation neural network (BNN)	79.38
Fast Fourier transform (FFT) [20] 1 channel	Genetic algorithm	GA	K-nearest-neighbor (KNN)	98.5

## References

1. A.J.F. Castro, J.N.P. Cruzit, J.J.C. De Guzman, J.J.T. Pajarillo, Development of a deep learning-based brain computer interface for visual imagery recognition, in *2020 16th IEEE International Colloquium on Signal Processing & its Applications (CSPA 2020)* (Langkawi, Malaysia, 28–29 Feb. 2020)
2. X. Xu, F. Wei, Z. Zhu, J. Liu, X. Wu, Eeg feature selection using orthogonal regression: application to emotion recognition 978-1-5090-6631-5/20/\$31.00 ©2020 IEEE
3. S. Ali Rammy, M. Abrar, S. Jabbar Anwar, W. Zhang, Recurrent deep learning for EEG-based motor imagination recognition 978-1-7281-4235-7/20/\$31.00 ©2020 IEEE
4. A. Al Shiam, T. Tanaka, Md.R. Islam, Khademul Islam Molla “978-1-7281-2297-7/19/\$31.00 ©2019 IEEE DOI <https://doi.org/10.1109/CW.2019.000047>
5. G. Tavares1, R. San-Martin, J'essica N. Ianof, Renato Anghinah, F.J. Fraga, Improvement in the automatic classification of Alzheimer's disease using EEG after feature selection, in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC) Bari*, Italy. October 6-9
6. S. Bayatfar, S.Seifpour, M. Asghari Oskoei, A. Khadem, An automated system for diagnosis of sleep apnea syndrome using single-channel EEG signal, 978-1-7281-1508-5/19/\$31.00 c 2019 IEEE.
7. A. Nandy, S. Alam, M. Ashik Alahe, A.-A.-Nahid, S.M. Nasim, Md. Abdul Awal, Feature extraction and classification of EEG signals for seizure detection, 978-1-5386-8014-8/19/\$31.00 ©2019 IEEE,
8. M. Hamza Bhatti, J. Khan, M. Usman Ghani Khan, R. Iqbal, M. Aloqaily, Y. Jararweh, B. Gupta, Soft computing based EEG classification by optimal feature selection and neural networks 1551–3203 (c) 2019 IEEE.
9. M. Koch,V. Geraedt, H. Wang Leiden, M.T. Leiden, T.Back Leiden, Automated machine learning for EEG-based classification of parkinson's disease patients, in *2019 IEEE International Conference on Big Data (Big Data)*
10. C. Maswanganyi, C. Tu, P. Owolawi, S. Du, Discrimination of motor imagery task using wavelet based EEG signal features, 978-1-5386-6477-3/18/\$31.00 ©2018 IEEE
11. S. Kumar, T. Reddy, L. Behera, EEG based motor imagery classification using instantaneous phase difference sequence, in *2018 IEEE International Conference on Systems, Man, and Cybernetics*, 2577-1655/18/\$31.00 ©2018 IEEE <https://doi.org/10.1109/SMC.2018.00094>
12. E. Septiana Pane, A.D. Wibawa, M.H. Purnomo, Channel selection of EEG emotion recognition using stepwise discriminant analysis, in *2018 CENIM*
13. D. Mzurikwao, C. Siang Ang, O. Williams Samuel, M. Grace Asogbon, X. Li, G. Li, fficient channel selection approach for motor imaginary classification based on convolutional neural network, in *Proceedings of the 2018 IEEE International Conference on Cyborg and Bionic Systems Shenzhen*, China, October 25–27 (2018)

14. B.A. Garro, R. Salazar-Varas, R. A. Vazquez, EEG channel selection using fractal dimension and artificial bee colony algorithm, in *IEEE Symposium Series on Computational IntelligenceSSCI 2018*
15. B.H. Liu, N.T. Nguyen, V.T. Pham, An efficient method for sweep coverage with minimum mobile sensor, in *2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing* (Kitakyushu, Japan, 2014), pp. 289–292. <https://doi.org/10.1109/IIH-MSP.2014.78>
16. B. Muthu, C.B. Sivaparthipan, G. Manogaran, R. Sundarasekar, S. Kadry, A. Shanthini, A. Dasel, IOT based wearable sensor for diseases prediction and symptom analysis in healthcare sector. *Peer-to-Peer Netw. Appl.* **13**(6), 2123–2134 (2020). <https://doi.org/10.1007/s12083-019-00823-2>
17. I.A. Sirajuddin, R. Septasurya, M.K. Sophan, N. Ifada, A. Muntas, Feature selection with genetic algorithm for alcoholic detection using electroencephalogram, in *2017 International Conference on Sustainable Information Engineering and Technology (SIET)*
18. K. Giannakaki, G. Giannakakis, C. Farmaki, V. Sakkalis, Emotional state recognition using advanced machine learning techniques on EEG data, in *2017 IEEE 30th International Symposium on Computer-Based Medical Systems*
19. P. Lahane, M. Thirugnanam, A novel approach for analyzing human emotions based on electroencephalography (EEG), in *International Conference on Innovations in Power and Advanced Computing Technologies [i-PACT2017]*
20. E.T. Alharbi, S. Rasheed, S.M. Buhari, Feature selection algorithm for evoked EEG signal due to RGB colors, in *2016 9th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics(CISP-BMEI 2016)*
21. S. Datta, A. Khasnobish, A. Konar, D.N. Tibarewala, A.K. Nagar EEG based artificial learning of motor coordination for visually inspired task using neural networks” EEG based artificial learning of motor coordination for visually inspired task using neural networks , in *2014 International Joint Conference on Neural Networks (IJCNN) July 6–11, 2014, Beijing, China*
22. D. Rivero, Using genetic algorithms and k-nearest neighbor for automatic frequency band selection for signal classification. *IET Sig. Proc. Inst. Eng. Technol.* **6**(3), 186–194. <https://doi.org/10.1049/iet-spr.2010.0215>
23. M.A. Hassan, A.F. Ali, M.I. Eladawy, Classification of the imagination of the left and right hand movements using Eeg, in Proceedings of the 2008 IEEE, Cibec'08, 978-1-4244-2695-9/08/\$25.00 ©2008 IEEE

# Deployment of Sentiment Analysis of Tweets Using Various Classifiers



Shatakshi Brijpuriya and M. Rajalakshmi

**Abstract** Twitter is a social media forum that permits individuals to share and express their perspectives regarding the matter and post messages. A great deal of examination has been done on the sentiment analysis of Twitter data. Our proposed methodology aims to incorporate the techniques of sentimental analysis, assessing the polarity of the tweets, and improving the accuracy of the model. This paper includes classifying of tweets into sentiments: positive and negative. For this paper, we have used different data preprocessing techniques for cleaning up the data by eliminating stop word, slangs, emoticons, and hashtags. To build the productivity and the precision of the model, we have used term frequency inverse data frequency, count vectorizer, stemming, and N-gram features. Several techniques have been used currently for sentiment analysis which is discussed in brief in this paper. Out of which, we have used supervised learning approach and compared different algorithms such as Naive Bayes, support vector machines, linear regression, decision tree, XGBoost, and random forest.

**Keywords** Twitter · Sentiment analysis · Machine leaning algorithm

## 1 Introduction

Analysis of sentiment is the work of seeking people's views and inclination on the specific topics of interest. Customer opinion matters a lot and it also influences the decision-making process. Social networking forums like Facebook, Twitter, etc., have become a place where people share their perspective. It also contains huge unstructured data which include feedback and responses. There are many applications of sentiment analysis of tweets such as stock market analysis of a company and movie recommendation. Tweet's emotion can be separated into numerous classes

---

S. Brijpuriya (✉) · M. Rajalakshmi  
SRM Institute of Science and Technology, Chennai 603203, India

M. Rajalakshmi  
e-mail: [rajalakm2@srmist.edu.in](mailto:rajalakm2@srmist.edu.in)

like positive, negative, supremely positive, supremely negative, and neutral. We have broadly categorized these sentiments as positive and negative [1]. There is a great deal of clamor in data, and thus, it is hard to achieve a good accuracy. The principal algorithms used in this paper are Naive Bayes and random forest classifier (as per results they are giving higher accuracy).

## 2 Literature Survey

There are various studies in the field of sentiment analysis, elaborated below: -

Desai and Mehta [2] studied an extensive exploration on different techniques for sentiment analysis. In addition, they used some identified parameter to differentiate between the different classifiers.

Preprocessing plays a significant part in cleaning the Twitter data. Tweets consist of hashtags, emoticons, username, punctuation, abbreviation, and URLs. Tajinder Singh and Madhu Kumari normalized the tweets and used SVM classifier together with n-gram for sentiment classification. The outcome distinctly predicted an enhancement in classification as well as accuracy [3].

Barnaghi et al. [4–6] conducted a study to upscale their model. They used a statistical technique called Bayesian logistic regression (BLR) for classification. They also used TF-IDF, unigram and bigram feature and a lexicons-based approach to identify the subjectivity of the tweets. Opinions were classified into two semantic orientations that are positive and negative.

A similar kind of features with n-gram and polarity classification was used by Jianqiang and Xiaolin [7]. They proposed an unsupervised learning method to acquire word embedding from a Twitter corpus along with semantic analysis and statistical characterization on tweets. The method used was deep convolution neural network for training and testing the data.

Jain et al. [1] used different techniques to extract emotion from the multilingual text. Dynamic keywords were used to construct an intellectual model. As emotion plays a crucial role in contributing to the semantic orientation of a word. Additionally, they preprocessed the data and used Naive Bayes algorithm and support vector machine to classify the tweets. Their result indicated that SVM outperformed Naïve Bayes algorithm in most of the cases.

Nagarajan and Gandhi [8] used a hybridization technique along with genetic algorithm, particle swarm optimization, and decision tree to improve the accuracy of sentiment analysis.

Goel et al. [9] suggested a new technique called Naïve Bayes classifier with SentiWordNet to improve the productivity of the model. SentiWordNet is a database of English words with definition according to part of speech.

Bilal et al. [10] discussed the meager researched in sentiment classification with respect to different languages except for English. They used Weka with three classifiers those are  $k$ -nearest neighbor, Naïve Bayes, and decision tree. For tokening, TF-IDF was used. Their experiment showed that Naïve Bayes algorithm performed

the best among the three having the highest accuracy, recall of decision tree is directly proportional to the size of dataset and precision of KNN is inversely proportional to the size of dataset.

Alsaedi and Zubair Khan [11] conducted a survey on machine learning and lexicon-based approach. The classifier they utilized were Naive Bayes, maximum entropy, and SVM. They concluded that machine learning techniques generated better precision when features like unigram and bigram were added. Overall, ensemble and hybrid-based Twitter sentiment classification algorithms performed better.

Mittal and Patidar [12] conducted a research and analyzed that Naïve Bayes classifier performs faster than any other classifier nonetheless, it does not produce the desired accuracy. Furthermore, they said that machine learning algorithm requires adequate dataset to produce good results.

Kharde and Sonawane [13] took a survey and found out that SVM and Naïve Bayes algorithm provide better outcome. Lexicon-based approach is useful in many applications. According to their study, bigram model predicts better accuracy.

Rout et al. [14] Williams proposed a model with supervised learning with unigram, bigram, and part of speech features. The multinomial Naive Bayes classifier produces the best outcome.

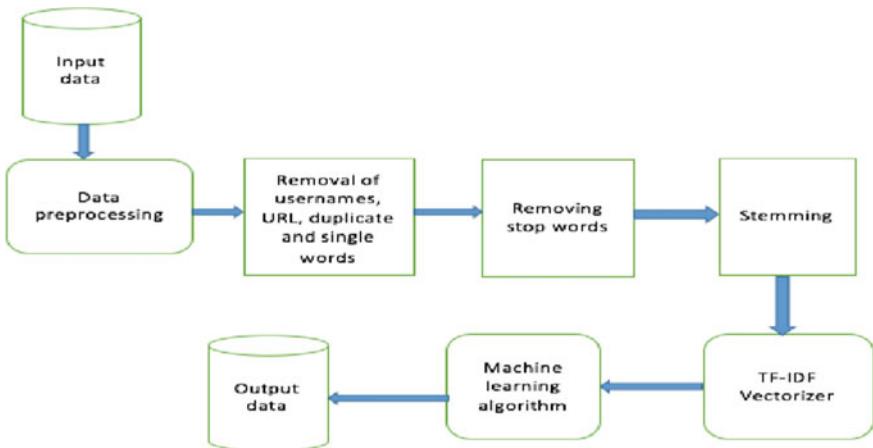
### 3 Methodology

The principle approach related with this experiment is the various information prehandling steps, the machine learning classifiers, and feature extraction process. The major algorithms utilized are Naive Bayes, support vector machines (SVM), linear regression, decision tree, XGBoost, and random forest. The guideline for information prehandling steps incorporates elimination of URL and username, Twitter slang, and stop words. And then finally, stemming the text. The general flow of the model is as follows, (Fig. 1)

#### 3.1 Preprocessing

Tweets have a character restriction is short length messages and have a most extreme length of 280 characters. This confines the measure of data that the client can impart to each message. Because of this explanation, clients utilize a great deal of abbreviations, hashtags, emojis, slang, and special characters and this make the data preprocessing as an important step in Twitter analysis [9].

##### Filtering



**Fig. 1** Architecture

### *Usernames*

Usernames are often used to give a references to other user by using the “@” sign. These again do not add to the sentiment and consequently are supplanted by the regular word.

USERNAME. Also users tend to include emoji which also start with “@.” [4]

### *URLs*

Twitter is a platform where people share their views as well as information. As the given length of tweets is small, one way of sharing the information is using links. A numerous amount of tweets includes links or URLs and these does not add to the sentiment of the tweet [9]. Henceforth, these were parsed and supplanted by a nonexclusive word, URL. Starting with http or https.

### *Duplicates and repeated character*

People on Twitter use plenty of informal language. For example, “cool” is used in the form of “cooool.” While this means the same word “cool,” is treated as two different words by the classifiers. To remove such inconsistency what’s more make words progressively like conventional words, such arrangements of rehashed letters are supplanted. In this manner, “cooool” would be substituted for “cool” [14].

### *Removing single characters*

We use plethora of single words such as vowels and adjective which actually does not contribute to the polarity of the sentence. Hence, they are eliminated.

### *Removing stop words*

There are some of the English words which are not mentioned are hard for the system to translate them. Thus, they are removed [14].

**Table 1** Example of abbreviation

S. No	Abbreviation	Actual word
1	OMG	Oh my god
2	k	Okay

**Table 2** Stemming of words

S. No	Original word	Stemmed word
1	Going	go
2	Goes	go

### Abbreviation Removal

Due to character restriction, people tend to use abbreviations for words. The abbreviations are replaced with the actual words that help to improve performance of the algorithms. Some of the examples of the abbreviation and the actual word can be referred from Table 1.

### Stop Words Removal

There are several English words that are used in sentences as conjunctions. For example, words such as the, and, before, while, etc., that do not add to the estimation of the tweet. Such terms likewise do not assist with recognition of the tweets since they show up in all tweets batches. These terms are eliminated from the data, so they are not utilized as feature. The stop words corpus is imported through NLTK [14].

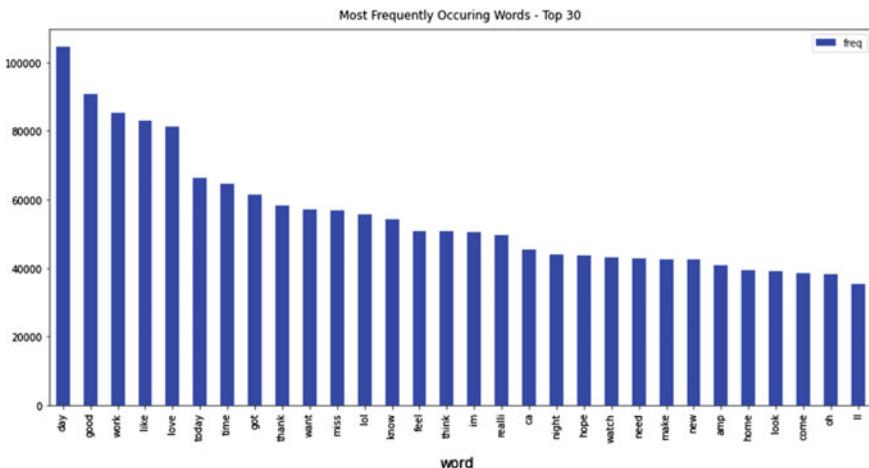
### Stemming

Stemming is a technique in which the word is reduced into its original root word (Table 2) [12]. For example, running to run. NLTK tool kit is used, which provides multiple packages for stemming words such as the PorterStemmer, Lancaster-Stemmer, and so on. The PorterStemmer is the most common technique used for stemming. In our analysis, we have also used the same technique. Most frequent words used in the dataset can be represented in a graph (Fig. 2).

Finally, after applying all the processes involved in preprocessing, we get a clean dataset (Fig. 3).

## 3.2 TF-IDF Vectorizer and Count Vectorizer

TF-IDF stands for “term frequency inverse data frequency.” It gives us the recurrence of the word in each report within the corpus. Term frequency is that the proportion of frequency of the word in a record to the total number of words in that record. Inverse data frequency (IDF) is equivalents to the weight of uncommon words across the reports within the corpus. A high IDF score is obtained if the frequency of the word



**Fig. 2** Most frequently occurring words frequency

	target	text	Clean_tweet
0	0.0	@switchfoot http://twitpic.com/2y1zl - Awww, t...	http://twitpic.com/2y1zl awww that's bummer you should...
1	0.0	is upset that he can't update his Facebook by ...	upset ca not updat facebook text might cri res...
2	0.0	@Kenichan I dived many times for the ball. Man...	dive mani time ball manag save the rest go bound
3	0.0	my whole body feels itchy and like its on fire	whole bodi feel itchi like fire
4	0.0	@nationwideclass no, it's not behaving at all....	not behav i'm mad ca not see
5	0.0	@Kwesidei not the whole crew	not whole crew
6	0.0	Need a hug	need hug
7	0.0	@LOLTrish hey long time no see! Yes.. Rains a...	hey long time see ye rain bit bit lol i'm fine...
8	0.0	@Tatiana_K nope they didn't have it	nope
9	0.0	@twittera que me muera ?	que muera

**Fig. 3** Snapshot of clean dataset

is rare. Initially, we import TF-IDF vectorizer and fit it into the model. This will convert it into tokens [5] (Fig. 4).

Count vectorizer is an easy procedure to tokenize the data, construct a new vocabulary, and encode the data according to the new vocabulary. The count vectorizer returns the value in the form of integer, and TF-IDF returns in the form of float, this is the single difference between them.

```
from sklearn.feature_extraction.text import TfidfVectorizer
```

**Fig. 4** Implementation of TF-IDF approach

### *N-gram feature*

It is a process of taking consequent words at the same time in a sentence for analysis. For reference, the value of N can be 1, 2, or 3 which is defined as unigram, bigram, and trigram, respectively. If  $N = 2$ , we analyze two sequential word in a sentence. This also affects the classification accuracy [4].

## **3.3 Machine Learning Algorithms**

### *Naïve Bayes*

The Naïve Bayes classifier is one of the most common algorithms which is used for sentiment analysis. The Bayes theorem is the bases of this algorithm. Naïve Bayes is fastest and straightforward [9] algorithm. We have used Bernoulli Naïve Bayes from scikit-learn library. In this, the weight of each word is equal to 1, if present or 0 if not [15].

### *Support Vector Machines*

Support vector machines are another common technique used for classification. To provide complete separation of data point, a hyperplane is constructed in a high-dimensional space. This is the clarification, the SVM is similarly called the maximum margin algorithm. The hyperplane recognizes certain models are closed to the plane which is referred as support vectors.

### *Random Forest*

In random forest technique, the classifiers are created from smaller subsets of the input file and subsequently their individual outcomes are collected dependent on a voting process to deliver the ideal yield of the input file set. We used random forest algorithm by using sklearn. Ensemble given by scikit-learn library.

### *XGBoost*

XGBoost is an algorithm that uses an ensemble of weaker trees. It uses gradient boosting framework to improve its performance. We have also tried to solve the problem with XGBoost classifier. We set the classifier in the default mode.

### *Decision Tree*

Decision trees are a type of supervised learning methods, so this algorithm requires training of data before implementation. It is very similar to text classification: Given a set of documents (for example, represented as TF-IDF vectors) together with their labels, the algorithm can determine how much each word correlates with a particular label. For example, if the word “good” frequently appears in data labeled it as positive, whereas the word “bad” will be labeled as negative. By integrating all these observations, it builds a model that can assign a label to any data.

### *Logistic Regression*

Logistic regression calculates the best weight such that the function is as similar as to all actual responses. Using the available observations, the method of determining the best weights is called model training or fitting. If there are one or more independent variables, it is used to evaluate the output [15]. The output value is in binary form.

## **3.4 Libraries**

### *Natural Language Toolkit*

The NLTK helps to build a Python program. It includes different library for text classification, labeling, stemming, tokenization, and parsing. NLTK was broadly used in this paper for tokenizing, stemming, and classification. Additionally, lemmatization was performed by using NLTK library. Lemmatization is a technique in which the words are grouped together so that they can be analyzed. Lemmatization and stemming are almost similar concept, except that lemmatization brings the context to the words [15]. Furthermore, to import Naïve Bayes classifier, NLTL is used.

### *Pandas*

Pandas is a software library used for data analysis and manipulation. It provides multiple data operations methods.

### *Scikit-learn*

Scikit-learn is an open source AI library. It bolsters different strategies like regression, clustering, and classification algorithms including support vector machines, logistic regression, Naive Bayes, random forests, gradient boosting, and k-means and is intended to work with the Python libraries, for example, SciPy and NumPy.

### *Matplotlib*

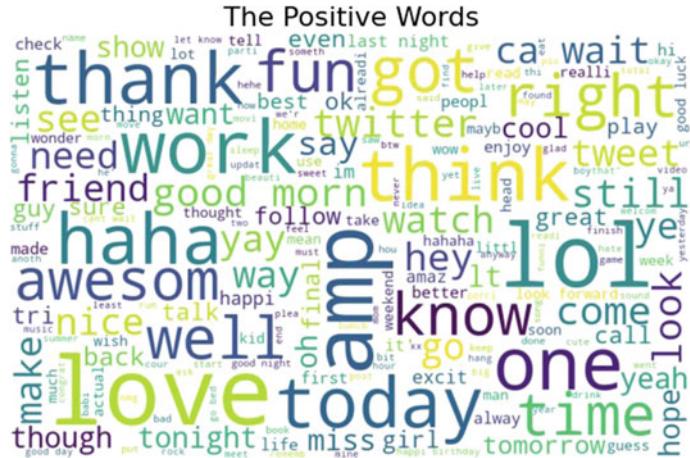
Matplotlib is an extensive Python library designed to construct animated, static, and interactive visualizations in Python. It is the mathematical extension to NumPy to map graphical GUI into the applications.

### *NumPy*

NumPy full form is numerical Python. It is used for building multi-dimensional arrays and large mathematical operations.

### *Word Cloud*

Word cloud is a method used for data representation. The size of the word represents its reoccurrence. A word cloud helps to highlight important textual data points. Modules such as— matplotlib, pandas, and wordcloud are required for using word cloud in Python. The figure shown below is the word cloud visualization of positive words (Fig. 5).



**Fig. 5** Visualization of positive words through word cloud

## 4 Result

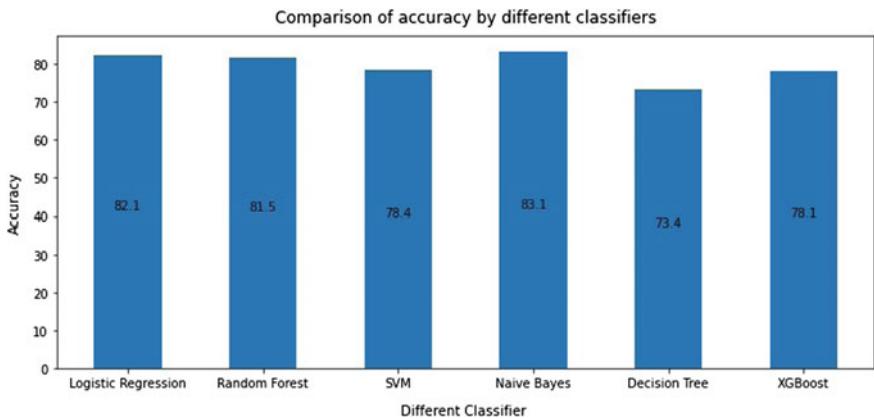
We performed the experiments using various classifiers. We used 20% of the training dataset for validation of our models to check against overfitting. The dataset used here is from Kaggle with 1,600,000 tweets with 800,000 positive and negative tweets, respectively. The fields from the dataset ids, flag, date, and user are removed. The target consists of sentiments of tweets which are classified as positive as four and negative as zero (Fig. 6).

All the tweets are extracted, and NLTK is used to describe each tweet in terms of the features it contains. Additionally, we create a set of words according to the frequency. We trained the classifier with this dataset. The classifier is used in the rest of 20% of the tweets to predict the accuracy.

target	ids	date	flag	user	text
0	0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_ @switchfoot http://twitpic.com/2y1z1 - Awww, t...
1	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton is upset that he can't update his Facebook by ...
2	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattyous @Kenchan I dived many times for the ball. Man...
3	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF my whole body feels itchy and like its on fire
4	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli @nationwideclass no, it's not behaving at all....
...	...	...	...	...	...
1599995	4	2193601966	Tue Jun 16 08:40:49 PDT 2009	NO_QUERY	AmandaMarie1028 Just woke up. Having no school is the best fee...
1599996	4	2193601969	Tue Jun 16 08:40:49 PDT 2009	NO_QUERY	TheWDBboards TheWDB.com - Very cool to hear old Walt interv...
1599997	4	2193601991	Tue Jun 16 08:40:49 PDT 2009	NO_QUERY	bpbabe Are you ready for your Mojo Makeover? Ask me f...
1599998	4	2193602064	Tue Jun 16 08:40:49 PDT 2009	NO_QUERY	tinydiamondz Happy 38th Birthday to my boo of all time!!!!...
1599999	4	2193602129	Tue Jun 16 08:40:50 PDT 2009	NO_QUERY	RyanTrevMorris happy #charitytuesday @theNSPCC @SparksCharity...

1600000 rows x 6 columns

**Fig. 6** Snapshot of the dataset



**Fig. 7** Comparison of accuracy by different classifiers

The preparation and test data we used contains instances dividing it into positive and negative. We have used n-gram feature with the value of  $N = 2$ , that is, bigram to improve the performance of the model. Additionally, we calculated the accuracy score from scikit-learn library along with F-score.

After evaluating the models, we get an accuracy score of 83.1% for Naïve Bayes algorithm and the second best of 82.1% from logistic regression Algorithm while using random forest, the model predicted the score of 81.5%.

The figure shown below is the comparison of accuracy by different algorithms (Fig. 7).

## 5 Conclusion

Naive Bayes classifier gives relatively good outcomes disregarding of the strong suppositions made about the independence among the features. Basically, it outflanks all the others classifiers.

Our model first collects the dataset of the Twitter and then refines the datasets and uses stemming technique to convert the words into its root words. Then, fields are extracted from these datasets which are supposed to be analyzed (time, favourite\_count, likes, frequency, retweets, etc.). And we used TF-IDF scheme to enhance the accuracy of the sentimental analysis. Finally, sentiment of the tweets is analyzed from these texts and are classified as positive and negative sentiments.

By expanding the size of the dataset in steps, uncovered that accuracy increases with the expansion in the size of the training dataset. Using bigram feature in sentiment analysis, it gives better accuracy. We can conclude that more refined dataset can result in accurate predication of sentiment.

While performing data analysis in real time, it aids the business associations to monitor their administrations and produces occasions to advance, publicize, and improve every once in a while.

In future, a large dataset can be used to predict the sentiment analysis which will furthermore increase the accuracy. Numerous tweets lack in any particular sentiment which can be improved. A better approach should be used to interpret emoticon's emotion rather than removing it. Additionally, deep learning approach can be used to enhance the accuracy of the model.

## References

1. V.K. Jain, S. Kumar, S.L. Fernandes, Extraction of emotions from multilingual text using intelligent text processing and computational linguistics. *J. Comput. Sci.* **21**, 316–326 (2017)
2. M. Desai, M. Mehta, Techniques for sentiment analysis of Twitter data: a comprehensive survey, in *2016 International Conference on Computing, Communication and Automation (ICCCA)* (IEEE, Noida, India, 2016), pp. 149–154
3. T. Singh, M Kumari, Role of text pre-processing in twitter sentiment analysis. in *12th International Conference on Communication Networks, ICCN 2016, 12th International Conference on Data Mining and Warehousing, ICDMW 2016 and 12th International Conference on Image and Signal Processing, ICISP 2016* (Procedia Computer Science, vol. 89, Bangalore; India, 2016), pp. 549–554
4. G.R. Nitta, B.Y. Rao, T. Sravani, N. Ramakrishiah, M. Balaanand, LASSO-based feature selection and naïve Bayes classifier for crime prediction and its type. *SOCA* **13**(3), 187–197 (2019). <https://doi.org/10.1007/s11761-018-0251-3>
5. D. Vu, T. Nguyen, T.V. Nguyen, T.N. Nguyen, F. Massacci, P.H. Phung, A convolutional transformation network for malware classification, in *2019 6th NAFOSTED Conference on Information and Computer Science (NICS)* (Hanoi, Vietnam, 2019), pp. 234–239. <https://doi.org/10.1109/NICS48868.2019.9023876>
6. P. Barnaghi, P. Ghaffari, J.G. Breslin, Opinion mining and sentiment polarity on twitter and correlation between events and sentiment, in *Proceedings - 2016 IEEE 2nd International Conference on Big Data Computing Service and Applications*, BigDataService 2016, art. no. 7474355 (IEEE, Oxford, 2016), pp. 52–57
7. Z. Jianqiang, G. Xiaolin, Z. Xuejun, Deep convolution neural networks for twitter sentiment analysis. *IEEE Access* **6**, 23253–23260 (2018)
8. S.M. Nagarajan, U.D. Gandhi, Classifying streaming of Twitter data based on sentiment analysis using hybridization. *Neural Comput. Appl.* **31**(5), 1425–1433 (2019)
9. A. Goel, J. Gautam, S. Kumar, Real time sentiment analysis of tweets using Naïve Bayes, in *Proceedings on 2016 2nd International Conference on Next Generation Computing Technologies, NGCT 2016*, art. no. 7877424 (IEEE, Dehradun, India, 2017), pp. 257–261
10. M. Bilal, H. Israr, M. Shahid, A. Khan, Sentiment classification of Roman-Urdu opinions using Naïve Bayesian, Decision Tree and KNN classification techniques. *J. King Saud Univ. Comput. Inf. Sci.* **28**(3), 330–344 (2016)
11. A. Alsaedi, M. Khan, A study on sentiment analysis techniques of twitter data. *Int. J. Adv. Comput. Sci. Appl.* **10**, 361–374 (2019)
12. A. Mittal, S. Patidar, Sentiment analysis on twitter data: a survey, in *Proceedings of the 2019 7th International Conference on Computer and Communications Management (ICCCM 2019)* (Association for Computing Machinery, New York, NY, USA, 2019), pp. 91–95
13. V.A. Kharde, S.S. Sonawane, Sentiment analysis of twitter data: a survey of techniques. *Int. J. Comput. Appl.* **139**(11), 0975–8887 (2016)

14. J.K. Rout, K.-K.R. Choo, A.K. Dash, S. Bakshi, S.K. Jena, K.L. Williams, A model for sentiment and emotion analysis of unstructured social media text. *Electron. Commer. Res.* **18**(1), 181–199 (2018)
15. S. Symeonidis, D. Effrosynidis, A. Arampatzis, A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Syst. Appl.* **110**, 298–310 (2018)

# Predictive Analysis on HRM Data: Determining Employee Promotion Factors Using Random Forest and XGBoost



D. Vishal Balaji and J. Arunnehrudayam

**Abstract** The size of companies has seen an exponential growth over the years. Corporations recruit anywhere from a few hundred to a few thousand employees every year. With such rates, human resource management in companies is proving to be more and more significant every day. Done manually, HRM is a laborious task, given the sheer quantity of employees. Luckily, over the years, data analytics in HR is emerging as an integral part in corporate operation. Yet, there remain a few tasks that involve human involvement, one of them being selecting candidates that are eligible for a promotion. This paper proposes a solution using decision tree-based machine learning algorithms to learn from past employee records to aid this decision-making process. It explores the usage of two machine learning algorithms, random forest, and XGBoost to predict whether an employee is eligible to receive a promotion or not and determine what factors are responsible for that prediction.

**Keywords** Human resource management · Machine learning · Random forest · XGBoost

## 1 Introduction

HR analytics [1] in corporations plays a major role in restructuring the operand of their HR department. A company of sizeable proportions deals with hundreds of employee records every day. Although HR analytics has been in operation in companies for years, some of these operations are still done manually. Automating such processes will aid in saving valuable time and increasing overall efficiency in the operation of the company. Eligibility for promotions depends on numerous criteria. These criteria vary between companies and even between departments within a company. Currently, there is no way to automate the process of determining why one employee should be promoted over the other, since such decisions require logical

---

D. Vishal Balaji · J. Arunnehrudayam (✉)

Department of Computer Science and Engineering, SRM Institute of Science and Technology, Vadapalani Campus, Chennai 26, India

reasoning and understanding the current environmental factors that computers are just not capable of.

Machine learning has been used to solve similar problems in different domains for several years now. In areas where some kind of human intervention is necessary, a well-trained machine learning algorithm has been proven to be an acceptable substitute, if not an ideal solution. Here, we will be using machine learning techniques to not only predict the promotion status of future employees but to also determine which of the provided attributes from the employees' data are most relevant to making this prediction. In this paper, we explore the application, the two different machine learning algorithms, namely random forest [2] and XGBoost [3] algorithms, to analyze a publicly available HR dataset and determine what factors help elevate an employee's chances of getting promoted.

In this paper, we have performed exploratory analysis on HRM data and used machine learning to find the factors that more commonly lead to an employee being considered for a promotion during an appraisal. The data required for this purpose were collected by an anonymous organization and made available to the public. This dataset consists of various differentiating features for the previous candidates who were shortlisted for a promotion and whether or not they were promoted. This dataset consists of 14 different attributes or columns, with 54,808 total observations or rows. A comprehensive summary of the dataset is as follows:

- **employee\_id(int)**: Unique id of the employee.
- **department(string)**: The department to which the employee belongs to. Possible values are *analytics, finance, HR, legal, operations, procurement, R&D, sales and marketing*, and *technology*.
- **region(string)**: Region of employment. Possible values are *region\_10, region\_11, region\_12, region\_13, region\_14, region\_15, region\_16, region\_17, region\_18, region\_19, region\_2, region\_20, region\_21, region\_22, region\_23, region\_24, region\_25, region\_26, region\_27, region\_28, region\_29, region\_3, region\_30, region\_31, region\_32, region\_33, region\_34, region\_4, region\_5, region\_6, region\_7, region\_8, region\_9*.
- **education(string)**: Describes the level of education of the employee. Possible values are *bachelor\*s, below secondary, master\*s, and above*.
- **gender(string)**: Gender of the employee. Possible values are *f* and *m*.
- **recruitment\_channel(string)**: Channel of recruitment of the employee. Possible values are *referred, sourcing, and other*.
- **number\_of\_trainings(int)**: Describes the number of training programs completed by the employee. Range: from *1* to *10*.
- **age(int)**: Describes the age of the employee. previous\_year\_rating (int): Employee rating from the previous year. Range from *1* to *5*.
- **length\_of\_service(int)**: The service length of the employee in years.
- **KPIs\_met > 80%(int)**: Describes whether the employee's *key performance indicators* scores are greater than 80%. Value is *1* if yes, else *0*.
- **awards\_won?(int)**: Whether the employee won any awards last year. Value is *1* if yes, else *0*.

- **avg\_training\_score(int)**: Employee's average training score in training evaluations.
- **is\_promoted(int)**: Whether the employee was promoted or not. Value is 1 if yes, else 0.

The remainder of this research paper is organized as follows: A literature review documenting past approaches and research similar to that in this paper are represented in Sect. 2. Section 3 provides a detailed description of the working dataset, outlines the preparation of the dataset, and elaborates on the two methodologies used in this paper, i.e., random forest, and XGBoost classifiers. In Sect. 4, we compare the outcomes of the two methods and elaborately discuss the results. We conclude this paper in Sect. 5 with references and future work.

## 2 Literature Review

Marler and Boudreau in [4] discuss the adoption of HR analytics by organizations and attempt to answer some key questions regarding its definition, inner workings, effectiveness, its impact on corporate operation, and its success factors by conduct evidence-based reviews of articles in peer-reviewed journals. Simbeck in [5] builds upon the previous literature on the discussion about the ethical implications of HR analytics.

Quinn, Rycraft, and Schoech as well as Liu et al. use logistic regression to predict employee turnover in [6–9]. The authors in [6] use this in conjunction with a neural network model from StatSoft. They use a dataset comprised of 15 variables and produce a model that can predict turnover with anywhere between 60 and 70% accuracy. The authors in [7] contrast logistic regression with other supervised learning approaches like random forest and AdaBoost algorithms in determining factors that influence an employee's promotion. From their analysis, the random forest classifier outperforms the other models with accuracy and AUC of 85.6% and 88.9%, respectively, along with a precision of 83.4%. Jin et al. in [10] predict employee turnover for companies using classification algorithms and survival analysis. They employ a variation of the random forest algorithm named RFRSF, which combines survival analysis for censored data processing and ensemble learning for turnover behavior prediction. The authors contrast the results their model with those of traditional machine learning techniques such as Naïve Bayes, decision tree, and random forest and their algorithm predicts employee turnover with 84.65% accuracy.

### 3 Methodology

#### 3.1 Preprocessing

The variable to be predicted by the model, i.e., the target variable is the `is_promoted` column, which indicates whether the employee with particular attributes has been promoted or not. Our algorithms will be trained to output a prediction of what the value of this variable will be based on the values of the other variables of a particular observation.

As mentioned in the dataset description, the target variable can have 2 values: 0, which indicates that the particular employee has not been promoted or 1, which indicates that the particular employee has been promoted.

The target variable is currently in integer encoding, which means that the possible values of this variable are in the form of a single integer, each representing one of the output classes. But, allowing the model to assume a natural order between these two categories may result in poor performance or unexpected results such as predictions halfway between the categories. So, we encode the target variable using a method called one-hot encoding. One-hot encoding is a method of representing categorical variables in a more expressive manner. It helps to indicate to the algorithm that the expected output is to be categorical and not continuous.

To perform one-hot encoding, we use the integer representation of our target variable and transform them into an array of binary digits whose length is equal to the total number of possible values (two in this case). The digit in the array whose position corresponds to our integer value is set to 1 while the other values are set to 0, i.e., 0 becomes [1, 0], 1 becomes [0, 1], etc.

To aid in better visualization of the data, however, the values of the target variable 0 and 1 are represented as *no* and *yes*, respectively, in all figures, where *no* means that the particular employee has not been promoted and *yes* means that the employee has been promoted.

#### 3.2 Feature Engineering

First off, we remove the `employee_id` column as it is just a column to distinguish records by and will not realistically impact the decision of whether an employee is to be promoted or not. This brings down our variable count to 13.

*Removing Duplicate Values:* Even though we have a large volume of data, we cannot be sure that all the observations in the dataset are unique. Looking at our data, we find that there are 118 duplicate records that are present in our dataset. Even though it does not seem like a significantly large number, duplicate records will negatively impact the training process of machine learning algorithms and may cause them to over-fit. After the removal of the duplicate values, we have 54,690 observations left.

*Removing Missing Values* To deal with missing values, we separate out the observations with a NULL value for at least one variable. We also notice that only two columns, such as education and previous\_year\_rating, contain null values in 2398 and 4062 observations, respectively.

We rectify this by filling in the missing attribute with the mode of that particular attribute in all the observations, where the value of the target variable is the same as in the column with the missing value. The steps to do so are as follows:

1. For each observation with a missing value, the mode of the non-missing values of the variable with the missing value is calculated for each value of the target variable. For example, in our dataset, the mode is found for all the available values of the **education** variable for all the observations where the value of the target variable is 0. The same is done for observations where the target variable is 1.
2. The missing values are then filled with the calculated mode corresponding to the value of the target variable.
3. This process is repeated for each of the columns where values are missing.

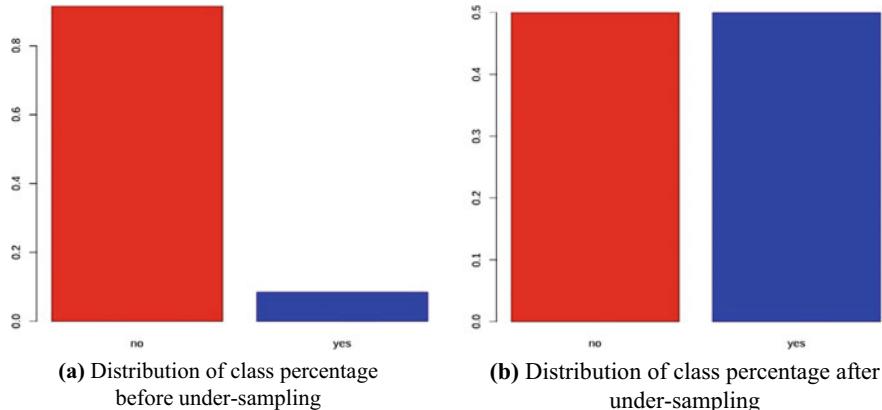
This approach allows us to eliminate observations with missing values while still retaining the size of our dataset.

### 3.3 *Balancing Classes*

In this dataset, there is a clear imbalance between the values of our target variables. Out of 54,690 observations, only 4665 observations are of class “no,” while 50,025 observations are of class “yes.” This is a significant issue, as the difference exceeds more than 50% of our total data. For a machine learning algorithm to be properly able to parse and understand the given data, there should ideally be an equal distribution of the number of examples with the different classes that the model is meant to classify data into. When one class takes precedence over the other class in the dataset, the algorithm is less likely to learn what the properties of each class are and tend to forget the less frequent class’ properties all together during training.

To ensure that this does not happen, we must make sure that there are equal numbers of examples for both the cases. Here, we randomly sample a subset of the data where the class is “yes,” as it is the class with higher frequency and append it to the observations with the class “no” to generate a new, minified training set with equal number of “yes” and “no” observations. This process is called undersampling. This brings down the size of our dataset to 9330 observations total. Even though it is only a fraction of the original 54,690, it is still a significant amount and should be enough to train our algorithms along with being balanced.

The distribution of class percentage before and after undersampling is shown in Fig. 1.



**Fig. 1** Distribution of *is\_promoted* class before and after undersampling

### 3.4 Training and Evaluation

The dataset is split into a training set and a testing set, where the testing set contains 1/5ths of the records in the dataset while the rest belong to the training set. The resultant training and testing sets will therefore have 7464 and 1866 observations, respectively.

A random forest (RF) and XGBoost (XGB) model will be trained on the training set and be evaluated using the K-fold cross-validation method [11, 12] for the testing set.

To minimize computational power, considering that the dataset is fairly large in size, as well as to reduce the bias of the algorithm, we need to select a small value for K. Here, we arbitrarily choose K to be five, since our dataset has 9330 observations, which can evenly be divided into five parts.

Confusion matrices will be generated for both models, which will help us determine the confidence with which the models classify observations as *yes* or *no*.

*Random Forest:* The random forest is a supervised learning approach that utilizes multiple decision trees, each representing a feature or label in the dataset in a random order, to arrive at its final conclusion. The final prediction of the RF model is dependent on the decision trees present in the model. The prediction from each individual tree stored and is polled in the end. The prediction at the end with the highest frequency is selected.

The main drawback with decision trees is their low bias and high variance.

The random forest algorithm uses this drawback to its advantage and utilizes multiple trees with slight variations. This helps prevent the model from overfitting and allows it to handle much more complex data than decision trees.

The standard recursive partitioning algorithm of a decision tree starts with all the data and does an exhaustive search over all variables and possible split points to find

one that best explains the entire data, thereby reducing the node impurity the most. However, for large trees the likes of which are used by the random forest algorithm, such an approach can get very computationally expensive. To avoid this problem, the random forest algorithm uses a variable called *mtry* for each split. The algorithm then randomly selects *mtry* predictor variables from the dataset, which ensures that not all the variables are included in the split, while also selecting a different set of variables for each split.

Three models were trained with three randomly assigned values for *mtry*. Out of those, the best performing model was that with *mtry* = 28 with an accuracy of approximately 82%.

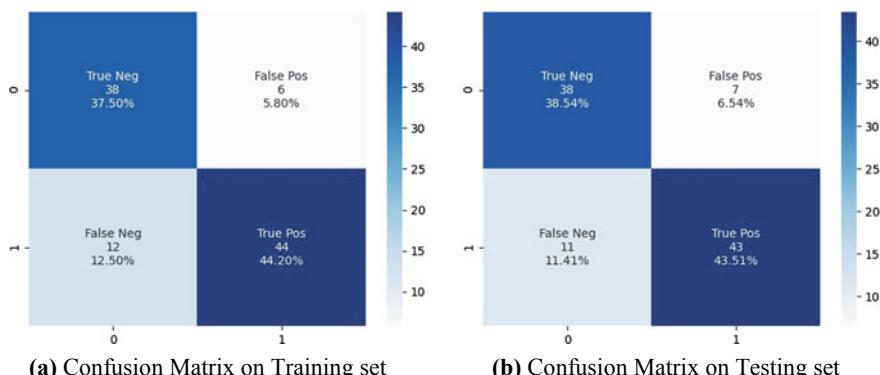
The training yielded tuning parameters for all three models which are shown in Table 1. The *accuracy* metric was used to select the optimal model using the largest value.

The confusion matrix plots for the predictions from the RF model can be seen in Fig. 2.

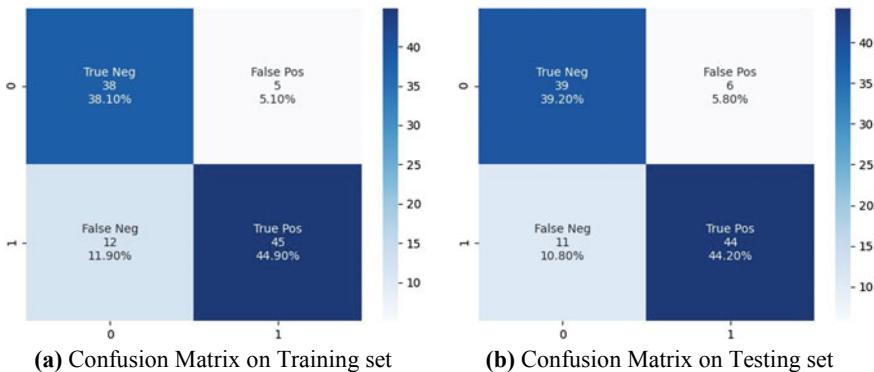
**XGBoost:** The XGBoost (short for extreme gradient boosting) algorithm is a variation of the random forest concept which utilizes the concept of gradient boosting to enhance the performance of the model. This algorithm performs best when being used to model small to medium sized structured data. The XGBoost algorithm is generally preferred over other conventional algorithms as it combines many different traits from other algorithms such as bagging, gradient boosting, random forest, and

**Table 1** Tuning parameters for RF model

Mtry	Accuracy	Kappa
2	0.7600484	0.5200968
28	0.8161811	0.6323622
54	0.8105528	0.6211057



**Fig. 2** Confusion matrices of RF model evaluated with 5-fold cross-validation



**Fig. 3** Confusion matrices of XBG model evaluated with 5-fold cross-validation

decision trees and makes improves upon them through system optimization and algorithmic enhancements such as regularization, sparsity awareness, weighted quartile sketch, and so on.

**Gradient Boosting:** Gradient boosting is a repetitive algorithm used to leverage the patterns of mistakes made by a model to strengthen the model using weak predictions. Basically, the data are modeled very simply and are analyzed for errors to identify data points that are difficult for the model to fit. The model is tweaked to fit better for those particular data points. Finally, this is combined with the original models for an optimal solution.

The confusion matrices plotted from the predictions of the XGB model on both the training and testing sets are shown in Fig. 3

## 4 Results

The statistics of the final trained models when evaluated on the testing set are shown in Table 2.

**Table 2** Final statistics of the trained RF and XGB models evaluated on the testing set

Attribute	RF	XGB
Accuracy	0.8199	0.8339
P-Value [Acc > NIR]	<2.2e-16	<2.2e-16
Kappa	0.6399	0.6677
Sensitivity	0.7706	0.7835
Specificity	0.8692	0.8842
Balanced accuracy	0.8199	0.8339

The above statistics are used to evaluate and describe the models. *Accuracy* is the percentage of correctly classified instances out of all instances. The *P-value* is a measure of the probability that an observed difference could have just by random chance. *kappa* or *Cohen's kappa* is similar to *accuracy*, but it is normalized at the baseline of random chance on the dataset. *Sensitivity* describes the model's ability to predict true positives, while *specificity* is a metric that evaluates the model's ability to predict true negatives. *Balanced accuracy* is calculated as the average of the proportion of correct predictions from each individual class.

The *accuracy* and *kappa* values of both models are very similar, indicating that both models perform very similarly for this particular dataset. Both models perform consistently over the testing set with regards to their performance on the training set and are able to classify the data points. However, the XGB model has an accuracy of 83.39% which is slightly higher than that of the RF model, which is approximately 82%. The XBG model also clearly outperforms the RF model in almost every other aspect, which was as expected.

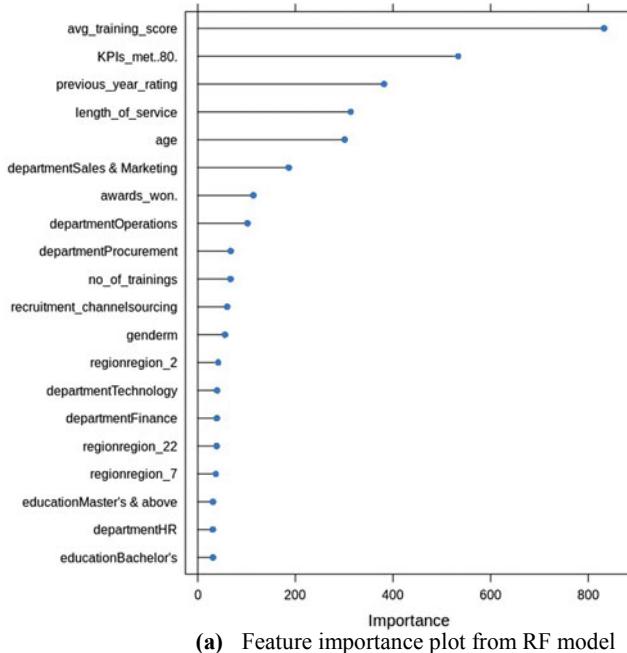
Both models assign a “feature importance” score to each attribute, which is calculated by the amount that the particular attribute’s split point improves the performance measure, weighed by then number of observations, the node is responsible for. In our case, the performance measure is the *Gini index*, which is used to select the splits points. The plots of the 20 most important attributes according to the RF model and the XGB model can be seen in Fig. 4a, b, respectively.

We notice that the list of the ten most important features according to both the random forest and XGBoost models are nearly identical with the exception of no\_of\_trainings and departmentHR. This tells us that the other attributes can be discarded to get improved performance from the models.

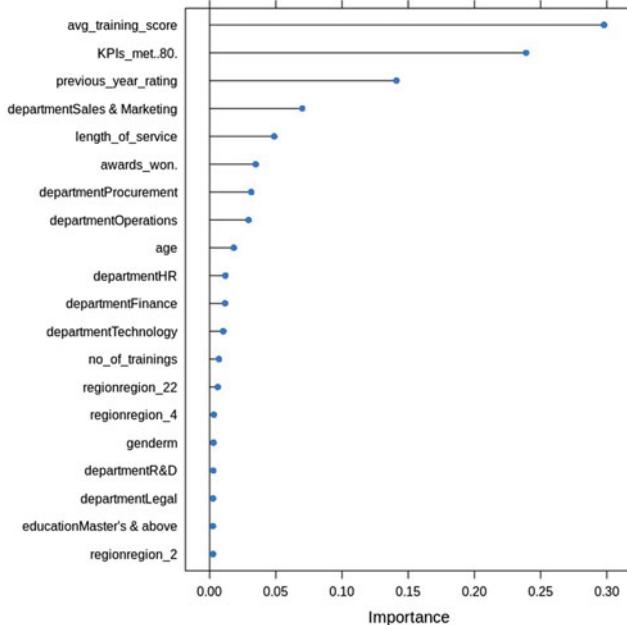
## 5 Conclusion

In this paper, we have demonstrated the use of the random forest and XGBoost machine learning algorithms in predicting an employee’s promotion status. We also used the trained models to determine which attributes have the most impact on said promotion status. Through this, we can see the use of machine learning as a predictive decision-making tool or at least a suggestive tool is a fully viable solution for the presented problem. With fairly limited data and computational resources, we were able to train algorithms to perform with significantly good accuracy.

Going forward, higher accuracy and more efficient models can be achieved with higher volumes of data and more careful tuning of the models’ parameters. We can also improve the models’ accuracy by retraining the models on just n-most important features as seen from Fig. 4. Furthermore, we can expand this research by applying various other machine learning approaches such as support vector machines and recurrent neural networks.



(a) Feature importance plot from RF model



(b) Feature Importance plot from XGB model

**Fig. 4** 20 most important features as scored by the RF and XGB models

## References

1. Q. Abdul, A. Mohammed, HR analytics: a modern tool in HR for predictive decision making. *J. Manag.* **6**, 51–63 (2019)
2. L. Breiman, Random forests. *Mach. Learn.* **45**, 5–32 (2001)
3. T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system. *CoRR*. abs/1603.02754 (2016)
4. J.H. Marler, J.W. Boudreau, An evidence-based review of HR analytics. *Int. J. Human Res. Manage.* **28**, 3–26 (2017)
5. K. Simbeck, HR analytics and ethics. *IBM J. Res. Devel.* **63**, 9:1–9:12
6. A.Q. MSW, J.R. Rycraft, D.S. PhD, Building a model to predict caseworker and supervisor turnover using a neural network and logistic regression. *J. Technol. Hum. Serv.* **19**, 65–85 (2002)
7. G. Manogaran, P.M. Shakeel, S. Baskar, C.H. Hsu, S.N. Kadry, R. Sundarasekar, P.M. Kumar, B.A.Muthu, FDM: Fuzzy-optimized data management technique for improving big data analytics. *IEEE Trans. Fuzzy Syst.* 1–1 (2020). <https://doi.org/10.1109/tfuzz.2020.3016346>
8. B.-H. Liu, N.-T. Nguyen, V.-T. Pham, Y.-X. Lin, Novel methods for energy charging and data collection in wireless rechargeable sensor networks. *Int. J. Commun. Syst.* **30**, e3050 (2017). <https://doi.org/10.1002/dac.3050>
9. J. Liu, T. Wang, J. Li, J. Huang, F. Yao, R. He, A data-driven analysis of employee promotion: the role of the position of organization, in *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*. pp. 4056–4062 (2019).
10. Z. Jin, J. Shang, Q. Zhu, C. Ling, W. Xie, B. Qiang, RFRSF: Employee turnover prediction based on random forests and survival analysis, in *Web information systems engineering – WISE 2020*. ed. by Z. Huang, W. Beek, H. Wang, R. Zhou, Y. Zhang (Springer International Publishing, Cham, 2020), pp. 503–515
11. J. Arunnehr, A.K. Nandhana Davi, R.R. Sharan, P.G. Nambiar, Human pose estimation and activity classification using machine learning approach. In: V. Reddy, V. Prasad, J. Wang, K. Reddy (eds) *Soft Computing and Signal Processing. ICSCSP 2019. Advances in intelligent systems and computing*, vol 1118. Springer, Singapore. (2020). [https://doi.org/10.1007/978-981-15-2475-2\\_11](https://doi.org/10.1007/978-981-15-2475-2_11)
12. J. Arunnehr, M.K. Geetha, Motion intensity code for action recognition in video using PCA and SVM. In: R. Prasath, T. Kathirvalavakumar (eds) *Mining intelligence and knowledge exploration. Lecture notes in computer science*, vol 8284. Springer, Cham. (2013). [https://doi.org/10.1007/978-3-319-03844-5\\_8](https://doi.org/10.1007/978-3-319-03844-5_8)

# A Review on Deaf and Dumb Communication System Based on Various Recognitions Aspect



G. Arun Prasath and K. Annapurani

**Abstract** It is very hard to communicate with deaf and dumb people without knowing the sign language. In the earlier years, deaf and dumb peoples will not come out and do jobs like nowadays, so the both normal and deaf –dumb peoples are in the need of better and efficient system which converts sign language into voice output. Though we have lots of translating methods, it is also having some barrier in efficiency, so to reduce the barriers in the existing systems, we need to adopt new technology for improving input and output quality of the system. The main objective of survey is making a man and machine communication based on recognition and input processing types in the gesture recognition system.

**Keywords** Sign language · Gesture recognition · Man–machine interface

## 1 Introduction

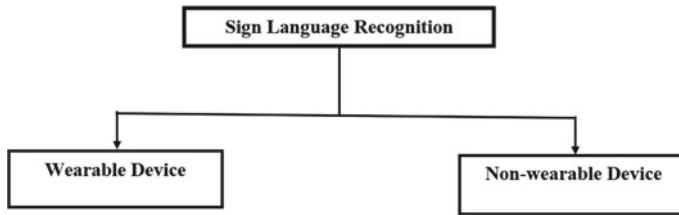
Around the world we have 466 million deaf and dumb people and 34 million of these are children. WHO says it will increase to 900 million by 2050. Hearing loss may result from genetic causes and complication at birth and so on [1]; sign language communication helps the sign community people to interact with normal community people. Within the family, deaf and dumb member will use different types of part to talk so there is no need for professional sign language gesture, but when he communicates with the deaf and dumb person, he or she should use the standard sign language gesture. In this platform, the communication is like a human computer interaction. When it comes to sign language recognition, it is very difficult to get proper input data due to the various reasons like environment and complications in the sign language data. Figure 1 shows the two types of the sign language recognition

---

G. Arun Prasath · K. Annapurani (✉)

Department of Networking and Communications, SRM Institute of Science and Technology,  
Kattankulathur 603203, India  
e-mail: [annapook@srmist.edu.in](mailto:annapook@srmist.edu.in)

G. Arun Prasath  
e-mail: [ag7678@srmist.edu.in](mailto:ag7678@srmist.edu.in)



**Fig. 1** Approaches of sign language recognition

method. Sign language recognition process can be classified into two major types based on the input method wearable.

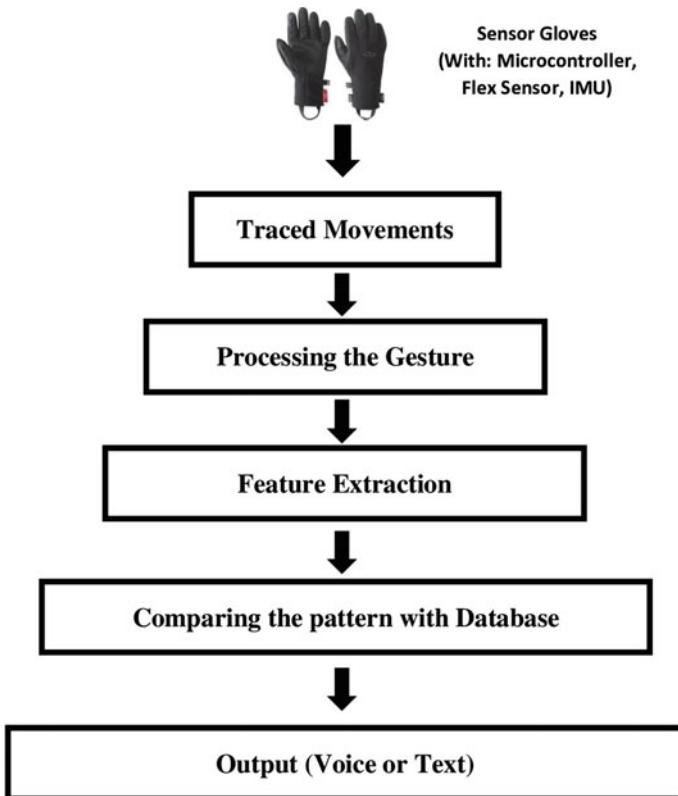
Hardware-based recognition and non-wearable based device, wearable hardware consist of sensors and transmitter to capture the hand movements it senses and track various part movements of the hand like finger and finger joint, and it send the data to processing system. But, it is very hard to wear the sensors and hardware's every day. Figure 2 shows the steps involved in the hardware-based gesture recognition method. In non-wearable methods, camera is predominant part to get the needed input from the user. The main advantage here is it eliminate the usage of sensor gloves. There are lots of non-wearable devices are available to recognize the sign language like webcam Microsoft Kinect sensor and many depth cameras like RGBD. Here, the deaf and dumb people no need to wear hardware that is why its became very popular nowadays [2].

## 2 Sign Language Recognition Methods

Based on the input type and output method, we can classify the sign language translation into three types letter-based recognition, word-based recognition, sentence-based continuous SLR.

### 2.1 Letter Based Recognition

Letter-based sign recognition very useful when we are addressing some name or place for this kind of things deaf and dumb sign language will not have signs [3]. In our Indian sign language dataset, we have totally 26 different hand shapes for every letter. The data can be obtaining from both sensor gloves and vision-based methods. To recognize the data, they have used statistical features and support vector machine. For letter-based recognition, the American sign language recognition obtained 94% accuracy for all 26 letters by the use of support vector machine they classified the letter into different meaning [4, 5].



**Fig. 2** Sensor gloves-based sign language recognition system

## 2.2 Letter-Based Recognition

The main purpose of the word-based recognition method is to give the output in the form of word. One sign for one word by this we can get the output like call, hi, and hello. Most of the sign language recognitions system have been implemented for word-based recognition. That are implemented by using CNN feature extraction, and then it was classified using support vector machine. CNN model with dataset of 20 classes 70% of sign data were used to train, and remaining 30% data were used for verification purpose [2]. The system has 97.28% accuracy, even though some of the sign images are similar with the algorithm that is called state-of-the-art. In existing model has accuracy of 96.58%, and 96.11% using Ycbr with CNN which is called feature fusion-based recognition system [2]. Similarly, [6] obtained input using Kinect sensor, which contain 18 different hand shapes with different angles; each and every shape have different meaning based on the projection of hand and hand joins which are classified by using support vector machine. Accuracy is 83.12% for 20 sign data when they use HOG-PCA. The hand shape descriptor will

have different types of representation for every sign language data [6]. Totally, 23 sign gestures and 3450 video-based sign gesture were used which are performed by three different users [7–9].

### 2.3 Sentence-Based Continuous SLR

This kind of approach will help to the real-time systems; mostly, the continuous sentence recognition is used in the video-based recognition methods. Using leap motion sensor, they have implemented the continuous sign language recognition system for feature extraction by two-dimension convolutional neural network. The long short-term memory (LSTM) network is controlling the input and output gates, so it consists of two gates [10]. In developed modified LSTM has four gates with reset model and the dimension are same. Totally, 157 sign sentences collected from the deaf and dumb people. The testing average accuracy is 72.3%, and the highest entire recognition data are 76%, and the lowest accuracy is 66%. Developed multimodal dynamic-based 3D and 2D long short-term memory network. It is a Chinese sign language data-based recognition system from the video clips. B3D model used to recognize the sign video sequence, and it consists of 17 convolutional layers with 2D long short-term memory network. It automatically extracts spatiotemporal data from the input. DEVISIGN-D and the SLR\_DATA set are used with 6000 video clips [11]. Both isolated and continuous sign gestures were used; they are stored in different class [12].

## 3 Three Major Part of SLR System

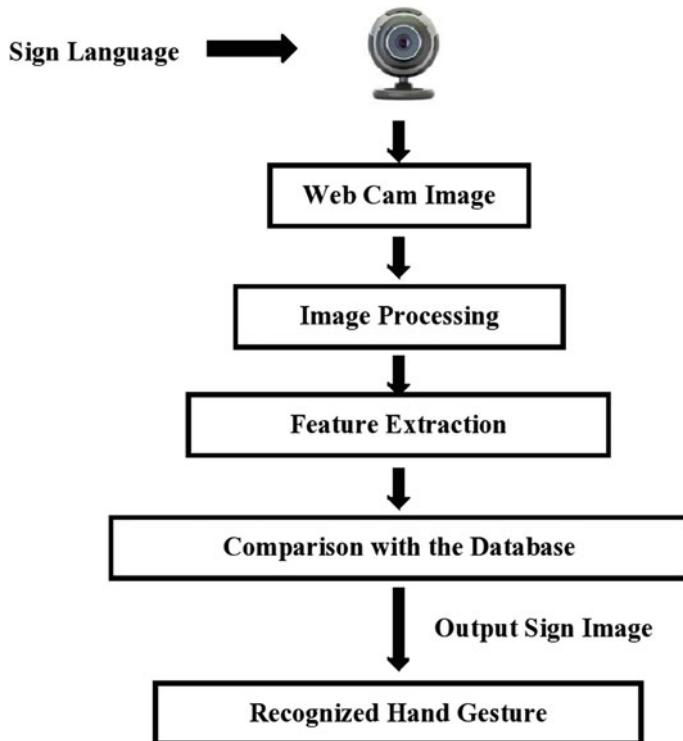
The sign language recognition system input will receive from the user; the outcome will be processed data. Input processing has done in two different ways that are as follows: (i) Wearable and (ii) Non-wearable or vision-based system. The three part of SLR system is input system, processing system, and output system.

### 3.1 Sentence-Based Continuous SLR

In sign language recognition system, the input can be taken in two different ways, which are wearable or sensor-based and non-wearable or vision-based input. The vision-based input takes the input from the camera and motion sensors like Microsoft Kinect and leap motion. The wearable input method will take the input from the sensor gloves which consist of microcontroller and flex sensor, hall sensor. The sensor gloves will measure and track the movement of the input part.

### 3.1.1 Vision-Based Input or Non-Wearable Device-Based Input

Vision-based and sensor-based input will have different kind of processing methods. They implemented the sign recognition system with camera-based input method. First, in the segmentation process of input captured from the camera which is the original hand gesture image of deaf and dumb people and which is converted from original RGB image to Ycbr image format. The pre-processing step is performed to convert the original image to binary form to find exact data from the complete image which will help to obtain the accurate input image. Firstly, the original image has been converted to HSV image. Morphological processing is done to the obtained HSV image to remove the unwanted small parts. Finally, skin-masked final image is obtained [2]. Figure 3 shows the steps involved in the vision-based input system. In hybrid system, camera is attached with both hardware and sensors. The neuromorphic sensor receives hand sign images and which is attached with camera. The neuromorphic sensor is connected through the USB protocol. And to enhance the image quality, a couple of led lights added and works with the driver of FPGA. In equation I, value is  $-1, 0, 1$  for filtration step. Noise filter used to remove the



**Fig. 3** Steps involved in the vision-based input system

unwanted data from the input image. The noise was removed by using pixel noise filter [3].

$$\bar{I}(x, y) = 1 \text{ if } \bar{I}(x + I, y + i) \neq 0$$

Similarly [6], the input has taken from the Kinect sensor, which is also a vision-based input system, but here, the input mode is video. Totally, 92 sign inputs were taken from single hand, and 5 inputs have taken from both hands, and while capturing the input, complete body part will be taken for the frame. The needed part will be segmented from the complete frame. Arabic sign language system uses different kind of body parts like face properties and hand state and depth image. While taking the input, hand shape will have different meaning based on the position and angle of the hand also considered for the input. So, it will give the output based on the combination of face properties, hand sign, and the hand angle. The data will be taken from the Kinect sensor in the form of 3D skeleton. The complete human body consists of 25 joins, but here, they only considered upper part of body which consist 16 joins. The leap motion sensor also has accurate input. The advantage of leap motion-based input system is consisting of 2 IR camera and 3LED's for the better input. This leap motion sensor can take 120–200 frame per second for the input, which using two hands, from the hand it takes data's like hand joins and finger and wrist, and the system can recognize both isolated and continuous sign gestures, total used sign are 35 and 15 times every sign are performed repeatedly. Since it is a continuous SLR method, it can also have more than two word for a single hand gesture and differentiated the input part like the finger joins and palm with color points. Normalizations has been used to get the needed part from the complete frame size [13]. It is a 3-dimension-based sign language recognition system [14]; the model learned by the 3D input dynamic pattern. 3D input system is proposed because of lack of accuracy in 1D system. The input area has both motion and non-motion parts of the body. It is between the highest correlation degree of two parts. In sign language recognition system, input data part is playing a very important role in the proper outcome of the system. Only with the hand part, we cannot get the proper output of the sign's, so here, both face and hand gestures are taken for input part. Eight infrared cameras with reflective masks used to capture the 3D input to get the full body part in the skeleton view. Totally, 500 sign language dataset which consist of our daily need communication like place, health, flowers, and agriculture. These 500 signs have one hand or both hands with face and chest region, then it coordinates in this format  $N_j * N_t * X_3$  matrix. The input is in the form of video frames; the conversion layer converts the input video into frames then the feature maps will map the layers. The proposed data from the ROI pooling classified based on the classifier [11]. The segmented data features are extracted from video, and it is identified by using predicted frame label. In the implemented Chinese sign language dataset, totally, 500 sign data were trained with 8 signers; one type is DEVISIN-D dataset, and the another one is general SLR dataset. Detection and segmentation are performed to identify the needed input area. The sign input taken from the hand area

is identified by the finger joint and skin color needed RGB image will produced from the dept. and skin mask. KCC'S (key frame centered clip) [15] is an input method, and it consists a greater number of frames in single word stream and extraction of the hand image, then finding the needed most meaning full frame from the sequence which is also called locating the key frame. From the represented frame, key frame centered clip is generated. The multi-model feature will have different class like CNN, DMM, and TRA. In Arabic sign language, the input device is two leap motion controllers (LMC); the system has dedicated field to take input from user in the form of cubic and isolated-based sign recognition method with 100 different sign data. Microsoft Kinect is used to get the input,  $D' = D_m + T - D$  (if  $D \neq 0$ ) it is very useful to find the hand area and brighter needed part of the object using threshold value they finding binary hand mask from the dept. image of the created model. Everything is calculated from hand orientation and the palm detection PCA net model will give the most representing part of the image from the input dept. image [4]. It segments layer 1 and layer 2 from output decimal image [16]. This input system is differing from other existing input systems. Non-touchable Kinect-based input method using the hand gesture by giving the character input to the system. The Kinect sensor will track the hand movements using the fingertip, and they were reduced the hand area to the frame. The fingertip detection algorithm will give points at the end of the fingertip, and the input has taken using virtual keypad [17].

### 3.2 Processing System

Processing system will process the input data using different type of approaches, classification of the data done after the input process. Sign language recognition method need to process different types of inputs like sensor-based inputs, camera-based inputs, and sensor gloves-based input. The pre-processed data are stored in the form of features. The hyperplane will separate the data based on the falling location; it contains multiple classes of SVM. After comparing the data with other classes, it sends the output like one versus rest. It helps to classifying the data with comparison of multiple class. One versus rest approach will classify the sign by maximum output. Processing system will improve the quality of the input, and it removes the unwanted data from image [2]. The input method is sensor-based system. Gestures are extracted from the input, and it is stored in gesture set table. The training dataset was stored in separate place, and it is trained using component classifier training, and extracted data are stored in the fuzzy k-means cluster; every sign gesture will have some unique code that code compared by the code matcher with gesture, and it produces the output [18]. The classification was done by ANN; they were used static neuron with an equation of  $a = [f(wk + Pk) + b]$ . ANN is implemented in the digital hardware using four digital buses, it established with processor and ANN. The characteristic array was passed to the processor, then data adjusting will be done. Classified data from the ANN layer will be compared by comparator data, and then, it generates the output. Every alphabet has some signature graph to differentiate the sign [3]. Two

descriptors, one is motion-based descriptor and another one shape-based descriptor. Once the pre-processing has done, the shape descriptor will store the details about hand shape. The motion-based descriptor has COV3DJ descriptor. It contains upper body skeleton, face property, and hand state details. The error analysis done for the same gesture, but it has small difference in the state. So, it is classified in the error analysis [6]. Feature extraction is done using convolutional neural network. Even though it has multiple hidden layers, these are consisting of input and output layers. The 2D CNN-based feature extraction process consisting of  $3 \times 5$  kernel and 3 feature maps of  $1 \times 8$  for time t. The input layer was connected with soft-max layer. Modified LSTM network has reset functionality which will help to remove the previous recognized word and the testing state. They trained both isolated sign language words and continuous sign sentences [10].

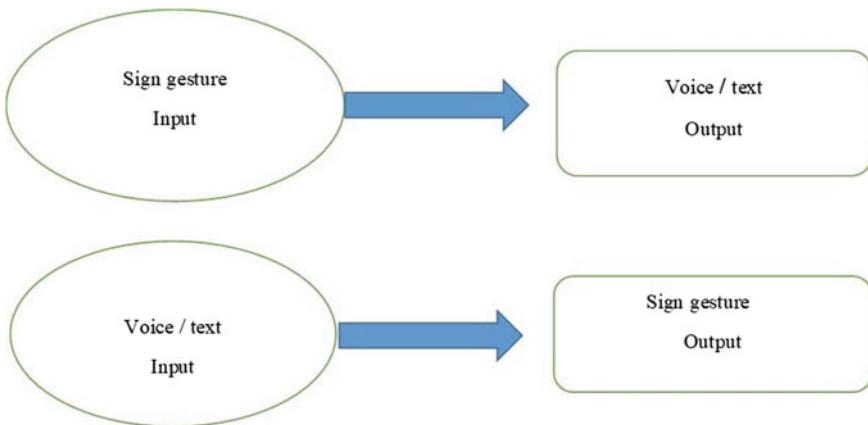
The sign data received from the user based on motion. It classified into two categories that are motion joins and non-motion joins. Once the motion is identified, it tracks the part for classification. Different type of body parts like hand, face, and chest based on the need. They were used hand, chest, and face combinations. For some sign, it requires only one hand and some signs require both hands and face properties. Adaptive kernel matching algorithm compares the source with database, and classified input data are arranged in the database with labels based on the matching score. It classifies the input data [13]. RBI pooling is used to classify the feature extracted data based on the hand gesture ROI will classified into different labels from that data are segmented into different class. Next step is video sequence feature extraction; here, the features are extracted from the video after that, it stored in the feature vector. Dynamic sign language recognition analyses the features vectors. Already have some predicted frame labels to compare the input. Video labels are extracted from the input video sequence based on the predicted frame and video label. Got the input and the objective localization done using faster R-CNN method. The faster R-CNN had generator and box classifier. Box classifier has multiway classification [11]. From the input, it identifies the most representative key frame gesture. Multimodal feature would help to classify the input data into different class; then, the network has been encoded and decoded to LSTM [15]. The training data model has been formulated to analysis the model. Two methods of analysis are Gaussian mixture model and linear discrimination model. Expectation maximization algorithm is used to identify the k which uses the Bayes features. Parameters are estimated for GMM model. Both Bayes and fishers linear are used to classify the data. The input is processed by using support vector machine; two strategies were used to process the input image. The first one is single PCA Net model, and another one is multiple PCA Net models. PCA Net model has 2 convolutional layer and it normalized using zero mean. Principal component analysis algorithm reduces the dimension of gesture. Convolutional layer used in single PAC Net model. If new user found, we need retrain. But in the multiple PCA model, the first convolutional layer learns the law of data feature, and next layer learns the high feature of day [19]. This conversion method using both sensor gloves and Kinect sensor. Both k-nearest neighbor and convolutional neural network are used for the classification process. The process is like binary search tree; the first input is check with graph, and the process goes till the leaf node, and then it displays

the output word [20]. The main objective is to find difference between the gestures. When similar type of gesture found for different meaning, it affects the accuracy of recognition. So, to find these kinds of difference, geodesic method is used. From the input, initial shape will have little different from the final shape in that case have to analysis the different. By geodesic path, similarity between the shapes are identified the orientation of the shape also consider for the comparison of the image. The robustness is calculated in two cases (i) inter individual different between shapes of the same class, (ii) there is only slight shape deformation between the different classes. The mean classification rate is obtained 99.8% [21]. This method consists of three-layers feature extraction, DLSTM for training phase and classification layer. That are as follows: (i) DLSTM have the input layer. (ii) From feature extraction data, n-number of hidden layers are found. (iii) The output layer for the classification. DLSTM higher level of abstraction is achieved by using deep long short-term memory network [22].

NPDA algorithm is used to find the neighboring pixels, and the local binary pattern histogram has been calculated for  $16 \times 16$  patterns. Uniform LBP histogram of  $n^{\text{th}}$  blocks will produce the combined histogram of all the blocks. As like many other classification techniques, it is also a support vector machine class, but multi-class classification is performed. Combination of all binary class used to achieve the multi-class SVM. One data compared with another line like a one versus other approach. The neighboring  $n$  distance is calculated and stored in a single vector [5]. Since, it is a virtual keypad input in the air. The processing model is little different from the other techniques. The virtual keypad is developed in Japanese and English language, and it is an alphabet-based recognition technique. Fingertip detection algorithm was used to obtain right and left hands gestures for the input. It consists of different letter and number, using this virtual keypad on the air we can select Japanese, English letters and numbers. The type will be different characters with voice input and semi-voice is also used for Japanese letters. The position between the input sensor and user must be 1–2 m [17]. From the input video, features are extracted then it aligns gesture using sequence learning model. Video sequence having the gesture label for the easy representation of the sign data [23].

### 3.3 Output System

When it comes to output system, it has two type of output system type (i) converting sing gesture into voice or text-based output and (ii) voice or text input to sign gesture output method. Figure 4 shows the communication is between normal people and deaf and dumb people used 18,000 sign image data with 20 classes, and 70% of the sign image data are used to train the machine, and remaining 30% has been used for testing phase. Using the multilayer SVM classifier, 97.28% accuracy has been achieved. Since it is an isolated word-based recognition, some words achieved highest accuracy [2]. Figure 4 shows the bi-directional way of sign language recognition system. Based on the size of training, the accuracy will be different here. Based



**Fig. 4** Bi-direction communication between normal people and deaf and dumb

on the mean standard, the accuracy will differ. Here, the subject is classified into five part, and the total sign language word here is 110 total component accuracy is 84.9% for all five subjects [18]. The total number of sign image used here is 720, and they proposed a confusion matrix with  $n$  column and  $n$  rows it represent the target and system response. Since it is an alphabet-based sign recognition, some letter has same representation like the letter 'L' 'Q' 'T' these letters has very less similarities compare to the other vertical properties. In this best case, it has 80% accuracy rate which is good in this ASL alphabets recognition, and some letter like B has worst accuracy rate. B has only 30% accuracy rate. Table 1 shows the different types of recognition models and input types. Obtained 91.79%, 95.96% for GCM and HGCM [12]. This model has deep CNN next to the temporal operation. Bi-LSTM used for the learning process [23].

## 4 Discussion and Conclusion

SLR has been a prolonged study that bring into being eras in the past, nonetheless not yet any other system developed for large scale. In deaf and hearing communication, SLR system has been certainly consuming a marvelous influence. Furthermost of the reviews testified attaining extraordinary recognition rates on the other hand, using certain dares that hold back the systems as of reaching a greater latent. The principal route problem for thinking through dependent taking place in the SLR category is that nonstop gesture-based communication acknowledgment is to slack than segregated word acknowledgment. In spite of the fact that secluded word acknowledgment indicated great execution, constant gesture-based communication acknowledgment is the one of the most in conveying an ongoing correspondence framework between hard of hearing quiet people and hearing people. By means of this, the linking breach

**Table 1** A summary of the SLR system existing work

Isolated/continuous	Recognition model	Input device	Dataset	Authors
Isolated	SVM	Camera	Self-built	Md Abdur Rahim [2]
CSL alphabets	Fuzzy k-means	SEMG&GYRO	Self-built	Wei et al. [18]
ASL alphabets	ANN	Camera	Self-built	Rivera-Acosta [3]
Isolated	PCA with hand descriptor	Kinect	Public	Elpeltagy et l. [6]
Continuous	Modified LSTM	Leap motion	Self-built	Mittal et al. [10]
Isolated	Motion-based classification	Multiple camera	Self-built	Kishore [13]
Continuous	Bi-directional LSTM	Camera	Public	LIAO [11]
Isolated	Encoder decoder network	Kinect	Self-built	Huang [15]
Isolated	ANN	Leap motion controller	Public	Chong and Lee [16]
Isolated	GMM model	LMC device	Self-built	Deriche [19]
Isolated	PCA net with SVM classification	Kinect	Public	Aly [24]
Isolated	CNN	Kinect	Self-built	Oliveira [20]
Isolated	Robust geodesics-based shape recognition	Kinect	Public	Nasreddine [21]
Isolated	DLSTM	LMC device	Public	Avola et al. [22]
Isolated	Multi-class SVM	Camera	Public	Saqlain Shah [5]

will be clean. Device handbags announced extraordinary precision yet for the most part utilized for detached word acknowledgment. Despite the fact that it is simpler to separate the pertinent highlights from the information gained commencing the devices, they are not prominently utilized. They are costly and awkward for the endorser. Greater part of SLR investigation uses the Kinect, with the additional profundity and direction data, precision seems to be expanded. All things considered, normal cameras ought to be acknowledged for the errand as the primary point is to build up a moderate and simple to-utilize SLR framework. As of now, (1) sentence-based recognition is great for deaf and dumb people and as well as for the normal people (2) compare to the wearable device, non-wearable-based SLR is cost effective.

## References

1. <https://www.who.int/news-room/fact-sheets/detail/deafness-and-hearing-loss>

2. Md. Abdur Rahim, Md. Rashedul Islam, J. Shin, Non-touch sign word recognition based on dynamic hand gesture using hybrid segmentation and CNN feature fusion. *Appl. Sci.* [MDPI] **Appl. Sci.** **9**, 3790. <https://doi.org/10.3390/app9183790>
3. M. Rivera-Acosta, S. Ortega-Cisneros, J. Rivera, F. Sandoval-Ibarra, *Sensors* [MDPI]. *Sensors* **17**, 2176 (2017). <https://doi.org/10.3390/s17102176>
4. W. Aly, A. Saleh, S. Almotairi, User-independent American sign language alphabet recognition based on depth image and PCANet Features. *IEEE Access, Dig. Object Ident.* <https://doi.org/10.1109/ACCESS.2019.2938829>
5. J. Shin, C. Min Kim, Non-Touch Character Input system based on hand tapping gestures using kinect sensor. *Dig. Obj. Ident.* <https://doi.org/10.1109/ACCESS.2017.2703783>.
6. M. Elpeltagy, M. Abdelwahab, M.E. Hussein, A. Shoukry, A. Shoala, M. Galal, Multi-modality-based Arabic sign language recognition. *IET Comput. Vis.* [The Institution of Engineering and Technology 2018] **12**(7), 1031–1039 (2018)
7. L. Chen, B. Muthu S. Cb (2020) Estimating snow depth inversion model assisted vector analysis based on temperature brightness for North Xinjiang region of China. *Eur. J. Remote Sens.* **1–10**<https://doi.org/10.1080/22797254.2020.1771217>
8. S.-I. Chu, B.-H. Liu, N.-T. Nguyen, Secure AF relaying with efficient partial relay selection scheme. *Int. J. Commun. Syst.* **32**, e4105 (2019). <https://doi.org/10.1002/dac.4105>
9. H. Wang, X. Chai, X. Chen, A novel sign language recognition framework using hierarchical grassmann covariance matrix. *IEEE Transactions Multimedia* **21**(11) (2019)
10. A. Mittal, P. Kumar, P. Pratim Roy, R. Balasubramanian, B.B. Chaudhuri, A modified LSTM model for continuous sign language recognition using leap motion. *IEEE Sens. J.* **19**(16) (2019)
11. Y. Liao, P. Xiong, W. Min, W. Min, J. Lu, Dynamic sign language recognition based on video sequence with BLSTM-3D residual networks. Special section on AI-driven big data processing: theory, methodology, and applications, Digital Object Identifier. <https://doi.org/10.1109/ACCESS.2019.2904749>
12. R. Cui, H. Liu, C. Zhang, A deep neural framework for continuous sign language recognition by iterative training. *IEEE Trans. Multimedia*, **21**(7)
13. P.V.V. Kishore, D. Anil Kumar, A.S. Chandra Sekhara Sastry, E. Kiran Kumar, Motionlets matching with adaptive kernels for 3-D Indian sign language recognition. *IEEE Sens. J.* **18**(8) (2018)
14. D. Anil Kumar, A.S.C.S. Sastry, P.V.V. Kishore, E. Kiran Kumar, M. Teja Kiran Kumar, S3DRGF: spatial 3-D relational geometric features for 3-D sign language representation and recognition. *IEEE Sig. Process. Lett.* **26**(1), (2019)
15. S. Huang, C. Mao, J. Tao, Z. Ye, A novel chinese sign language recognition method based on Keyframe-centered clips. *IEEE Sig. Proc. Lett.*, **25**(3)
16. T.-W. Chong, B.-G. Lee, American sign language recognition using leap motion controller with machine learning approach. *Sensors* [MDPI]. *Sensors* 2018, **18**, 3554. <https://doi.org/10.3390/s18103554>
17. S. Aly, W. Aly, DeepArSLR: a novel signer-independent deep learning framework for isolated Arabic sign language gestures recognition. *IEEE Access, Dig. Obj. Ident.* <https://doi.org/10.1109/ACCESS.2020.2990699>.
18. S. Wei, X. Chen, X. Yang, S. Cao, X. Zhang, A component-based vocabulary-extensible sign language gesture recognition framework. *Sensors* [MDPI]. *Sensors* **16**, 556. <https://doi.org/10.3390/s16040556>
19. M. Deriche , S.O. Aliyu, M. Mohandes, An intelligent arabic sign language recognition system using a pair of LMCs with GMM based classification. *IEEE Sens. J.* **19**(18) (2019)
20. K. Nasreddine, A. Benzinou, Shape geodesics for robust sign language recognition. *IET Image Proc.* **13**(5), 825–832, IET Image Process (2019)
21. D. Avola, M. Bernardi, L. Cinque, G. Luca Foresti, C. Massaroni, Exploiting recurrent neural networks and leap motion controller for the recognition of sign language and semaphoric hand gestures. *IEEE Trans. Multimedia*, **21**, (1) (2019)
22. S. Muhammad Saqlain Shah, H. Abbas Naqvi, J. I. Khan, M. Ramzan, Zulqarnain, H. Ullah Khan, S. Based Pakistan Sign language categorization using statistical features and support

- vector machines. IEEE Access, Dig. Object Ident. <https://doi.org/10.1109/ACCESS.2018.2872670>
- 23. Y. Li, X. Wang, W. Liu and Bin Feng, Pose Anchor: A Single-Stage Hand Keypoint Detection Network. *IEEE Trans. Circ. Syst. Video Technol.* **30**(7) (2020)
  - 24. T. Oliveira, N. Escudeiro, P. Escudeiro, E. Rocha, F. Maciel Barbosa, The virtual sign channel for the communication between deaf and hearing users. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, **14**(4), (2019)

# Refined Search and Attribute-Based Encryption Over the Cloud Data



M. S. Iswariya and K. Annapurani

**Abstract** In recent years, cloud technology is used everywhere to store data. For data security reasons, the data are encrypted, but this adds complexity in searching the data in cloud which is in encrypted form. Generally, searching a data is done using the single keyword search, but the data retrieved is not precise. Also, this results in high searching cost due to scanning of huge amount of data. The search result will be adulterated search and lack of security. The time token methodology also adds more process time to the computation. Lack of rank mechanism is also a major drawback here, as user takes long time to select the data. We have proposed multi-keyword ranked search, and attribute-based encryption. This technique will help us to give a refined search and also, it encrypts the keyword that is used for searching in such a way that the keywords are not exposed for any privacy concerns.

**Keywords** Multi-keyword ranked search · Trapdoor · Co-ordinate matching · Searchable encryption · Attribute-based encryption

## 1 Introduction

Cloud computing is a type of Internet-based computing which provides many data and shared resources to computers and other devices. Cloud computing provides flexibility and also cost efficient. Cloud computing has a number of benefits like rapid elasticity, on-demand, resource pooling services, broad network access and pay as you use. Cloud computing is a shared pool of resources to access from or store data to a remote place. Generally, the data are stored with keyword indices; therefore, the data users can make keyword search over them and retrieve the desired data. The query frequency analysis of keywords, in many database applications, is

---

M. S. Iswariya · K. Annapurani (✉)

Department of Networking and Communications, SRM Institute of Science and Technology,  
SRM Nagar, Kattankulathur, Kanchipuram, Tamil Nadu 603203, India  
e-mail: [annapook@srmist.edu.in](mailto:annapook@srmist.edu.in)

M. S. Iswariya  
e-mail: [im9456@srmist.edu.in](mailto:im9456@srmist.edu.in)

very important to improve user experience and increase user volume [1]. Searching of a data can be done more precisely by enhancing the keyword settings [2]. In order to make search faster, the data storage is optimized so that most habitually query will generate.

Cloud computing has some issues but that has to be handled efficiently. The two main issues are privacy and security. Privacy is the obstruction that proscribe the adoption of the cloud by many end users. Even though sensitive data can be protected by firewalls, intrusion detection systems and many other tools, privacy is still not fully attainable. With the approach of cloud computing, the data owners out their sensitive data from locals to the cloud. Encryption of the data prior to re-appropriating is the best methodology, to secure the data confidentiality. The current methods on keyword based data recovery cannot be applied legitimately on encoded information. Downloading and decrypting the data locally is exceptionally illogical. To have successful data recovery, a lot of archives request the cloud server to perform applicable relevant ranking. The rank-based search system empowers data user to locate the most appropriate information quicker. This can wipe out undesirable network traffic by sending back the most significant information. To improve the exactness of the search result and to upgrade the end user searching experience, it is essential for such positioning framework to help multi-keyword search since single keyword search yields extremely coarse outcomes. A protected tree-based inquiry over the encoded cloud data is proposed to underpin multi-keyword ranked search over cloud information for end users.

## 2 Literature Survey

### 2.1 *Different Encryption Techniques Used Over Cloud Data*

It is prudent to keep data on storage servers like document servers and mail servers in encoded structure to lower protection dangers and security. In any case, this normally infers one's needs to forfeit usefulness for security and report recovery. The proposed system enables the effective searching of encrypted data [3]. A basic path is to likewise encode the record pointers in each index of the list. Thus, when Bob looks for E(W) and finds a match, he returns Alice the encoded index of coordinating situations from the list. Alice may decode the encrypted passages and send Bob another request to recover the applicable records. To recover the reports, Alice needs to invest an extra round trip time is the greatest disadvantage. In the event that Alice would not like to hang tight for an extra round trip time, or if Alice might want Bob to consolidate the result of search queries. For instance, she may encode the index of record pointers in the file utilizing a key kW. Consequently, when Alice needs to look for the word W, she will uncover to Bob. It is smarter to keep the arrangements of pointers in a fixed-size list to keep Bob from doing factual investigation on the file. Accessible symmetric encryption (SSE) [4] conveys the information to another in a private way,

while keeping up the capacity to specifically look over it. Dynamic examination and developments have been proposed to reduce the difficulties raised around security. We think of new ill-disposed models for SSE which is versatile and nonadaptive. Firstly, versatile considers foes that pick their inquiries as an element of recently acquired secret entryways and search results. Also, nonadaptive just considers foes that make their inquiry inquiries without considering the hidden entrances and search results of past pursuits. All past work on SSE falls inside the non-versatile setting (with the exemption of neglectful RAMs). The suggestion is contradictory to the regular utilization of accessible encryption depicted in, these definitions just assurance security for clients that play out the entirety of their hunts without a moment's delay.

The IPE (Inner Product Encryption) [5] schemes are malleable, secure and feeble attribute-hiding. We put forward the inner product encryption shows full attribute-hiding and security. The ABE is a mainstream strategy for authorizing access control approaches through cryptographic methods. Fundamentally, this strategy permits substances with legitimate qualifications to decode a cipher text that was encoded according to an entrance manage approach [6].

Cloud technology are used in many fields to store large amount of data. Though it gets rid of heavy data management problem, security and privacy is the biggest concern of [7] cloud computing. This will cause damage to industrial enterprise. To achieve that we proposed an attribute-based encryption (ABE) scheme to protect the privacy of user during key issue and store the data. The ABE scheme make use of key generation centre and attribute auditing centre. This paper protects the user privacy during the generation of key. Near duplicate detection (NDD) is utilized for resource utilization and traffic issue in network. Despite the fact that data protection in network is fruitful by encoding, yet it needs finding encrypted near duplicate data [8]. So we propose a safe and compelling near duplicate detection system. We utilize locality sensitive hashing to hash comparative info things into a similar input item with high likelihood. Since comparable things end up in similar bucket, this method can be utilized for data clustering and closest neighbour search. In numerous IR undertakings, archive closeness alludes to semantic importance among records, which could be linguistically totally different however pertinent.

## 2.2 *Different Searching Techniques Used in Cloud*

The PHRs are developing rapidly day by day, individuals outsource PHR systems to the cloud to facilitate management. The MEIM mechanism is employed to accomplish efficient [9] encryption of data on cloud. The advantage of MEIM is the PHRs can be efficiently retrieve by issuing one encrypted query. Index privacy is considered to know the content of data stored in platform and to generate valid queries. At present, the searchable encryption is separated into two types as SSE and PKE. Searchable encoding [10] approach as a two-layered system depends in the symmetric

encryption. Boneh et al. from the work of Song et al. proposed asymmetric encryption system. It introduced an approach of searching for conjunctive keywords to accomplish multi-keyword searching. Nevertheless, all of the above strategies were based on accurate matching so that the efficiency of usability could be minimized. To solve the challenges, the fuzzy keyword [11] monitoring is carried to replace the matching keywords with the maximum similarity score. Designing a cipher text search is more difficult. So, we propose a hierarchical clustering method to support more search semantics. This will cluster the documents and partition the resulting cluster into sub-cluster. Hybrid encryption is a type of encryption that combine many encryption system. Hybrid encryption secures the data but encrypting [12]. In place of single encryption, we make use of hybrid encryption because it secures the data when compared to other mechanism. Using AES and FHE approach, the user can protect data confidentiality, integrity and privacy from hackers.

A secure index empowers to check whether the substance of semantically secure cipher texts incorporates a predetermined keyword without applying decoding operation or [13] revealing the keyword value. The data structure allows a trapdoor and index with the querier which will not reveal any information about its contents. Here, the trapdoors can be generated with a secret key. Searchable encryption protects users' data. Searchable encryption (SE) [14] allows the server to search encrypted data without leaking information in plaintext. The searchable encryption hold up only Boolean search, which allows the users to combine keywords with operators [15]. To improve the search efficiency different keyword search methods have been used. But these methods show large overhead, such as computational cost by linear map or communication cost by sharing secret. A key role of encryption is to secure the data. In an encrypted data searching [16], a document is a challenging one, to overcome this problem many techniques are available. Here, we develop a search index dependent on the term frequency and the vector space model to have high output. To improve the proficiency of search, a tree-based index structure is proposed, so the search will be far superior to the search pursuit strategy. Here, two secure index schemes are used; they are known background model and known cipher text model.

### **2.3 Access Control in Cloud Security**

Access control is a security technique, i.e. for protection and privacy in cloud data. Access control will prevent unauthorized users' to enter into cloud data. Data authentication, authorization and auditing were affected by an ineffective access control process. We present another technique for giving secured access control in cloud computing. This model gives a safe access control in cloud computing. To give more secured access control, it embraces a hierachal structure and it utilises a clock. Utilizing this, we can undoubtedly transfer, download and erase records from the cloud. The conventional model for access control is application-centric access control, where every application monitors its assortment of users and oversees them, is not practical in cloud-based structures. Since in this strategy, we need a ton of

memory for putting away the user details, for example, username and secret word. So cloud requires a client-centric access control where each user solicitation to any service provider is packaged with the user identity and qualification data [17]. Discretionary access control need not be strictly controlled, so it achieves a flexible access control for cloud users. Instead of artefacts, the system administrator is responsible for control access so that the policy focused on trust rather than integrity. Access privileges were allocated to subjects in the RBAC model on the basis of their positions and roles in the system rather than their identity [18]. Because of the absence of different parts of the points, the role-based access control causes downside. To conquer these issues attribute-based access control was proposed. The access rule was done dependent on the attribute analysis of articles and subjects [19]. With its careful thought during verification, ABAC's major advantage was significant. Authentication process for attribute-based access control is a tedious activity.

## 2.4 Taxonomic Relations

Thesauri, taxonomies or idea progressions are a basic part of numerous applications inside the semantic Web, information retrieval, knowledge management, text clustering and natural language processing and information systems all in all. Indeed, there has been a long custom in artificial intelligence and related fields, for example, natural language processing or information retrieval to consequently take in scientific categorizations from data [20]. Most analysts have attempted to learn scientific classifications by literary contribution as text reports are hugely accessible. This paper investigate the chance of learning ordered relations by joining the assertion from various methods and sources utilizing a characterization approach [7, 21, 22]. The urgent inquiries we address in this paper are (I) how to change over the aftereffects of various methodologies and the various wellsprings of proof into first-request highlights which can be utilized by a classifier, (ii) which classifiers perform best on the errand and (iii) which techniques are generally reasonable to manage the uneven datasets that we consider.

## 3 Proposed System

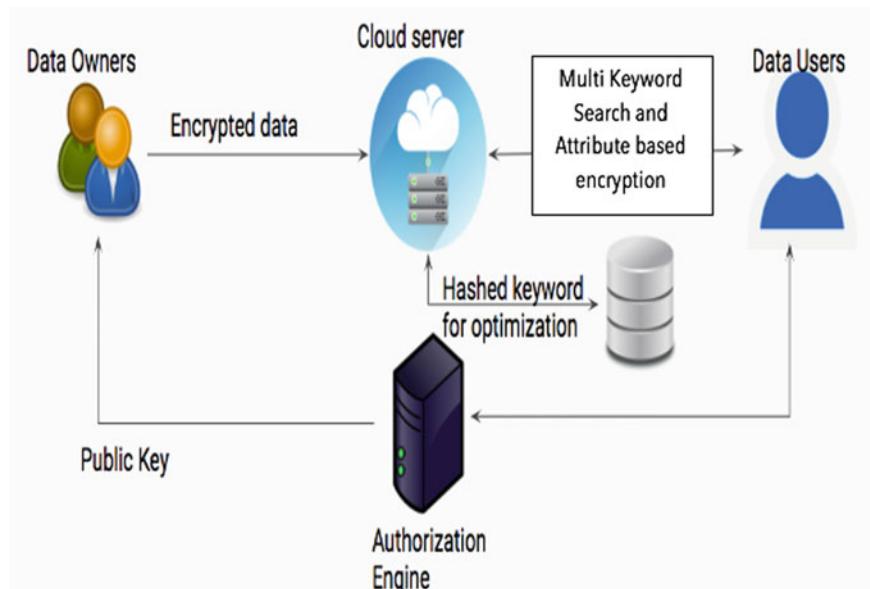
### 3.1 Architecture of the Attribute-Based Encryption System

The proposed system is to allow MRSE and encode search keyword for additional security. Generally, the encrypted data are searched using single keyword; the idea here is to use multiple keyword searches in which it gives a refined search result and improves the efficiency of searching, i.e. time taken to search a data is less when compared to single keyword search. The paper [23] discusses about the time token

issued from the time server to the authorized user to search the data. Due to it, the time taken to retrieve the data is increased. In our paper, instead of time token, we have used multiple keywords that will retrieve the appropriate data and in less time. Also, encryption of the keyword is done based on attribute encryption in such a way that the keywords are not exposed for privacy concerns and also adds additional security. Multi-keyword ranked search is used to configure search scheme which permit multi-keyword query and give positioning to successful data recovery. Privacy preserving is utilized to keep the cloud server from knowing extra information from the dataset just as record and to meet protection. Attribute-based encryption is a public key encryption where the secret key and cipher text rely on the characteristics.

Figure 1 shows the architecture of the refined search and encrypted keyword over cloud. Data owner encrypts the data with public key which is provided by the authorizer, and then the encrypted data is uploaded in the cloud server. The authorization engine authorize the data user to make search over the encrypted data [23]. The attribute-based encryption technique will encrypt the search keyword so that it is not exposed for any privacy concern which gives additional security. The multi-keyword search allows us to give a refined search. The hashed keyword helps us find optimized keyword which is stored in a database.

The various modules of the system are encryption of the data and keyword, multi-keyword search. The encryption of data and keyword are finished using attribute-based encryption in which the secret key and cipher text rely on the attributes, so decoding of cipher text is just conceivable only if the set of attributes of the user key matches the attributes of cipher text. Multi-keyword search is the point at which a



**Fig. 1** Architecture of refined search and encrypted keyword over cloud data

user searches and records for numerous varieties of same keyword. Among different multi-keyword semantics, we pick the effective rule of coordinating matching, for example, to discover how many matches could be reasonably expected, to catch the similitude between search query and data reports.

## 4 Result and Discussion

Summarizing the above referred papers, it is found that it has limited security for the data that are stored over the cloud and also has less efficiency. In the existing papers, authorization are based on the time token to validate the authorized user, but the drawback of using token is it takes more time for the overall process. Thus, in this paper, the proposed model will encrypt the keyword that is used for searching so that the search data is hidden and cannot be accessed by unauthorized user. Quicker searching of data because of multiple search keyword.

## 5 Conclusion

This paper elaborates the existing methodology that is used to search the encrypted data and the retrieval of queried keyword and the authorization process for the data users. Several models based on MRSE and attribute-based encryption have been discussed in this paper. Though existing models are achieving results, researchers still have great opportunity to explore more in this area and come up with models resolving the limitations of the existing models. We implement multi-keyword rank search in order to give a refined search and also a time-saving search technique. In addition to that, our proposed method encrypts the keyword that is used for searching in such a way that the keywords are not exposed for privacy concerns and also adds additional security.

## References

1. B.S. Kavya, P.V. Shrilakshmi, N. Sushmitha, Yamuna, Multi-keyword search methodology for cloud data. *IJARCCE Int. J. Adv. Res. Comput. Commun. Eng.* **6**(4)
2. V. Kulkarni, P. Pise, Secure multi-keyword ranked search over encrypted cloud data. *Int. J. Adv. Res. Comput. Commun. Eng.* **5**(12) (2016)
3. D. Song, D. Wagner, A. Perrig, Practical techniques for searches on encrypted data, in *Proceeding 2000 IEEE Symposium on Security and Privacy. S&P 2000* (IEEE, 2000), pp. 44–55.
4. G. Poh, J. Chin, W. Yau, K. Choo, S.M. Mohamad, Searchable symmetric encryption: designs and challenges. *ACM Comput. Surv.* **50**(3), 40 (2017)
5. Y. Yang, M. Ma, Adaptive keyword search over the cloud data. *IEEE Trans. Inf. Forens. Sec.* **11**(4), 746759 (2016)

6. Q. Zheng, S. Xu, G. Ateniese. VABKS: verifiable attribute-based keyword search over outsourced encrypted data. INFOCOM 2015, IEEE Computer Society, pp. 522–530
7. M. Qiu et al., Proactive user-centric secure data scheme using attribute-based semantic access controls for mobile clouds in financial industry. *Fut. Generat. Comp. Syst.* **80**, 421–429 (2018)
8. H. Cui, X. Yuan, Y. Zheng, C. Wang, Enabling secure and effective near-duplicate. Detection over encrypted in-network storage, in *IEEE INFOCOM 2016 - The 35th Annual IEEE International Conference on Computer Communications*, vol. 6, no. 9, (July 2016)
9. X. Yao, Y.L. Qin Liu, S. Long, Efficient and privacy preservation search in multi-source personal health record clouds, in *IEEE Symposium on Computer and Communications* (2015)
10. J. Koodi, G. Srinivasachar, Privacy-preserving multi-keyword ranked search over encrypted cloud data. *Int. Res. J. Eng. Technol. (IRJET)* **02**(03) (2015)
11. J. Li, Q. Wang, C. Wang, N. Cao, K. Ren, W. Lou, Fuzzy keyword search over encrypted data in cloud computing, in *2010 Proceedings IEEE INFOCOM* (IEEE, 2010), pp. 1–5
12. L. Kumar, N. Badal, A review on hybrid encryption in cloud computing, in *4th International Conference on Internet of Things: Smart Innovation and Usages* (2019)
13. E.-J. Goh, Secure indexes, *Cryptology ePrint*, March 16, 2004
14. K.D. Sonam, M.K. Kulkarni, Multiuser multi-keyword ranked search over encrypted cloud using MHR andKP- ABE. *Int. J. Comput. Sci. Trends Technol. (IJCST)* **4** (4) (2016)
15. G. Karthika Priya Dharshini, D. Viji, K. Saravanan, Seclusion search over encrypted data in cloud storage services. *Int. J. Comput. Sci. Mob. Comput.* **4**(3) (2015)
16. J. Francis, R. Bansod, C. Getme, P. Bagde, A secure and encrypted cloud data with multi-keyword rank search and revocation of user. *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* **6**(9) (2016)
17. B.K. Onankunju, Access control in cloud computing. *Int. J. Sci. Res. Publ.* **3**(9) (2013)
18. J. Lopez, J. Rubio, Access control for cyber-physical systems interconnected to the cloud. *Comput. Netw.* **134**, 46–54 (2018)
19. L. Hu, N.-T. Nguyen, W. Tao, M.C. Leu, X. Frank Liu, Md.R. Shahriar, S.M. Nahian Al Sunny, Modeling of cloud-based digital twins for smart manufacturing with MT connect. *Procedia Manuf* **26**, 1193–1203 (2018). ISSN 2351-9789. <https://doi.org/10.1016/j.promfg.2018.07.155>
20. P.T. Hong Doan, N. Arch-int, S. Arch-int, A semantic framework for extracting taxonomic relation from corpus. Corpus ID:211566071, *Comput Sci., Int. Arab. J.Inf. Technol.* (2020)
21. S. Zhu, V. Saravanan, B. Muthu, Achieving data security and privacy across healthcare applications using cyber security mechanisms. *Electron. Libr.* **38**(5/6), 979–995 (2020). <https://doi.org/10.1108/el-07-2020-0219>
22. P. Cimiano, A. Pivk, S.S. Schmidt-Thieme, Learning taxonomic relations from Source of evidence. Corpus ID:648922, *Computer Science* (2005)
23. X. Lingling, W. Li, F. Zhang, R. Cheng, S. Tang (2019) Authorized keyword searches on public key encrypted data with time controlled keyword privacy. *IEEE Trans. Inf. Forens. Secur.* **15**(04)

# Facial Recognition Techniques and Their Applicability to Student Concentration Assessment: A Survey



Mukul Lata Roy, D. Malathi, and J. D. Dorathi Jayaseeli

**Abstract** In recent times, the daily life of the people belonging to educational institutions has shifted online. With the benefit of Internet and modern technology, it has been made possible to conduct classes and examinations remotely. However, this kind of online learning lacks the benefit of interactivity and communication that classroom learning has. In order to improve the experience of online learning, teachers may find it helpful to have some way of being automatically alerted when a student appears to stop paying attention during sessions. Machine learning techniques applied to fields like object detection and face expression recognition (FER), in conjunction with handcrafted features like histogram of oriented gradients (HOG) and local binary pattern (LBP), have seen great success in the past. This paper takes a look at several such techniques and their merits in trying to create a solution for this problem of dwindling student concentration due to lack of human supervision during online classes.

**Keywords** Face expression recognition · Student concentration · Student attention · Drowsiness detection

## 1 Introduction

The onset of the COVID-19 pandemic has forever changed the way that institutions of business and education operate. Prior to the pandemic, many situations involving

---

M. L. Roy (✉) · D. Malathi · J. D. Dorathi Jayaseeli

Department of Computer Science and Engineering, SRM University, Chennai, India

e-mail: [mr9991@srmist.edu.in](mailto:mr9991@srmist.edu.in)

D. Malathi

e-mail: [malathid@srmist.edu.in](mailto:malathid@srmist.edu.in)

J. D. Dorathi Jayaseeli

e-mail: [dorathij@srmist.edu.in](mailto:dorathij@srmist.edu.in)

online video calls existed, from cross-continental business meetings, to cyber classrooms, to simple personal video chats. However, now nearly the entirety of student life and daily work has been moved to the Internet.

On one hand, there are clear benefits to this mode of learning: students can easily access materials and contact their teachers via an Internet connection, which is widely available, and they can go about their work from the safety of their homes, and so avoid the dangers of being infected by the virus. However, these circumstances also make it difficult for teachers to judge the engagement of students in their sessions, as classroom learning provides a level of connection and interaction that cannot be replaced.

Distractions and disinterest become quite common among young adults who attend online classes from home. This may detrimentally affect not only the teachers, who must struggle to do their due diligence, but also the students, who lose out on the material that is taught as well as the opportunity to engage with the teachers and clear their doubts. It would be helpful for teachers to have some kind of solution so that they would be alerted to a student's lack of alertness during their sessions.

To this end, this paper will aim to survey various existing FER techniques that can help toward developing a human concentration level recognition model to judge whether a student is drowsy or distracted during online video classes using machine learning techniques.

## ***1.1 Face Recognition and Face Expression Recognition***

The faces of humans and animals are arguably their most significant identifying feature. In different research fields, face recognition has generated a lot of interest and presented many challenges. As a part of the field of pattern recognition, face recognition has been extensively applied in many areas, including in human computer interaction, security footage analysis and threat detection, driver's license databases, criminal justice systems, and so on. This interest has resulted in a large number of proposed techniques and algorithms.

A face recognition system is one that is able to identify a human face from the medium of a digital image. Although there are many different methods using which this might be done, the main steps consist of extracting select facial features from an image, and comparing them to pre-existing images in a database. It can be compared to other biometric techniques such as iris or fingerprint recognition; although, its accuracy is lower than its counterparts. Recently, it has become a popular tool adopted for commercial applications, image indexing, video surveillance, and even photo filters on phone apps.

For humans, expressions that are visible on our faces is a significant non-verbal means of communicating our inner feelings and intentions to other humans and is thus an important part of society. In 1974, Mehrabian [1] indicated that, of the signals related to emotions and thoughts, approximately 55% is contributed via facial expressions. Further, 7% of the same is expressed verbally through words, and the

remaining are what is known as “paralinguistic,” that is, the manner in which the words are spoken.

The rise of artificial intelligence has led to intensive study of automatic recognition of expressions in recent years [2]. FER has gained much attention from academics and professionals from different fields, including computer science, criminal justice, and psychology. It is a technology that has been incorporated in several domains, such as virtual and augmented reality, driver assistance systems, education, and entertainment. In general, techniques surrounding FER can be categorized based on how features are extracted from a facial image. The two methods of feature extraction are either a conventional FER approach or a deep learning FER approach.

In literature, there are several existing methods regarding the different types of FER. These methods can be broadly classified as follows:

- Conventional approach
- Deep learning-based approach

In a conventional FER approach, the extraction of features from an image of a face is done manually. The general process of FER consists of image preprocessing, feature extraction, and classification of expression. Since the conventional FER methods involve manual extraction of features, they have the benefit of not needing to depend as much on data and hardware. This makes them better for small data sample analysis. On the other hand, manual extraction of features is a grueling task and will fall short when faced with a larger data sample.

On the other hand, in deep learning-based FER approaches, feature extraction is achieved by generation through the output of neural networks. This type of approach reduces the difficult task of manually extracting features as it employs something of an end-to-end learning process which directly produces an end result from an input image. This type of approach is best suited to massive datasets, which is both a blessing and a curse. Using a smaller dataset could easily cause overfitting. However, a larger dataset will provide a more efficient model as it will have a large amount of data to use for learning.

When considering general FER techniques in terms of the available training data, there are several existing image datasets that are commonly used by researchers. The most popular of these seems to be the JAFFE (Japanese Female Facial Expressions) dataset, consisting of a total of 213 images, with 7 expressions from 10 subjects [3]. The Cohn-Kanade (CK) and the Extended Cohn-Kanade (CK + ) [4] databases contain 486 images with 6 emotions and 593 images with 8 emotions, respectively. The Multimedia Understanding Group (MUG) dataset [5] contains 1462 high resolution unobstructed images with 7 expressions. The AR face dataset [6], MMI dataset [7] and Taiwanese Facial Expression Image Database (TFEID) [8] datasets contain 4000, 250 and 7200 images, respectively.

## 1.2 The Standard FER System

A standard facial expression recognition system has three main stages:

- Detect the face from an image or video frame
- Extricate the main features of the face
- Classify the expression from the features

The classification works by comparing the extracted information from the input data to the pre-existing trained and labeled data.

In most existing papers, these are the steps that are followed, with the only variations being the individual methods that can be used at each stage. Besides these steps, the input images may also be preprocessed to prepare them for easier classification, as data is often infused with noise or other irrelevant information.

## 1.3 Organization of the Paper

Section 2 of this paper consists of a discussion of existing methods for FER, various techniques for gauging driver fatigue, as well as previously reported student concentration algorithms. Section 3 will present the inferences from the reviewed techniques in their possible application for student concentration detection, while Sect. 4 discussed a possible framework to detect the alertness of students during online classes. Finally, Sect. 4 concludes the paper.

## 2 Review of State-of-the-Art Methods

Image pre-processing eliminates irrelevant information and enhances the detection ability of relevant information from the input image. This step includes methods such as noise reduction, face detection, normalization, and histogram equalization. The technique of extracting relevant information from the image is known as feature extraction.

These resulting “non-picture” representations of face features are obtained from extraction techniques like Gabor feature extraction, local binary pattern (LBP), optical flow method, Haar-like feature extraction, feature point tracking, and so on. Another vital aspect for influencing the rate of recognition is the way to choose the fitting classifier that can effectively learn and predict the relevant features in a face. kNN (k-nearest neighbors), SVM, AdaBoost, and Bayesian classifier are a few examples of popular classifiers that are often incorporated in FER frameworks.

Among conventional FER approaches in the literature, several techniques have been proposed. In [9], the authors discuss Haar features and local binary patterns (LBP), both used in face detection techniques. They utilized three image databases,

namely the Color FERET, MIT CBCL, and Taarlab databases, to present a comparison between the techniques. After their experiments, the authors concluded that LBP surpassed Haar features in terms of both detection speed and number of detected faces. A discriminative spatiotemporal local binary pattern based on an integral projection is proposed to resolve the problems of STLBP for micro-expression recognition, which often miss the shape attribute of face images and ignore the discriminative information between two micro-expression classes that has been proposed in [10]. The authors of [11] offer an efficient new way to represent face images by using LBP for the extraction of features. In this paper, the authors proposed segmenting the image into a collection of regions and then applying the LBP for extraction, after which they are combined into an improved feature vector to be used as a descriptor for faces. This technique is also applicable to other object recognition and detection jobs.

The paper in [12] presents face detection and recognition algorithm to identify/recognize the wanted person in a surveillance video. Face detection is achieved via the Viola–Jones algorithm, and clustering is then applied to cluster the face parts. HOG and LBP features are obtained for the detected face parts. The contribution of the work is to implement HOG and LBP for surveillance video and combine both the features to address the issues such as pose variations, illumination changes, expression changes, and occlusion for face recognition. The SVM classifier is then applied to distinguish strong vs. weak features, of which the strong features are chosen as the descriptors. The authors of [13] use local binary patterns for recognition. The major contribution from this technique is the method of selection of features, where the pixels selected to describe faces are LBP pixels with high variance. This decision to choose the high variance pixels has resulted in vastly improved recognition rates. The BU-3DFE database was used for this project.

A multilevel ellipse detector, as well as the usage of SVM as a verifier, distinguishes the paper in [14], allowing for a face and eye detection system that is highly precise. This algorithm resulted in improved accuracy in the detection of eyes from good quality images. Since the goal of any solution to a problem requires the maximum results in the smallest time period with minimal errors, the authors in [15] succeeded in optimizing the parameters to the Viola–Jones algorithm. They proposed that using a combination of cascading classifiers instead of a single definite one allowed the algorithm to gain higher accuracy in a shorter time of operation. In [16], the authors offer a technique to not only recognize, but also monitor the emotions of a student within an e-learning environment. Additionally, the algorithm provides a technique to generate feedback in real-time to further enhance e-learning aids. The goal is to perform continuous monitoring of the movements of the eyes and head to ascertain the level of focus of the student. The authors of [17] use Grayscale, Gaussian filter, and valley extraction to detect and localize both irises from color images.

Various techniques along the lines of a deep learning-based approach to FER have also been created. The authors of [18–20] perform both face detection as well as recognition using a technique based on a (CNN) which is able to outperform older methods. They use a region proposal network (RPN) to detect faces, which

was trained using the Labeled Faces in the Wild (LFW) dataset. In [21], an attempt was made to categorize the emotions of the subjects by using the EEG signals from the DEAP database. They also used principal component analysis to perform feature reduction of the processed EEG data and evaluated the efficiency of the classification by using the CNN algorithm. In [22], the authors use local binary patterns histogram (LBPH) descriptors for feature reduction, multi-kNN, and a backpropagation neural network (BPNN) to employ enhanced human face recognition.

In [23], face detection is conducted with a multi-stage model that consists of an integrated algorithm of the Viola–Jones algorithm, Gabor filters, principal component analysis, and artificial neural networks (ANN). The Carnegie Mellon University dataset was used for training and testing. The authors of [24] use the Viola–Jones algorithm for the detection of face and eyes and the neural network for detection of glasses in order to define a more robust algorithm for face and glasses detection. To perform FER using images, the work in [25] proposes a hybrid convolution recurrent neural network technique. This model is capable of extricating the correspondences within images of faces. This is possible due to the consecutive convolution layers, and then the usage of a recurrent network (RNN), which considers the temporal aspects of the images during classification.

The authors of [26] propose a multi-stream CNN model that uses three manually generated features using LBP for extraction and the Sobel function for the detection of edges. This algorithm was created in order to solve the problem of limited data, in an attempt to improve its execution. In [27], the authors based their proposed model on convolutional neural network in order to anticipate emotion in real-time from a digital image. The parameters of this CNN-based model are reduced, and tests were performed using eight different datasets. In [28], the authors investigate the results of experimenting with CNN variables like kernel size and the number of filters, and the effect they may have on the accuracy of classification. The model was applied using the FER-2013 dataset.

When it comes to detecting drowsiness, there have been many published papers dedicated to the detection of fatigue in drivers of vehicles. Several multinational car establishments are always investigating different techniques for modeling driver inattention. Famous car firms like Toyota, Nissan, and Volkswagen all have their own driver and road safety systems, as investigated by the authors of [29].

The techniques for detecting fatigue differ from automotive companies to third parties. Due to the nature of their business, such car companies will base their models around the specifications and attributes of their vehicle products. Comparatively, others tend to concentrate on the physical characteristics of the drivers themselves, using movements like rapid blinking, the turning of heads, yawning, and so on [29, 30]. However, the path to a highly precise technology remains as yet undiscovered.

There have been several papers proposing various models for recognizing student concentration. However, many of these techniques are not related to images [32–34]. To elaborate a few examples, the authors in [33] use the Structural Similarity Index Method (SSIM) in order to gauge the changes between frames from a live video feed and used the results to detect drowsiness in students during lectures. In [36], the data from the Microsoft Kinect One Sensor was used to assess attention levels of students.

The authors used EEG signals from an EEG headset to evaluate the auditory attention in visually impaired students in [37]. Using heart activity as captures through the use of smart watches, the authors of [38] to measure the attention levels of the students. The analysis of learner attention was conducted using fuzzy logic in [39].

### 3 Challenges Posed to the Design of a Student Concentration Level Recognition System

For situations like online classes where the students are visible from cameras attached to their phones, laptops or desktop computers, a few issues come up due to factors like Internet connectivity, camera quality, position of cameras, environmental features, and so on.

The following are some of the challenges that will come up in the face of creating a face recognition system for student concentration detection keeping all these issues in mind:

- The task of creating an FER system that can adapt to various situations remains an arduous task, particularly in situations or environments which are unpredictable. Significant issues that prevent this kind of flexibility include the obstruction of the face in the image as well as changes in pose and gait. Both of these things effectively hinder the process of face expression identification, seen in [38, 39]. In the case of online learning environments, there are often differences in the quality of webcam feeds along with possible unreliable internet connectivity in a single class of several students. Additionally, the different backgrounds and positioning of each student may offer significant challenges in the feature extraction stage of the FER system.
- FER is an activity highly reliant on available data. The training of deep neural networks demands training data that is great both in terms of quality as well as quantity, as there is a need to capture more minute details in the changes of the facial muscles related to expressions. As it happens, there is a real scarcity of such training data, and any existing data remains lacking in consistency and numbers [42, 43]. This is more so the case for an attention level detection system, as most datasets have generic labels like joy, sadness, neutral, and so on for their training images.
- The burden of processing high-volume data is enormous. The amount of processing power required for an extensively trained model significant enough that other options should be looked into for reducing the pressure on researchers, like compression algorithms. For data compression, either the minimization or discarding of irrelevant noise or other information from the input data would be worthwhile.
- While individual FER can achieve promising performance according to visible facial images, it may be possible to provide additional information and further enhance reliability when coupled with other models into an integrated system.

- Huang et al. in [2] suggested that the different factors that can help correctly classify emotions, like paralinguistic information [44], light, and depth information from 3D data, and usage of the Valence-Arousal model [45] may be combined to produce a system that is more effective than the individual models.
- All things considered, quite a wide scope of proposed FER techniques depend on image data of good quality but have next to zero stress on securing their clients' visual privacy. In order to have both privacy while maintaining useful data for FER frameworks, there is a need for better security measures [46–48].

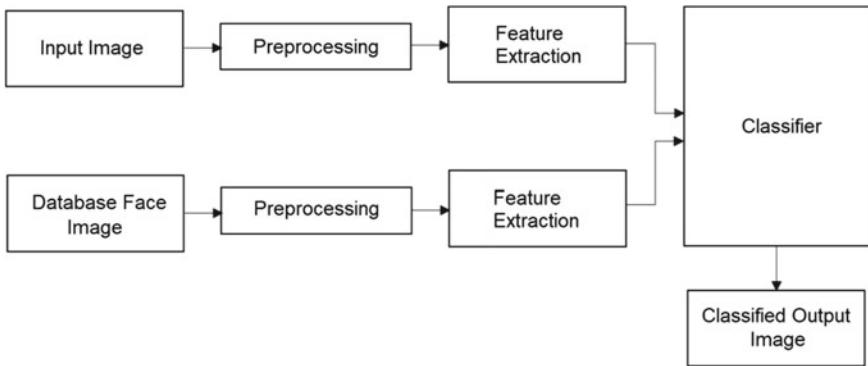
It is essential to choose the correct techniques for creating a system as computationally intensive as an expression recognition system. Some suggestions for doing the same in the application of student concentration evaluation include using Action Units [49] and Facial Action Coding System (FACS) [48] to incorporate the movements of the facial muscles as well as head pose variations into the system; using a combination of existing datasets and custom datasets for training, both to identify and discard situations involving generic expressions as well as to include manually labeled drowsiness, fatigue, or distracted expressions; performing classification using powerful classifiers such as SVM or CNN; and using good preprocessing techniques including noise reduction filters like Gaussian, Bilateral or Median filters, face detection [51, 52], and image dataset normalization [53–56].

## 4 Suggested Framework to Detect Drowsiness in Students

With the way the world has changed with the global pandemic, a huge part of the public life of humans has shifted online, such as academia, business, and workforce. In particular, the entirety of student life has started to be conducted through online classes. These circumstances make it difficult for teachers to judge the engagement of students in their sessions, as classroom learning provides a level of connection and interaction that cannot be replaced.

In order to improve the experience of online learning, teachers may find it helpful to have some way of being automatically alerted when a student appears to stop paying attention during sessions. Machine learning techniques applied to fields like object detection and face expression recognition (FER), in conjunction with hand-crafted features like histogram of oriented gradients (HOG), and local binary pattern (LBP), has seen great success in the past. Keeping in mind this problem of dwindling student concentration due to lack of human supervision during online classes, a Human Concentration Level Recognition model is proposed.

In Sect. 1, the major steps involved in a standard FER system were listed. They include: detection of the face from an image or video frame, extricating the main features of the face, and finally classifying the expression from the features. The general frame of this FER system is shown in Fig. 1. In the case of an FER system to detect student drowsiness during online classes, we have the existence of a webcam, which we can use as the input.



**Fig. 1** Framework of a general FER system

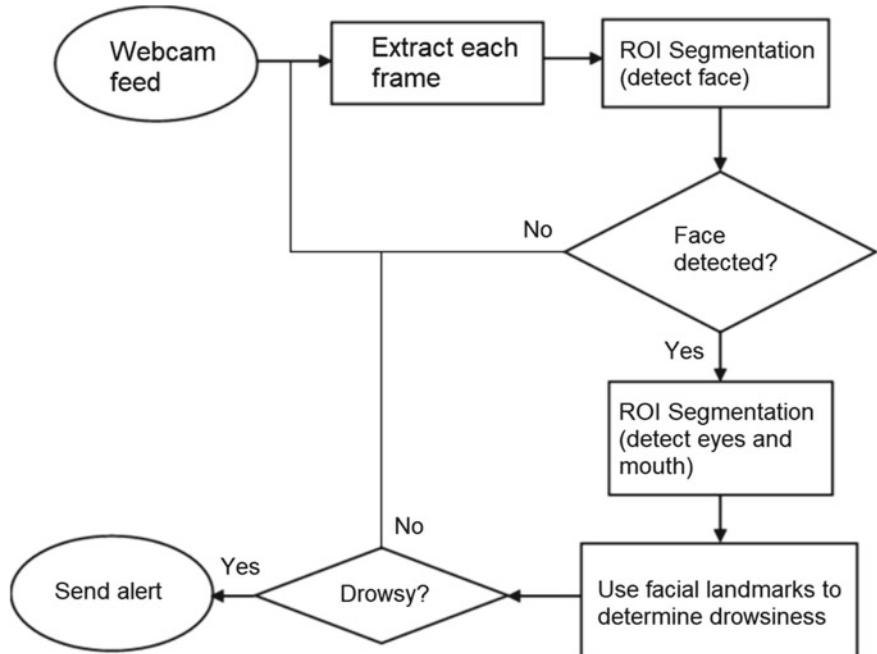
In the case of the proposed system, each of the steps from the standard FER system will be wrapped up in the frame-by-frame captures of the webcam feed. The framework can be divided as the following modules:

- Preprocessing—ROI Segmentation with Viola–Jones algorithm. The Viola–Jones algorithm is a popular, fast and straightforward algorithm which is easy to implement. Its drawback is that it is most effective on frontal images. However, since most students will be facing a webcam during online classes, this should not be a problem in case of the proposed system.
- Feature extraction—Application of Gabor filters, along with feature reduction using principal component analysis. Gabor filters are a powerful tool in texture analysis, cited by many as having a similar effect as would be perceived by humans [54]. PCA can be used to extract only the most relevant features to improve calculation time.
- Classification—Application of a machine learning classifier along with calculations using facial landmarks. There are several existing powerful classifiers which would be effective for categorizing the expression, including SVM and CNN.

The diagram of the proposed framework is given in Fig. 2.

## 5 Conclusion

This paper presents a survey of existing models for face expression recognition, driver drowsiness and fatigue, and student concentration level recognition to which kinds of techniques may best be suited to create a model to detect student drowsiness and distractedness during classes conducted online via video calls. The various challenges and limitations to this application were presented. A suggested system for detecting student drowsiness is also given.



**Fig. 2** Framework of proposed system

## References

1. A. Mehrabian, J.A. Russell, *An Approach to Environmental Psychology* (Press, M.I.T, 1974)
2. Y. Huang, F. Chen, S. Lv, X. Wang, Facial expression recognition: a survey. *Symmetry* (Basel) **11** (2019)
3. M.J. Lyons, M. Kamachi, J. Gyoba, Coding facial expressions with gabor wavelets (IVC Special Issue)
4. P. Lucey, J.F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, I. Matthews, The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression, in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, CVPRW 2010* (2010), pp. 94–101
5. N. Aifanti, C. Papachristou, A. Delopoulos, The MUG facial expression database, in *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10* (2010), pp. 1–4
6. A.M. Martinez, The AR face database. CVC Tech. Report24 (1998)
7. MMI Facial Expression Database—Home <https://mmifacedb.eu/>
8. L.-F. Chen, Y.-S. Yen, *Taiwanese Facial Expression Image Database* (Brain Mapp. Lab. Inst. Brain Sci. Natl. Yang-Ming Univ, Taipei, Taiwan, 2007)
9. K. Kadir, M.K. Kamaruddin, H. Nasir, S.I. Safie, Z.A.K. Bakti, A comparative study between LBP and Haar-like features for Face Detection using OpenCV. In: *2014 4th International Conference on Engineering Technology and Technopreneurship, ICE2T 2014. 2014-Augus* (2015), pp. 335–339
10. X. Huang, S.J. Wang, X. Liu, G. Zhao, X. Feng, M. Pietikainen, Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition. *IEEE Trans. Affect. Comput.* **10**, 32–47 (2019)

11. T. Ahonen, A. Hadid, M. Pietikainen, Face description with local binary patterns: application to face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **28**, 2037–2041 (2006)
12. S. Kokila, B. Yogameena, Face recognition based person specific identification for video surveillance applications, in *Proceedings of the Third International Symposium on Women in Computing and Informatics - WCI '15* (ACM Press, New York, New York, USA, 2015), pp. 143–148
13. A.N. Ekweariri, K. Yurtkan, Facial expression recognition using enhanced local binary patterns, in *Proceedings - 9th International Conference Computer Intelligence Communication Networks, CICN 2017. 2018-Janua* (2018), pp. 43–47
14. M. Kawulok, J. Szymanek, Precise multi-level face detector for advanced analysis of facial images. *IET Image Process.* **6**, 95–103 (2012)
15. A.D. Egorov, D.U. Divitskii, A.A. Dolgih, G.A. Mazurenko, Some cases of optimization face detection methods on image (Using the Viola-Jones method as an example), in *Proceedings of the 2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering, ElConRus 2018* (Institute of Electrical and Electronics Engineers Inc. 2018), pp. 1075–1078
16. L.B. Krithika, G.G. Lakshmi Priya, Student emotion recognition system (SERS) for e-learning Improvement based on learner concentration metric. *Procedia Comput. Sci.* **85**, 767–776 (2016)
17. T.Y. Chai, B.M. Goi, Y.H. Tay, Y.H. Khoo, Vote-based Iris detection system. *ACM Int. Conf. Proc. Ser. Part F1479*, 114–118 (2019)
18. Y. Wang, B.A. Muthu, C.B. Sivaparthipan, Internet of things driven physical activity recognition system for physical education. *Microproc. Microsyst.* **81**, 103723 (2021)
19. D. Do, T. Anh Le, T. N. Nguyen, X. Li, K.M. Rabie, Joint impacts of imperfect CSI and imperfect SIC in cognitive radio-assisted NOMA-V2X Communications. *IEEE Access* **8**, 128629–128645 (2020). <https://doi.org/10.1109/ACCESS.2020.3008788>
20. M.Z. Khan, S. Harous, S.U. Hassan, M.U. Ghani Khan, R. Iqbal, S. Mumtaz, Deep unified model for face recognition based on convolution neural network and edge computing. *IEEE Access* **7**, 72622–72633 (2019)
21. G. Cao, Y. Ma, X. Meng, Y. Gao, M. Meng, Emotion recognition based on CNN, in *Chinese Control Conference, CCC.IEEE Computer Society* (2019), pp. 8627–8630
22. M.A. Abuzneid, A. Mahmood, Enhanced human face recognition using LBPH descriptor, multi-KNN, and back-propagation neural network. *IEEE Access* **6**, 20641–20651 (2018)
23. M. Da'San, A. Alqudah, O. Debeir, Face detection using Viola and Jones method and neural networks, in *2015 International Conference on Information and Communication Technology Research ICTRC 2015* (2015), pp. 40–43
24. M.N. Chaudhari, M. Deshmukh, G. Ramrakhiani, R. Parvatikar, Face detection using viola jones algorithm and neural networks, in *Proceedings - 2018 4th International Conference on Computing, Communication Control and Automation, ICCUBEA 2018* (Institute of Electrical and Electronics Engineers Inc. 2018)
25. N. Jain, S. Kumar, A. Kumar, P. Shamsolmoali, M. Zareapoor, Hybrid deep neural networks for face emotion recognition. *Pattern Recognit. Lett.* **115**, 101–106 (2018)
26. J.A. Aghamaleki, V. Ashkani Chenarlogh, Multi-stream CNN for facial expression recognition in limited training data. *Multimed. Tools Appl.* **78**, 22861–22882 (2019)
27. S. Jaiswal, G.C. Nandi, Robust real-time emotion detection system using CNN architecture. *Neural Comput. Appl.* **32**, 11253–11262 (2020)
28. A. Agrawal, N. Mittal, Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy. *Vis. Comput.* **36**, 405–412 (2020)
29. G. Sikander, S. Anwar, Driver fatigue detection systems: a review. *IEEE Trans. Intell. Transp. Syst.* **20**, 2339–2352 (2019)
30. L. Bretzner, M. Krantz, Towards low-cost systems for measuring visual cues of driver fatigue and inattention in automotive applications, in *IEEE International Conference on Vehicular Electronics and Safety* (IEEE , 2005), pp. 161–164
31. C. Morimoto, D. Koons, A. Amir, M. Flickner, Pupil detection and tracking using multiple light sources. *Image Vis. Comput.* **18**, 331–335 (2000)

32. V. Rothoft, J. Si, F. Jiang, R. Shen, Monitor pupils' attention by image super resolution and anomaly detection, in *2017 International Conference on Computer Systems, Electronics and Control ICCSEC 2017* (2018), pp. 843–847
33. W. Li, F. Jiang, R. Shen, Sleep gesture detection in classroom monitor system, in *ICASSP 2019–2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE 2019), pp. 7640–7644
34. S.M. Yang, C.M. Chen, C.M. Yu, Assessing the attention levels of students by using a novel attention aware system based on brainwave signals, in *Proceedings of - 2015 IIAI 4th International Congress on Advanced Applied Informatics, IIAI-AAI 2015* pp. 379–384 (2016)
35. N. Krishnan, S. Ahmed, T. Ganta, G. Jeyakumar, A video analytics based solution for detecting the attention level of the students in class rooms, in *Proceeding of Confluence 2020: 10th International Conference on Cloud Computing, Data Science and Engineering* (2020), pp. 498–501
36. J. Zaletelj, A. Košir, Predicting students' attention in the classroom from Kinect facial and body features. *Eurasip J. Image Video Process.* **2017** (2017)
37. H. Ghasemy, M. Momtazpour, S.H. Sardouie, Detection of sustained auditory attention in students with visual impairment, in *27th Iranian Conference on Electrical Engineering* (2019), pp. 1798–1801
38. Z. Zhu, S. Ober, R. Jafari, Modeling and detecting student attention and interest level using wearable computers, in *14th International Conference on Wearable and Implantable Body Sensor Networks*, BSN 2017 (2017), pp. 13–18
39. K.A. Hwang, C.H. Yang, Attentiveness assessment in learning based on fuzzy logic analysis, in *Proceeding of 8th International Conference on Intelligent Systems design and Applications, ISDA 2008*. vol. 3, 142–146 (2008)
40. Y. Li, J. Zeng, S. Shan, X. Chen, Occlusion aware facial expression recognition using CNN With attention mechanism. *IEEE Trans. Image Process.* **28**, 2439–2450 (2019)
41. Lai, Y.H., Lai, S.H.: Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition, in *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018* (Institute of Electrical and Electronics Engineers Inc. 2018), pp. 263–270
42. S. Li, W. Deng, Deep emotion transfer network for cross-database facial expression recognition, in *Proceedings—International Conference on Pattern Recognition* (Institute of Electrical and Electronics Engineers Inc. 2018), pp. 3092–3099
43. S. Li, W. Deng, A deeper look at facial expression dataset bias. *IEEE Trans. Affect. Comput.* (2020)
44. S. Ramakrishnan, I.M.M. El Emary, Speech emotion recognition approaches in human computer interaction. *Telecommun. Syst.* **52**, 1467–1478 (2013)
45. W.Y. Chang, S.H. Hsu, J.H. Chien, FATAUVA-Net: an integrated deep learning framework for facial attribute recognition, action unit detection, and valence-arousal estimation. *IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Work.* 2017-July, 1963–1971 (2017)
46. J. Chen, J. Konrad, P. Ishwar, VGAN-based image representation learning for privacy-preserving facial expression recognition, in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops* (IEEE Computer Society, 2018), pp. 1651–1660
47. Y. Rahulamathavan, M. Rajarajan, Efficient privacy-preserving facial expression classification. *IEEE Trans. Dependable Secur. Comput.* **14**, 326–338 (2017)
48. E.M. Newton, L. Sweeney, B. Malin, Preserving privacy by de-identifying face images. *IEEE Trans. Knowl. Data Eng.* **17**, 232–243 (2005)
49. Y.L. Tian, T. Kanade, J.F. Cohn, Recognizing upper face action units for facial expression analysis. *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* **1**, 294–301 (2000)
50. P. Ekman, W. Friesen, *Facial Action Coding System: Investigator's Guide Consulting* (Psychologists Press, Palo Alto, CA, USA, 1978)
51. P. Viola, M.J. Jones, Robust real-time face detection. *Int. J. Comput. Vis.* **57**, 137–154 (2004)
52. R.L. Hsu, M. Abdel-Mottaleb, A.K. Jain, Face detection in color images. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**, 696–706 (2002)

53. S. Du, R. Ward, Wavelet-based illumination normalization for face recognition, in *Proceedings—International Conference on Image Processing, ICIP*. (2005), pp. 954–957
54. W. Chen, M.J. Er, S. Wu, Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain. *IEEE Trans. Syst. Man, Cybern. Part B Cybern.* **36**, 458–466 (2006)
55. S. Shan, W. Gao, B. Cao, D. Zhao, Illumination normalization for robust face recognition against varying lighting conditions. Presented at the April 23 (2004)
56. B.A. Olshausen, D.J. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607–609 (1996)

# Malware Detection Using API Function Calls



Bashar Hayani and E. Poovammal

**Abstract** The battle between cybersecurity specialists and malware developers is endless. Malware developers always come with new ways to evade anti-virus software by developing more sophisticated malware, and since signature-based methods that are used by anti-virus software cannot identify sophisticated malware quickly and effectively, the researchers recently have started to employ artificial intelligence techniques that prove their effectiveness in detecting malware faster than traditional methods. Two common ways are used in detecting malware using AI techniques, static analysis and dynamic analysis. In this work, we present a system to detect malware using artificial intelligence techniques based on statistical analysis of portable executable files. The system depends on API function calls extracted from portable executable file that form inputs to the system to detect whether the file is malicious or not. We applied SVM, KNN, Decision Tree, Random Forest, and Stacking Ensemble algorithms on the dataset and the system was able to classify the data with high performance rates up to 98% for some algorithms.

**Keywords** Malware detection · Machine learning · Static analysis · API function calls

## 1 Introduction

The Internet nowadays is connected with every domain around the world. A lot of data including sensitive data are transferred through the Internet, moreover, a lot of sensitive data are stored and available online. In addition, each device needs to be connected to the Internet which makes all devices exposed to get infected with malware.

---

B. Hayani (✉) · E. Poovammal

Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur 603203, India

E. Poovammal

e-mail: [poovamme@srmist.edu.in](mailto:poovamme@srmist.edu.in)

Malware is the general name for a number of malicious software types such as backdoors, ransomware and worms. Malware is a programming code written by cyber attackers for purposes of causing harmful damages to data and systems or to gain unauthorized access to a system or network.

In the early 1970s, the documentation of viruses started to emerge. In 1971, Creeper Worm was developed by Bob Thomas which is considered the first documented virus in history. Creeper would move between computers through ARPANET network and display the message, “I’m the creeper, catch me if you can!” Thomas’s intention was not to harm other computers. He wanted to confirm his experiment whether it is possible to transfer a program from one computer to another one through network, and he achieved that purpose. Creeper was totally a benign program [1, 2]. Then in 1974, another virus emerged which is called “Wabbit”. Wabbit is a program that aims to crash the system installed on it ultimately. It works by replicating itself causing choke in the system and making all system resources to be utilized by Wabbit to crash the system.

However, the term “virus”, that we know today, was introduced for the first time in the mid-eighties by Fred Cohen who defined the computer virus for the first time in his Ph.D. thesis. From these simple starts, a massive industry has emerged and since then it has been evolved dramatically. To resist malware, cybersecurity analysts have to improve the defenses against malware and come up with new ways to increase the protection of computer systems [3].

Signature-based malware detection is the most popular method used by anti-virus software. It depends on generating signatures for malware and storing those signatures in a database. Signature is a series of bytes that identifies each malware uniquely. This method works nicely for previously discovered malware. The drawback of this method is that new malware cannot be detected on user device until it has been discovered by the anti-virus company, a signature for that newly discovered malware has been uploaded to the signature database by the anti-virus company and finally the user has updated his anti-virus software to get that signature [4]. Nowadays, cyber attackers have started using automated malware generation tools which depend on different techniques such as encryption, obfuscation and packing techniques. By using these tools, the previously written malware can be altered by cyber attackers to generate newly written malware which has new signatures that might not be identified by anti-viruses.

Heuristic Based Detection anti-virus usually simulates the suspicious file on a virtual safe environment to detect whether the file is malicious or not by observing the behavior of the program during its run time.

Whenever a new development is discovered by cybersecurity specialists toward detecting malware, cyber attackers come with new methods to evade detection systems [5]. So, the battle between cyber attackers and cybersecurity specialists are always going on, and there is always a pressing need to develop alternative analysis and detection methods to deal with the limitations of signature-based detection systems.

In the company of persistent growth in number of malicious programs, traditional malware detection systems gradually have shown its limitations. Signature-based

malware detection fails to detect new viruses until their signatures are available in the signature database, also signature-based malware can be evaded by using techniques such as encryption, obfuscation and packing. Heuristic Based Detection fails to detect new viruses. In addition, different detection rules need to be established for different malware. As a result, a lot of organizations and anti-virus vendors have started recently to leverage artificial intelligence techniques toward detecting developing robust malware detection systems.

Both traditional and AI-based malware detection systems require the suspicious file to be analyzed. The purpose of analyzing the suspicious file is to know its characteristics and behaviour [6]. Dynamic analysis and static analysis are the types of malware analysis. Dynamic analysis is a process of examining malware by running it and observing its behavior, and this process is always done on a secure controlled environment to keep the system safe from any potential risks. Contrarily, static malware analysis is a process of inspecting the given malware without actually executing the binary code of malware.

Both types of malware analysis have their merits and demerits, and they supplement each other. Compared with dynamic analysis, static malware analysis is considered to be faster, but it is ineffective against sophisticated malware which is successfully obfuscated using obfuscation techniques and that could cause malware to avoid detection. Contrarily, obfuscated and sophisticated malware hardly evades dynamic analysis because it observes and examines the program during its run time.

## 2 Related Work

In this section, we take a quick look on some researches that have been developed to detect malware depending on artificial intelligence methods by using static analysis features extracted from the portable executable.

A method for detecting whether a file is malicious or benign using n-gram attribute similarity was proposed by Fuyong and Tiezhu [7]. The first step they did is to extract all n-grams from code for each file in the training dataset, then information gain was computed for each n-gram of the training samples. After that, the maximum T n-gram that got the highest information gain was chosen to form features for the proposed method, then the average for each attribute in malware and benign samples was computed independently. The decision whether the test sample is benign or malicious detected based on attributes similarity between the test sample and the average attributes calculated for the benign and malicious samples in the training dataset. The best accuracy rate that they achieved is 81.07%.

In [8], they proposed a framework to detect whether the Portable Executable is malicious or not based on API calls. The first step is to analyze the Portable Executable to extract the list of imported API function calls, then they apply Clospan algorithm to reduce the binary feature vector. The resulted vector which contains the subset of features is used to train the proposed system using Random forest algorithm. They have achieved detection rate (recall) of 99.7% with an accuracy rate 98.3%.

In [9], a method to detect malware was developed based on entropy calculation. The system has two stages. The first stage is the segmentation stage. In this stage, entropy calculations with wavelet analysis are used to divide the file into segments in order to detect the notable changes in entropy that happen in the areas of the file. In the second stage, the similarity between two given files is calculated by using Levenshtein distance which depends on the segments of the files to calculate the similarity. Therefore, an unknown file will be classified into a class that matches the class of the file in the training dataset that has the highest similarity with that unknown file. The area under curve that they achieve is 93%.

Deep learning techniques are also used in static detection of malware, in [10], they proposed a system to detect malicious portable executable by using convolutional deep neural network. The input for the convolutional neural network is raw sequences of bytes of the files and labels only. The accuracy of the proposed system is around 96%.

Yuxin and Zhu [11] introduced a system to detect malware using Deep Belief Network which was used as a feature extractor for extracting the features from opcode sequences and decreasing the number of features in the input feature vectors. The final hidden layer of Deep Belief Network gives the feature vector of the n-gram vector that forms input to the classification layer of DBN. The classification accuracy of that model was better compared with Decision Tree, KNN and SVM algorithms. The precision that they achieved is 98.6%.

### 3 Data Collection

The dataset that we used to train the applied algorithms collected from [12]. It contains static analysis data of top 1000 API function calls imported in the portable executables extracted from the “pe\_imports” elements of Cuckoo Sandbox reports. The malware samples were downloaded from virusshare.com. Whereas the portable executables for goodware samples were downloaded from portableapps.com and from Windows 7 x 86 directories.

The dataset contains 47,580 samples. Among these samples, there are 1929 goodware and 45,651 malwares.

There are 1002 features in the dataset. The first one is a string of 32 bytes represents the MD5 hash of the sample. The next 1000 features correspond to the top 1000 API function calls imported in portable executable. Each portable executable in the dataset is represented as a binary vector, namely  $X$ , where  $X_i = 1$  if the portable executable file has imported  $i$ th API function call and  $X_i = 0$  if the corresponding portable executable file has not imported  $i$ th API function call where  $i$  goes from 1 to 1000. The output feature has two values 0 if the file is goodware and 1 if the file is malicious.

## 4 Experiments

We have split the dataset into training and testing sets. The training set forms 75% of the dataset and the testing set forms 25% of the dataset. Since the training set has 1482 and 34,203 malware, the training set is imbalanced. The minority class is goodware whereas the majority class is malware. In order to make it balanced, we use Over-Sampling method with replacement for that purpose. After applying the Over-Sampling method, we increased the number of goodware samples from 1482 goodware samples into 35,000 samples. So, the balanced training set after applying Over-Sampling method contains 34,203 malware and 35,000 goodware.

Four experiments have been done:

- Experiment 1:** Classifying data without feature selection and extraction.
- Experiment 2:** Classifying data with feature selection using Chi-squared.
- Experiment 3:** Classifying data with feature selection using Mutual Information.
- Experiment 4:** Classifying data with feature extraction using PCA algorithm.

The methods that have been used in the four experiments are:

- **Method 1:** KNN Algorithm with  $K = 3$ .
- **Method 2:** SVM Algorithm.
- **Method 3:** Random Forest Algorithm.
- **Method 4:** Decision Tree Algorithm.
- **Method 5:** Stacking classifier with KNN, Decision Tree, SVM and Random Forest as base classifiers and Logistic Regression as a meta classifier
- **Method 6:** Stacking classifier with KNN, Decision Tree, SVM and Random Forest as base classifiers and Backpropagation neural network as a meta classifier.
- **Method 7:** Stacking classifier with KNN, Decision Tree, SVM and Random Forest as base classifiers and Random Forest algorithm as a meta classifier.

### 4.1 Experiment 1: Classifying Data Without Feature Selection and Extraction

Table 1 shows the classification results for this experiment. As we can see Methods 3, 4, 5 and 6 give better results compared with Methods 1 and 2.

**Table 1** Classification results for experiment 1

Method #	Precision	Recall	F1-score	Accuracy
1	0.97	0.97	0.97	0.97
2	0.98	0.96	0.96	0.96
3	0.98	0.97	0.97	0.97
4	0.97	0.97	0.97	0.97
5	0.98	0.97	0.97	0.97
6	0.98	0.97	0.97	0.97

**Table 2** Classification results for experiment 2

Method #	Precision	Recall	F1-score	Accuracy
1	0.97	0.97	0.97	0.97
2	0.97	0.95	0.96	0.95
3	0.97	0.96	0.97	0.96
4	0.97	0.96	0.96	0.96
5	0.98	0.97	0.97	0.97
6	0.97	0.97	0.97	0.97

#### **4.2 Experiment 2: Classifying Data with Feature Selection Using Chi-Squared**

In this experiment, we applied Chi-squared method to select the features that are more dependent of the target variable (class attribute) and remove features which are less dependent of the class attribute from the dataset. After applying Chi-squared on our dataset, the 227 features which are strongly more dependent of the class attribute have been selected to make the dataset that contains 227 features. Table 2 shows the classification results for this experiment.

#### **4.3 Experiment 3: Classifying Data with Feature Selection Using Mutual Information**

Mutual information is computed between two variables to quantify the mutual dependence between those two variables. Mutual information is always greater than or equal to zero. The greater value of mutual information, the stronger relationship is between those two variables. When mutual information is equal to zero that means the two variables are independent. In this experiment, we applied Mutual Information to measure the dependency between each feature and the class attribute, then the 227 features that are strongly more dependent of the class attribute have been selected to make the dataset that contains 227 features. Table 3 shows the classification results for this experiment.

#### **4.4 Experiment 4: Classifying Data with Feature Extraction Using PCA Algorithm**

PCA algorithm computes the eigenvectors and eigenvalues of the covariance matrix. After that, the N highest eigenvalues with their corresponding eigenvectors are used to put the data in a new space that has N dimensions which is lower than or equal

**Table 3** Classification results for experiment 3

Method #	Precision	Recall	F1-score	Accuracy
1	0.97	0.98	0.97	0.98
2	0.97	0.95	0.96	0.95
3	0.97	0.96	0.97	0.96
4	0.97	0.96	0.96	0.96
5	0.98	0.96	0.97	0.96
6	0.98	0.97	0.97	0.97
7	0.96	0.95	0.95	0.95

**Table 4** Classification results for experiment 4

Method #	Precision	Recall	F1-score	Accuracy
1	0.98	0.97	0.97	0.97
2	0.97	0.95	0.96	0.95
3	0.97	0.97	0.97	0.97
4	0.97	0.96	0.96	0.96
5	0.97	0.97	0.97	0.97
6	0.97	0.97	0.97	0.97

to the original number of dimensions of data. After applying PCA algorithm on our dataset, we selected the 120 components that have the highest variances to form the new dataset that has 120 features, then we applied the classification methods on the new dataset. Table 4 shows the classification results for this experiment.

## 4.5 Results Discussion

The tables show the results that we got after training and testing all methods in the four experiments on the dataset. As we can see from the results, the applied algorithms perform well on balanced dataset. The recall is the measure of our model correctly identifying True Positives. It is an important metric when evaluating a malware detection model because when it is high that means the number of False Negatives is low which indicates a high malware detection rate.

The precision is the ratio between the True Positives and all the Positives. A high precision implies a low number of False Positive. The Accuracy represents the ratio of the total number of data points that have been classified correctly and the total number of data points. F1-score is the Harmonic mean of the Precision and Recall.

We can notice from the experiments that the experiments achieve 98% Precision, Recall, F1-score and Accuracy in some applied methods. However, the experiment 1 uses the dataset without any feature reduction or selection which means it uses 1000

features as inputs to the applied methods and the complexity of those models are too high and led to a long training time compared with the other experiments. In contrary, the experiment 4 uses PCA algorithm on the original dataset to extract features. As a result, the new dataset after applying PCA algorithm has only 120 features. Those 120 features formed the inputs to the applied methods and we got almost the same performance for some applied methods compared with the experiment 1 and a little bit better performance than the experiment 2 and experiment 3 for some applied methods. The advantage of feature selection and reduction that has been done in the experiments 2,3 and 4 is that complexities of the applied methods are lower than the applied method in the experiment 1 and the training time for the applied methods in the experiments 2,3 and 4 is less compared with the experiment 1.

The experiments 2 and 3 achieve upto 98% Precision, Recall, and Accuracy in some applied methods. In those experiments, we have chosen the most 227 dependent features on the target variable. The degradation in the performance in those experiments is only upto 1% compared with experiments 1, but the complexity and training time for the applied methods in the experiments 2 and 3 are significantly less than the experiment 1.

## 5 Conclusion

In this work, we presented a malware detection system based on API function calls extracted from portable executable. These API function calls form inputs to the proposed system. To reduce the complexity and training time of the applied methods, we leveraged some of feature selection and extraction techniques and we were able to acquire high detection rate for all applied experiment and applied methods.

## References

1. T.M. Chen, J.M. Robert, The evolution of viruses and worms (2004). <https://doi.org/10.1201/9781420030884.ch16>
2. O. Olowoyo, P. Owolawi, Malware classification using deep learning technique, in *2nd International Multidisciplinary Information Technology and Engineering Conference (IMITEC)*, (2020), pp. 1–6. <https://doi.org/10.1109/IMITEC50163.2020.9334071>
3. F. Cohen, Computer viruses: theory and experiments. *Comput. Security* **6**, 22–35 (1987)
4. Ö. A. Aslan, R. Samet, A comprehensive review on malware detection approaches. *IEEE Access* **8**, 6249–6271 (2020). <https://doi.org/10.1109/ACCESS.2019.2963724>
5. D. Gibert, C. Mateu, J. Planes, The rise of machine learning for detection and classification of malware: Research developments, trends and challenges, *J. Netw. Comput. Appl.* **153**, 102526 (2020). ISSN 1084-8045. <https://doi.org/10.1016/j.jnca.2019.102526>
6. M. Sikorski, A. Honig, Practical malware analysis: the hands-on guide to dissecting malicious software. no starch press, (2012)
7. Z. Fuyong, Z. Tiezhu, Malware detection and classification based on N-Grams attribute similarity. in *2017 IEEE International Conference on Computational Science and Engineering*

- (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), Guangzhou (2017), pp. 793–796. <https://doi.org/10.1109/CSE-EUC.2017.157>
- 8. S. Ashkan, B. Yadegari, H. Rahimi, N. Peiravian, S. Hashemi, A. Hamze, Malware detection based on mining API calls. in *Proceedings of the 2010 ACM Symposium on Applied Computing*. (Association for Computing Machinery, 2010), pp. 1020–1025
  - 9. B. Donabelle, L. Richard, S. Mark, Structural entropy and metamorphic malware. *J. Comput. Virol. Hacking Techniq.* **9**(2013). <https://doi.org/10.1007/s11416-013-0185-4>
  - 10. M. Krc̄el, O. Švec, M. Bælek, O. Jařek, Deep convolutional malware classifiers can learn from raw executables and labels only (2018). [Online]. Available <https://openreview.net/forum?id=HkHrmM1PM>
  - 11. D. Yuxin, S. Zhu, Malware detection based on deep learning algorithm. *Neural Comput. Appl.* **31** (2019). <https://doi.org/10.1007/s00521-017-3077-6>
  - 12. A. Oliveira, Malware analysis datasets: top-1000 PE imports. *IEEE Dataport November* 7, (2019). <https://doi.org/10.21227/004e-v304>

# Implementing Blockchain Technology for Health-Related Early Response Service in Emergency Situations



Nemanja Zdravković, Milena Bogdanović, Miroslav Trajanović, and Vijayakumar Ponnusamy

**Abstract** Throughout late 2019 and in 2020, we have witnessed the COVID19 epidemic, where the self-isolation of newly and potential infected people helped stop the virus spreading. The detection of potential patients in the early disease phase has proven crucial, as did monitoring patients with early symptoms. For any country, it was challenging to monitor people in self-isolation, and also to detect, predict, and observe changes in their health in the early disease phase. The lack of control and untimely access to such information increased the number of patients and victims. For the purpose of monitoring already confirmed infected cases, countries had to open and maintain temporary hospitals. In this paper, we present a model on how losses in financial, logistic, equipment, and human resources could be avoided by applying blockchain technology. On the basis of the immutability and security properties of blockchain, data loss and patients' malfeasance would be impossible.

**Keywords** Blockchain · Data protection · Diagnostic devices · Healthcare · IoT · Reliability · Sensors

---

N. Zdravković (✉) · M. Bogdanović

Faculty of Information Technology, Belgrade Metropolitan University, Belgrade, Serbia  
e-mail: [nemanja.zdravkovic@metropolitan.ac.rs](mailto:nemanja.zdravkovic@metropolitan.ac.rs)

M. Bogdanović

e-mail: [milena.bogdanovic@metropolitan.ac.rs](mailto:milena.bogdanovic@metropolitan.ac.rs)

M. Trajanović

Faculty of Mechanical Engineering, University of Niš, Niš, Serbia  
e-mail: [traja@masfak.ni.ac.rs](mailto:traja@masfak.ni.ac.rs)

V. Ponnusamy

Department of ECE, SRM Institute of Science and Technology, Kattankulathur, India  
e-mail: [vijayakp@srmist.edu.in](mailto:vijayakp@srmist.edu.in)

## 1 Introduction

The coronavirus (COVID-19) outbreak which started in the last quarter of 2019 has caused a global health emergency [1]. In mere months, the number of new people infected by the coronavirus worldwide rose to more than one million infected. In addition, the rapid spread of the virus lead to new cases being reported globally, on a daily basis. As a result, many countries around the world enforced lock-downs and social distancing guidelines to stop the spread of COVID-19, promoting self-isolation as a mean to stop (or at least decrease) the spread of the virus. In addition, the process of contact tracing has been applied, where all people whom the patient (or in some cases a potential patient) has come in contact with are identified.

To aid with tracking the self-isolation of patients, tracking apps on smart wearable devices and medical (Internet of Things) IoT became a promising solution to in the prevention of virus transmission, and at the same time maintaining data quality and data integrity. Furthermore, to track the progress of the pandemic, tracking valid data became vital. All stakeholders, such as healthcare and government officials, researchers, as well as tech companies, started using mobile apps and devices with contact-tracing options, often using Bluetooth- or NFC- based proximity tracing, and even geolocation tracking options to track COVID- 19 cases [2, 3]. Good data is required to understand the pandemic dynamics in order to predict disease spreading rate, the effectiveness of countermeasures, and the overall impact it has on the peoples' lives. This availability of the data by itself is not sufficient; the data acquired online is in general not secure as it is susceptible to data manipulation [4].

Here Blockchain technology (BCT) emerged as a key technology that will transform the way in which healthcare-related information is shared [5–8]. Block-chain, paired with medical IoT and tracking apps, has the potential to reduce the spread of the disease, while keeping the data valid and immutable.

The remainder of the paper is presented as follows. Section 2 gives related work in Blockchain-aided healthcare solutions regarding emergency situations such as the ongoing pandemic. In Sect. 3, we identify the similarities and differences in the proposed solutions, as well as possible bottlenecks that frequently occur. On the basis of previous models, we propose our system model for an early response system based on BCT. Finally, in Sect. 4 we draw conclusions and set up a foundation for our research.

## 2 Related Work

### 2.1 A Quick BCT Overview

Blockchain imposes fundamental changes to the way personal data, especially medical data, are currently being processed, and can improve current data security solutions. A Blockchain is a shared, append-only distributed ledger, in which all

transactions (describing some events) are stored in linked blocks [9]. Every transaction, part from the data, contains a unique cryptographic signature, ensuring the ledger resilient to modifications. In addition, this ledger is hence simultaneously shared across all network member nodes, resulting in real-time node update. A block can be viewed as a data structure consisting of a set of transactions, and a header which connects the new block to the previous one. All blocks hence form a chain, and can trace back to the first block, called the genesis block. A blockchain relies on peer-to-peer networks, public-key cryptography, and distributed consensus. The combination of these three concepts is what secures blockchain transactions. Unlike a centralized system, no single entity should be able to control the process of adding a block to the chain: all member nodes share equal rights, and every single block is at all time managed by all member nodes. This management system is accomplished with distributed consensus. This process establishes an agreement among the nodes in the blockchain network in the validation of each data block to be added to the chain. Depending on the consensus algorithm, nodes can either compete for correct transaction validation, be chosen randomly, or apply a different algorithm altogether.

It is important to note that Blockchains are a class of technology; the term refers to different forms of distributed databases with variations in their technical and governance arrangements and complexity. One significant advantage of using BCT is that it can reform interoperability of healthcare databases, providing authorized access to patient medical records, and to other hospital assets [10]. Blockchain as a class of technology is showing enough opportunities to become an integral part of fighting against COVID-19 as it could enable: (a) efficient tracking and monitoring solutions, (b) transparent supply chain of vital products and donations, and (c) secure payments [4].

## 2.2 Related Work

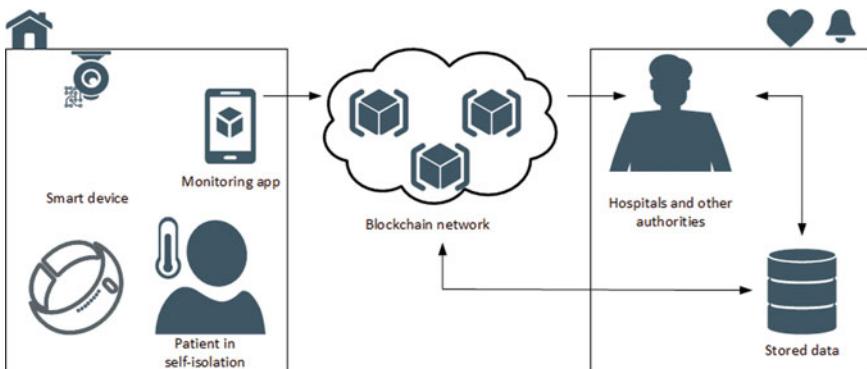
In the last five years, various Blockchain-aided healthcare solutions have been proposed [11–19]. The majority of them are based on executing transactions in the form of adding new data, or permitting/revoking access to the patient's electronic health record (EHR), while a smaller number of papers have gained attention in the last couple of years by using a model that includes healthcare IoT devices, mostly in Remote Patient Monitoring (RPM). As of writing this paper, there are very few papers that tackle the application of BCT to the COVID-19 pandemic. The authors of [4] proposed and evaluated a blockchain-based tracking system for validating the COVID-19 data from diverse sources to mitigate the spread of falsified or modified data, based on Ethereum smart contracts. The authors of [20] propose a blockchain-based system for issuing immunity certificates in order to mitigate the falsification of test reports and encourage people to come in contact with individuals having immunity-based licenses. In [21, 22], the general challenges that have arisen during the COVID-19 pandemic are highlighted, and the authors argue the applicability of blockchain as a key enabling technology to tackle those challenges, identifying

potential use cases and a high-level view of how blockchain can be leveraged and discuss the expected performance.

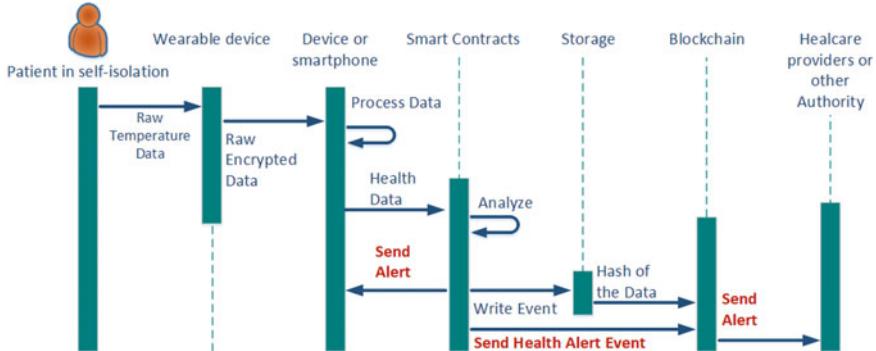
### 3 Proposed System Model

Building upon existing Blockchain-with-IoT solutions presented in literature, we consider a model which can be grouped into three parts, as shown in Fig. 1. The first part is at the patient's home, or place of staying during self-isolation. The patient would be equipped with medical IoT devices, such as smart wristbands or similar wearable devices, which have location tracking features and, if possible, for measuring body temperature. If such wearable devices are not available, the place of stay during self-isolation should be equipped with RPM devices with temperature monitoring features. These devices have an interface to the blockchain network (directly, or through a smartphone with an appropriate app), and a possibility to execute transactions.

The app should be setup in such a way that it periodically sends a transaction to the blockchain network, and to send a transaction when an alarm condition is met, which we termed a health alert event. A health alert event may include when the sensing device measures high body temperature or when a patient has moved beyond his self-isolation location. These transactions, along with patient ID, patient location, and any other relevant data, including a timestamp are then validated through the blockchain network. Sensitive data should be anonymized, with authority institutions containing the mapping to the real data. The blockchain network uses a fast consensus mechanism that requires low processing power [20]. After a transaction has been validated and has entered the blockchain, it can be easily tracked back. All transactions can be accessed by the authorities (such as hospitals or similar authorities), and can further store data on their own databases, e.g. health information system



**Fig. 1** Blockchain-based early response system for patients in self-isolation



**Fig. 2** Sequence diagram for a health alert event

databases. In addition, these databases are connected to the blockchain as well. If a direct modification to the database is attempted, it will trigger a modification transaction to the blockchain. Transactions which include health alert events should be flagged as special transactions, and can trigger additional activities from the authority stakeholder. A sequence diagram of the health alert event can be shown in Fig. 2.

The sequence starts with the patient, interfacing with an authorized device within the network. It is assumed that the patient has previously authorized healthcare providers access to their data. Data from the patient in raw form is sent to the device, which is then encrypted using a lightweight digital signature algorithm. Encrypted raw data is hence sent to a device acting as a node for formatting and processing. The formatted data is sent to the appropriate smart contract and only when analyzed, the node will write an event (transaction) to the Blockchain. The data itself can be stored externally, while the hash of the data is written (when verified) on the Blockchain(s). A health alert will reach the healthcare provider and the patient as well, notifying that a health alert event has been triggered, after which an appropriate authority can take needed actions.

## 4 Conclusion

The described early response system pointed out the possibilities of easily implementing blockchain technology in emergencies. It can help expand the blockchain research field due to its immutability properties, identifying other use-cases as well. With appropriate interfaces, this system may be integrated with multiple healthcare systems, or outpatient treatment applications, for patients suffering from various diseases. Secure remote medical care can therefore be easily obtainable. Using BCT, sensitive data can be managed and stored securely and reliably. Hence, blockchain paired with IoT has the potential to be used in many application areas within healthcare. It can also be used for elderly person care, or with people with disabilities or

special needs. We have shown that this system can be applicable and easily deployable to any situation that requires direct and indirect observation. It can be made scalable to handle a wide range of populace.

The goal of this paper was to identify the potential of using Blockchain paired with RPM and IoT in emergency situations which require monitoring a large populace, and yet is still able to acquire data that is tamper-proof and transparent to those with authorized access. This paper therefore presents the first step in our research in BCT-aided healthcare solutions, and is a foundation for future development.

## References

1. World Health Organization (WHO): Rolling updates on coronavirus disease (covid-19). <https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen>. [Online], Accessed July (2020)
2. P. Bischoff, Covid-19 app tracker: Is privacy being sacrificed in a bid to combat the virus? <https://www.comparitech.com/blog/vpn-privacy/coronavirus-apps/>. [Online] Accessed Oct (2020)
3. K. Warner, S. Nowais, Coronavirus: doctors urge public to help track COVID-19 cases with tracing app. <https://www.thenationalnews.com/uae/health/coronavirus-doctors-urge-public-to-help-track-covid-19-cases-with-tracing-app-1.1012267>. [Online] Accessed Oct (2020)
4. D. Marbouh, T. Abbasi, F. Maasmi, I.A. Omar, M.S. Debe, K. Salah, R. Jayaraman, S. Ellahham, Blockchain for COVID-19: review, opportunities, and a trusted tracking system. Arabian J. Sci. Eng. 1–17 (2020)
5. P.P. Ray, D. Dash, K. Salah, N. Kumar, Blockchain for COVID-19: blockchain for IoT-based healthcare: background, consensus, platforms, and use cases. IEEE Syst. J. (2020)
6. F.A. Khan, M. Asif, A. Ahmad, M. Alharbi, H. Aljuaid, Blockchain technology, improvement suggestions, security challenges on smart grid and its application in healthcare for sustainable development. Sustain. Cities Soc. **55**, 102018 (2020)
7. N. Zdravkovic, M. Damnjanovic, D. Domazet, V. Ponnusamy, M. Trajanovic, Implementing blockchain technology for data protection in health-IoT and diagnostic devices: overview, potential, and issues. in *Proceedings of the 10th International Conference on Information Society and Technology* (ICIST 2020) (Mar 2020), pp. 211–215
8. V. Grkovic, J. Jovic, N. Zdravkovic, M. Trajanovic, D. Domazet, V. Ponnusamy, Usage of blockchain technology for sensitive data protection—medical records use case. in *Proceedings of the 10th International Conference on Information Society and Technology* (ICIST 2020) (Mar 2020), pp. 216–221
9. Z. Zheng, S. Xie, H. Dai, X. Chen, H. Wang, An overview of blockchain technology: architecture, consensus, and future trends. in *2017 IEEE International Congress on Big Data* (BigData congress) (IEEE, 2017), pp. 557–564
10. K.N. Griggs, O. Ossipova, C.P. Kohlios, A.N. Baccarini, E.A. Howson, T. Hayajneh, Healthcare blockchain system using smart contracts for secure automated remote patient monitoring. J. Med. Syst. **42**(7), 130 (2018)
11. G. Zyskind, O. Nathan, Decentralizing privacy: using blockchain to protect personal data. in *2015 IEEE Security and Privacy Workshops* (IEEE, 2015), pp. 180–184
12. M. Mettler, Blockchain technology in healthcare: the revolution starts here. in *2016 IEEE 18th international conference on e-health networking, applications and services* (Healthcom) (IEEE, 2016), pp. 1–3
13. C. Esposito, A. De Santis, G. Tortora, H. Chang, K.K.R. Choo, Blockchain: a panacea for healthcare cloud-based data security and privacy? IEEE Cloud Computing **5**(1), 31–37 (2018)

14. A. Azaria, A. Ekblaw, T. Vieira, A. Lippman, Medrec: Using blockchain for medical data access and permission management. in *2016 2nd International Conference on Open and Big Data (OBD)* (IEEE, 2016), pp. 25–30
15. W.J. Gordon, C. Catalini, Blockchain technology for healthcare: facilitating the transition to patient-driven interoperability. *Comput. Struct. Biotechnol. J.* **16**, 224–230 (2018)
16. G. Rajmohan, C.V. Chinnappan, A.D. William, S.C. Balakrishnan, B.A. Muthu, G. Manogaran, Revamping land coverage analysis using aerial satellite image mapping. *Trans. Emerg. Telecommun. Technol.* (2020). <https://doi.org/10.1002/ett.3927>
17. D.V. Pham, G.L. Nguyen, T.N. Nguyen, C.V. Pham, A.V. Nguyen, Multi-topic misinformation blocking with budget constraint on online social networks. *IEEE Access* **8**, 78879–78889 (2020). <https://doi.org/10.1109/ACCESS.2020.2989140>
18. Y. Rahulamathavan, R.C.W. Phan, M. Rajarajan, S. Misra, A. Kondoz, Privacy-preserving blockchain based IoT ecosystem using attribute-based encryption. in *2017 IEEE International Conference on Advanced Networks and Telecommunications Systems (ANTS)* (IEEE, 2017), pp. 1–6
19. A.D. Dwivedi, G. Srivastava, S. Dhar, R. Singh, A decentralized privacy-preserving healthcare blockchain for IoT. *Sensors* **19**(2), 326 (2019)
20. A. Bansal, C. Garg, R.P. Padappayil, Optimizing the implementation of COVID-19 “immunity certificates” using blockchain. *J. Med. Syst.* **44**(9), 1–2 (2020)
21. A. Kalla, T. Hewa, R.A. Mishra, M. Ylianttila, M. Liyanage, The role of blockchain to fight against COVID-19. *IEEE Eng. Manage. Rev.* **48**(3), 85–96 (2020)
22. W. Wang, D.T. Hoang, P. Hu, Z. Xiong, D. Niyato, P. Wang, Y. Wen, D.I. Kim, A survey on consensus mechanisms and mining strategy management in blockchain networks. *IEEE Access* **7**, 22328–22370 (2019)

# Interpreting Chest X-rays for COVID-19 Applying AI and Deep Learning: A Technical Review



A. Veronica Nithila Sugirtham and C. Malathy

**Abstract** The outbreak of novel coronavirus disease was raised in China in December 2019 and since then it has escalated worldwide becoming a major health concern on an international level. Nearly, going to be one year since the virus came into this world but still there is neither exact medicine to cure the sick patients nor any vaccine to prevent it. This scenario has given researchers a lot of opportunity to explore the virus, its detection, and treatment. Vaccine trials are being carried out in different countries around the world. The present laboratory tests are time-consuming and a lack of availability of kits delay the diagnosis. The respiratory system is the part of the human body stricken the most by the pathogen and thus use of X-ray images of the chest appears to be a more methodical way than the ongoing thermal imaging. Motivated by the modern advancements of artificial intelligence and machine learning and their application in various areas many researchers have emphasized its contribution in responding to the pandemic. In this survey paper, several research works have been compared and work done and limitations or future work of each has been described. After reviewing the papers, we have proposed to develop a hybrid model using convolutional neural network (CNN) and Bayesian neural network (BNN) that uses X-ray images for the detection of coronavirus accurately, considering the uncertainty in the predictions and further classify X-rays as of corona, viral pneumonia, or normal case.

**Keywords** COVID-19 · SARS-CoV-2 · Chest radiographs · Computerized tomography (CT) · Convolutional and Bayesian neural network

---

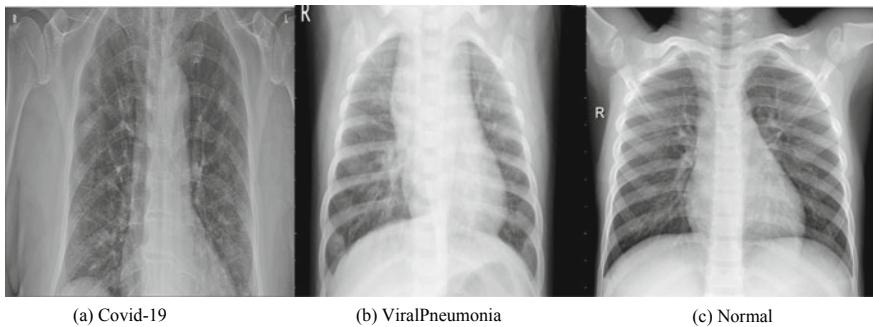
A. V. N. Sugirtham · C. Malathy (✉)

Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur 603203, Tamil Nadu, India

A. V. N. Sugirtham  
e-mail: [vn9428@srmist.edu.in](mailto:vn9428@srmist.edu.in)

## 1 Introduction

Coronavirus disease has turned the world upside down not only affecting the health of human beings but also various other dimensions of life such as transportation, education, travel, supply chain, and politics. Caused by RNA virus named as SARS-CoV-2. Human-to-human transmission via droplets, contaminated hands, or surfaces has been observed with an incubation period of 2–14 days. From China, now it has traveled to more than 188 countries. In India, the first positive case was reported on January 30th, 2020. India currently has the largest number of positive cases in Asia and stands second all over the world after the USA and overtaking Brazil for having the maximum number of confirmed cases. The World Health Organization (WHO) has given few guidelines on how we can protect ourselves from the novel coronavirus and prevent further spread of the disease such as using face masks, hand gloves, and PPE Kit. Maintaining cleanliness. Regularly washing hands with soap and water, in case not available using alcohol-based sanitizer. Social distancing is another major factor which can help in controlling the spread of the virus as it is said the infection is spread through nasal droplets from an infected person, so the person standing the closest would catch the infection quicker than the one standing at least 6-feet apart. Cleaning and disinfecting routinely are some other measures which can help. The most affected part of the human body is the respiratory system leading to cough, fever and major symptoms like shortness of breath and low oxygen saturation which becomes the reason of fatality in most cases. The major cause of fatalities is known to be preexisting health conditions such as blood pressure, diabetes, cardiac problem, kidney, and lung issues. When diagnosing the novel coronavirus disease, the important information to keep note of is the patterns of anomalous obscurities in the lungs, as well as the spread of murk across the entire lung. Applying AI to the analysis can help automate a process, which normally requires doctors to physically check hundreds to thousands of chest X-ray or CT images, and can help to reduce the load on doctors and support speedy decision making for effective and timely treatment of the patients, thus reducing the number of deaths. Many people are asymptomatic but carry the virus and pass it to others and later are diagnosed positive. Nowadays, cases can also be seen where the patient complains of shortness of breath but has no fever and his/her samples are negative for the virus, in such cases early detection of the virus using this AI method based on X-ray images of the lungs can help doctors start the treatment on time and save their lives. High body temperature, respiratory symptoms such as shortness of breath, decreased count of white blood cells, pneumonia, low oxygen level in blood are the classic clinical attributes of coronavirus cases. For screening suspected cases, the standard method is considered to be the reverse-transcription polymerase chain reaction testing. However, sensitivity of this method screening is not up to the mark. Also, physically visiting the laboratory or hospital to give the sample maybe exhausting for patients as they are advised to take rest and be in home quarantine and can also be the reason for others exposure to the virus. As transmission of virus is through the droplets released when an infected person coughs or sneezes and these droplets present in the air can enter the lungs of a



**Fig. 1** Sample chest X-ray images of various patients stored in Kaggle COVID-19 radiography database

healthy person when he inhales the same air. Though people are wearing masks still it is believed that the virus may stay on the top surface of the mask and can ultimately infect the person. In comparison to conventional imaging techniques that heavily stresses on human efforts, more secure, precise, and effective imaging answers are provided by artificial intelligence. When suspected cases are diagnosed accurately as early as possible, they play a vital role in timely isolation and start of medication, which is incredibly important for patients' prospect, the regulation of this pandemic, and therefore the overall health safety of the public (Fig. 1).

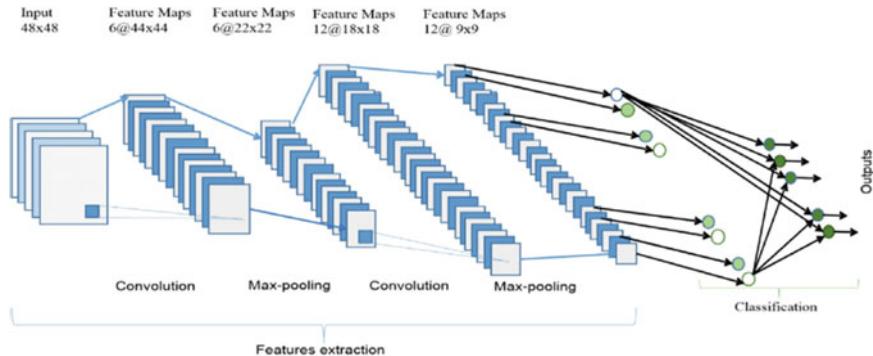
## 2 Literature Survey

Several AI models have been put forward for classifying COVID-19 using chest X-ray images as well as CT images. Proposed that the impacted areas of lungs can be located by radiologists using heatmaps. Developed a model called DarkCovidNet model to assist clinicians in order to make quicker and precise diagnostic decisions [1]. ResNet-34 and ResNet -50 models were trained using an approach called transfer learning, and results were reported where it stated that ResNet-50 architecture exhibited a better performance than ResNet-34, with a precision rate of 72.38% [2]. X-ray deviations strongly suggested that patients have caught the infection after analyzing both X-ray and CT images of lungs. The presence of lesions in the lower part of chest, and blur surfaces of opacity (also known as ground glass opacity) and fluid filled inside lungs instead of air also known as consolidations that were circumferential and two-sided in nature were observed [3]. Ground glass opacity, enlargement of blood vessels, and thickening of pleura are typical CT signs monitored in the CTs. Most of the cases show the involvement of both sides and numerous spotting. It was presumed that the combined use of imaging techniques and thermal screening resulted in better diagnosis. Also, positive CT manifestations were shown by 97% of coronavirus patients [4]. A three-dimensional DeCoVNet deep CNN for detecting

coronavirus from volumes of computerized tomography was proposed unnecessarily elucidating the lesions in CT volumes for training, strong classification performance was obtained from weakly supervised deep learning framework and resulted in good lesion localization. The deep learning-based detection model used in this study is effective [5]. To manifest the effectiveness of AI-powered medical vision for coronavirus, two techniques were used X-rays and CT. The whole channel of medical imaging and analysis practices involved with coronavirus, which includes of acquisition of image, segmentation, diagnosis, and follow-up has been covered [6]. Put up an approach based on transfer learning CNN for computerized sensing of coronavirus. Applying lung X-ray images eight famous previously found to be effective CNN-based deep learning algorithms different from one another were prepared and evaluated for the identification of healthy versus patients having pneumonia. CNN models were trained by using image augmentation [7–9]. To learn the representation hierarchy of attributes from the chest X-ray dataset and segregate them into normal, COVID, and pneumonia categories the VGG and Inception-V3 models are optimal [10]. Model aimed at pruning of dataset, model integration, and classification of lung images. According to the precision and error value, architectures chosen were Res Net 101 and Res Net 152 with fine fusion effect, and ratio of their weights was dynamically improved through the training process. During evaluation, the model could achieve high accuracy on the test set around ninety-six percentage [11]. Proposed an intelligent healthcare structure B5G enabled to combat infectious pandemics such as COVID-19. The structure is composed of a cloud layer, an edge layer, and a shareholder layer. Social distancing among people, wearing of facial masks, and high temperature of body sensing were incorporated into the structure for mass surveillance. Utilized the modern generation of eminent computing systems and analyzed hospital test dataset and signs of human vitals. The proposed method could be applied to any communicable disease. It would therefore contribute to decrease crowding in hospitals and to identify coronavirus negative patients [12]. Proposed a deep forest guided by feature or attribute selection adaptive for coronavirus versus community-acquired pneumonia (CAP) classification by making use of the lung CT images. Specifically, the sophisticated depiction built on the surface-peculiar attributes were learned using deep learning. Also, to decrease the repetition of attributes an adaptive attribute selection was employed ground on the forest trained [13]. Proposed a model called Corona-19 Net molded out of MOBILE-NET V2 architecture using transfer learning approach [14] (Fig. 2 and Table 1).

### 3 Result and Discussion

Summarizing the above-referred papers, it is found out that as limited amount of X-ray images of chest of coronavirus positive patients were present, so they used a learning approach called transfer learning where in pre-trained model is taken and the model is tuned finely with own dataset which is known as transfer learning approach. This is based on the notion that the pre-trained model will be present as a feature



**Fig. 2** Convolutional neural network

extractor, and only train the last layer on the current task. Some of the pre-trained models used for image classification were:

1. Very Deep Convolutional Networks for Large-Scale Image Recognition (VGG-16)
2. Inceptionv3
3. ResNet50
4. MobileNetV2
5. DenseNet

Also, the existing models are implemented in the local system and not placed in the cloud. Thus, in this paper the proposed model is a hybrid of CNN and BNN and will be developed from scratch without applying any transfer learning approach and will try to achieve the maximum accuracy possible and correctly classify COVID-19 vs. pneumonia vs. normal X-rays.

## 4 Dataset

The dataset intended to be used in our proposed work is an open-source coronavirus radiography database available in Kaggle consisting of 1143 X-ray images of chest of corona positive cases, 1341 images of healthy lungs, and 1345 chest radiographs of viral pneumonia cases collected by a team of researchers of Qatar University, Doha (Fig. 3).

## 5 Conclusion

This paper elaborates on how various researchers have found AI and machine learning to be playing a major role in detecting COVID-19 and thus helping in combating it.

**Table 1** Summary of the literature survey comparing the existing works and their limitations

Title	Year and Publisher	Work done	Limitations / Future work
Automated detection of COVID-19 cases using deep neural networks with X-ray images [1]	2020 Elsevier Computers in biology and medicine Volume 121	Proposed that the impacted areas of lungs can be located in the X-ray images of the chest by radiologists using heatmaps. Developed a model called DarkCovidNet model to assist clinicians in order to make quicker and precise diagnostic decisions	Limitation of this study was a lack of public image data. In the future, intended to incorporate more images to validate the model. This AI model could also be extended on cloud storage providing immediate diagnosis and guidance to affected patient's rehabilitation
Computer vision and radiology for COVID-19 detection [2]	2020 international conference for emerging technology (INCET) Publisher: IEEE	Using transfer learning approach taught ResNet-34 and ResNet-50 and results reported where it stated that ResNet-50 architecture exhibited a better performance than ResNet-34, with a precision rate of 72.38%	The lack of data quality is one of the downsides of this paper. The currently available dataset was too limited to obtain exemplary performance and therefore replace the thermal imaging technique
AI-Driven COVID-19 tools to interpret, quantify lung images [3]	2020 IEEE Pulse (Volume: 11, Issue: 4)	X-ray deviations strongly suggested that patients have caught the infection after analyzing both X-ray and CT images of lungs. The presence of lesions in the lower part of chest, and blur surfaces of opacity (also known as ground glass opacity) and fluid filled inside lungs instead of air also known as consolidations that were circumferential and two-sided in nature were observed	When Qure.ai (founded in 2016) began this work a larger dataset of lung X-rays for training qXR (the name of the model) was not available

(continued)

**Table 1** (continued)

Title	Year and Publisher	Work done	Limitations / Future work
The role of imaging in the detection and management of COVID-19: a review [4]	2020 IEEE reviews in biomedical engineering	Ground glass opacity, enlargement of blood vessels, and thickening of pleura are typical CT signs monitored in the CTs. It was presumed that the combined use of imaging techniques and thermal screening resulted in better diagnosis	To ascertain the current findings studies from multiple centers with large size of sample images are still required
A Weakly supervised Framework for COVID-19 classification and lesion localization from chest CT [5]	Aug 2020 IEEE transactions on medical imaging (Volume: 39, Issue: 8)	DeCoVNet, a deep CNN for detecting coronavirus from volumes of Computerized Tomography was proposed unnecessarily elucidating the lesions in CT volumes for training, strong classification performance was obtained from weakly supervised deep learning framework and resulted in good lesion localization	The data used was collected from a single hospital and did not validate across centers. Community-acquired pneumonia (CAP) CT images data was not gathered during experiment
Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation, and Diagnosis for COVID-19 [6]	April 2020 IEEE reviews in biomedical engineering (Early Access)	To manifest the effectiveness of AI-powered medical vision for coronavirus, two techniques were used X-rays and CT. The whole channel of medical imaging and analysis practices involved with coronavirus, which includes of acquisition of image, segmentation, diagnosis, and follow-up has been covered	The dataset of images in these models can have labels which are neither complete, exact nor accurate, proving to be a hurdle for training a precise model. Encouraging the scrutinizing of deep learning with self-supervision and transfer learning approach is the fact that manually labeling imaging data is costly and consumes lot of time

(continued)

**Table 1** (continued)

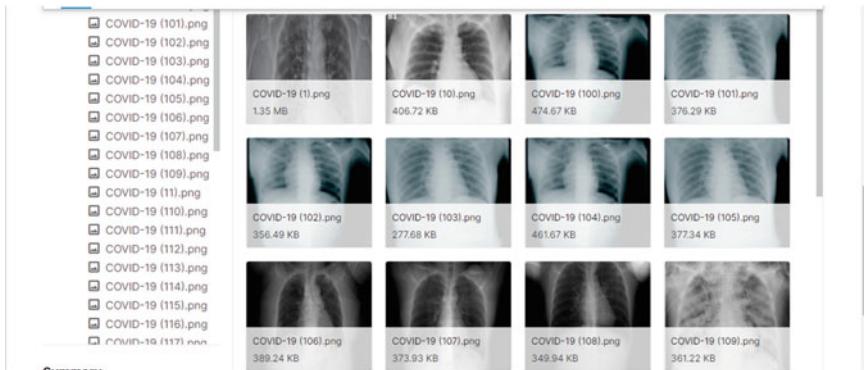
Title	Year and Publisher	Work done	Limitations / Future work
Can AI help in screening Viral and COVID-19 pneumonia? [7]	2020 IEEE Access (Volume 8)	Using lung X-ray images eight famous previously found to be effective CNN-based deep learning algorithms different from one another were prepared and evaluated for identification of healthy versus patients having pneumonia	The outcomes of this study cannot be incorporated into real-time applications until the performance reported on smaller database in this study is evaluated on a larger database of X-rays
Iteratively pruned deep learning ensembles for COVID-19 detection in chest X-rays [8]	2020 IEEE Access (Volume 8)	To learn the representation hierarchy of attributes from the X-ray images and segregate them into normal, COVID, and pneumonia categories the VGG & Inception-V3 models are optimal	In a deep model redundant network, parameters (neurons) are present which do not yield to improve the prediction performance. A quicker and compact model having equivalent or better performance than the models which are not pruned will be a resultant of identifying and removing the neurons with lesser activations
Deep learning for the detection of COVID-19 using transfer learning and model integration [9]	2020 IEEE 10th international conference on electronic information and emergency communications (ICEIEC) Publisher: IEEE	Model aimed at pruning of dataset, model integration, and classification of lung images. According to the precision and error value, architectures chosen were Res Net 101 and ResNet 152 with fine fusion effect and ratio of their weights were dynamically improved through the training process. During evaluation, the model could achieve high accuracy on the test set around ninety-six percentage	Throughout the preparation, the error value was not stable because of the variable deviations of weights in the two models. The percentage of accuracy of evaluating a model was high in the previous round, but not mandatory for the next round, and the improvement of weight in the previous round, resulted in slow convergence of loss

(continued)

**Table 1** (continued)

Title	Year and Publisher	Work done	Limitations / Future work
Explainable AI and mass surveillance system-based healthcare framework to combat COVID-19 like pandemics [10]	2020 IEEE Network (Volume: 34, Issue: 4)	Proposed an intelligent health care structure B5G enabled to combat infectious pandemics such as COVID-19. Social distancing among people, wearing of facial masks, and high temperature of body sensing was incorporated into the structure for mass surveillance	In the future, deep learning algorithms and proteinase pattern study would be evaluated as part of the overall structure. Additionally, edge programming could be used for the decreased delay and enhanced safety
Adaptive feature selection guided deep forest for COVID-19 classification with chest CT [11]	2020 IEEE Journal of Biomedical and Health Informatics (Early Access)	Proposed a deep forest guided by feature or attribute selection adaptive for coronavirus versus community-acquired pneumonia (CAP) classification by making use of the lung CT images. Specifically, the sophisticated depiction built on the surface-peculiar attributes were learned using deep learning	In the coming years, more data with multiple diseases are planned to be consolidated. Also, expecting to enhance the presented method for hike in performance deep learning algorithms will be used to learn the features
CORONA-19 NET: transfer learning approach for automatic classification of Coronavirus infections in chest radiographs [12]	ICIPCN 2020. Advances in intelligent systems and computing, vol 1200. Springer,	Proposed a model called Corona-19 Net molded out of MOBILE-NET V2 architecture using transfer learning approach	To excel and prove as an universal technique for the prior detection of the virus further feedback and larger dataset are required

Several models based on transfer learning of pre-trained image classification models have been discussed in this paper. Though existing models are achieving results, researchers still have great opportunity to explore more in this area and come up with models resolving the limitations of the existing models. Despite achieving results in almost all classification tasks, neural networks still make overconfident decisions. Missing in existing neural network architectures is a measure of uncertainty in the prediction. To make the model vulnerable to over-fitting issues, incredibly careful training and regularization of weights are needed. To address both these concerns, our



**Fig. 3** A sample of COVID positive chest X-ray images

proposed model introduces Bayesian learning to convolutional neural network adding a measure of uncertainty and regularization in their predictions. A hybrid model of convolutional neural network and Bayesian neural network will be developed using the Google Colab platform. The output from one will be fed into the other. Will be experimenting both ways either from CNN to BNN or from BNN to CNN whichever gives the best result will be taken as the final model to achieve the maximum accuracy. Since this proposed work is being undertaken during the peak time of corona, a large database of chest radiographs of corona positive cases, viral pneumonia besides images of normal lungs are available to train our model effectively and test the model accurately.

## References

1. T. Ozturk, M. Talo, E.A. Yildirim, U.B. Baloglu, O. Yildirim, U. Rajendra Acharya, Automated detection of COVID-19 cases using deep neural networks with X-ray images. *Comput. Biol. Med.* **121**, (Elsevier, 2020)
2. R. Punia, L. Kumar, M. Mujahid, R. Rohilla, Computer vision and radiology for COVID-19 detection. in *2020 International Conference for Emerging Technology (INSET)*, Belgaum, India (2020), pp. 1–5. <https://doi.org/10.1109/INSET49848.2020.9154088>
3. L. Mertz, AI-Driven COVID-19 tools to interpret, quantify lung images. *IEEE Pulse* **11**(4), 2–7 (2020). <https://doi.org/10.1109/MPULS.2020.3008354>
4. D. Dong et al., The role of imaging in the detection and management of COVID-19: a review. *IEEE Rev. Biomed. Eng.* (2021). <https://doi.org/10.1109/RBME.2020.2990959>
5. X. Wang et al., A weakly-supervised framework for COVID-19 classification and lesion localization from chest CT. *IEEE Trans. Med. Imaging* **39**(8), 2615–2625 (2020). <https://doi.org/10.1109/TMI.2020.2995965>
6. F. Shi et al., Review of artificial intelligence techniques in imaging data acquisition, segmentation and diagnosis for COVID-19. *IEEE Rev. Biomed. Eng.* (2021). <https://doi.org/10.1109/RBME.2020.2987975>
7. S.-I. Chu, B.-H. Liu, N.-T. Nguyen, Secure AF relaying with efficient partial relay selection scheme. *Int. J. Commun. Syst.* **32**, e4105 (2019). <https://doi.org/10.1002/dac.4105>

8. G.R. Nitta, T. Sravani, S. Nitta, B. Muthu, Dominant gray-level based K-means algorithm for MRI images. *Heal. Technol.* **10**(1), 281–287 (2019). <https://doi.org/10.1007/s12553-018-00293-1>
9. M.E.H. Chowdhury et al., Can AI help in screening viral and COVID-19 Pneumonia? *IEEE Access* **8**, 132665–132676 (2020). <https://doi.org/10.1109/ACCESS.2020.3010287>
10. S. Rajaraman, J. Siegelman, P.O. Alderson, L.S. Folio, L.R. Folio, S.K. Antani, Iteratively pruned deep learning ensembles for COVID-19 detection in chest X-rays. *IEEE Access* **8**, 115041–115050 (2020). <https://doi.org/10.1109/ACCESS.2020.3003810>
11. N. Wang, H. Liu, C. Xu, Deep learning for the detection of COVID-19 using transfer learning and model integration. in *2020 IEEE 10th International Conference on Electronics Information and Emergency Communication (ICEIEC)*, Beijing, China, (2020), pp. 281–284. <https://doi.org/10.1109/ICEIEC49280.2020.9152329>
12. M.S. Hossain, G. Muhammad, N. Guizani, Explainable AI and mass surveillance system-based healthcare framework to combat COVID-19 like pandemics. *IEEE Netw.* **34**(4), 126–132 (2020). <https://doi.org/10.1109/MNET.011.2000458>
13. L. Sun et al., Adaptive feature selection guided deep forest for COVID-19 classification with chest CT. *IEEE J. Biomed. Health Inform.* **24**(10), 2798–2805 (2020). <https://doi.org/10.1109/JBHI.2020.3019505>
14. S. Sharma, S. Ghose, S. Datta, C. Malathy, M. Gayathri, M. Prabhakaran, CORONA-19 NET: transfer learning approach for automatic classification of coronavirus infections in chest radiographs. in *Image Processing and Capsule Networks. ICIPCN 2020. Advances in Intelligent Systems and Computing*, vol. 1200, ed. by J.Z. Chen, J. Tavares S. Shakya, A. Iliyasu (Springer, Cham, 2021). [https://doi.org/10.1007/978-3-030-51859-2\\_48](https://doi.org/10.1007/978-3-030-51859-2_48)

# Technical Survey on Covid Precautions Monitoring System Using Machine Learning



K. Lekha, Nisha Yadav, and T. Y. J. Naga Malleswari

**Abstract** The world hit by the Covid-19 pandemic now has made public safety in community spaces a major task for officials, diverting them from other major activities. To prevent this, the whole surveillance process can be automated with a single good quality camera. Complete surveillance would recognize and detect the following conditions on a person crossing the camera: face mask recognition, safe distance detection, body temperature detection, and blood oxygen concentration. To automate this process of monitoring of people, the computer needs to perform face occlusion detection and the other three conditions; and classify based on the results. Thus, in this paper, we present a comparative study of recent trends on tackling this problem, and in addition, we aim to perform a complete literature survey to look for the most efficient methods to implement checking of all four features together in single surveillance technology.

**Keywords** Computer vision · Machine learning · Neural networks · Support vector machine · Pulse oximetry · Thermal imaging · Template matching · Sparse network of winnow · Naive bayes classifier · Occluded face recognition (ORF)

## 1 Introduction

Computer Vision is defined as “an interdisciplinary field that deals with how computers can be made to gain high-level understanding from digital images or videos”. The advances in the field of artificial intelligence, specifically in deep learning and neural networks have given a boost to computer vision technologies

---

K. Lekha (✉) · N. Yadav · T. Y. J. Naga Malleswari  
Department of CSE, SRMIST, Kattangulathur, India  
e-mail: [lk5160@srmist.edu.in](mailto:lk5160@srmist.edu.in)

N. Yadav  
e-mail: [na2098@srmist.edu.in](mailto:na2098@srmist.edu.in)

T. Y. J. Naga Malleswari  
e-mail: [nagamalt@srmist.edu.in](mailto:nagamalt@srmist.edu.in)

and have been able to outshine even us at tasks such as detecting and labeling objects [1]. Computer vision is dramatically redefining the tech world these days in various fields such as Self-Driving Cars, Augmented Reality and Mixed Reality, and Healthcare. Computer vision systems are widely used for public space surveillance because it can reduce the ineffectiveness and sources of error to a large extent. The sources of error mentioned here are getting distracted during surveillance and becoming bored. A computer algorithm will never succumb to these errors [2].

One of the main applications of computer vision is face detection and recognition. Face detection is a technology that identifies human faces in a given frame or digital image and it can be further extended to perform face recognition by matching it with a specific person's facial features. There are a lot of application areas for face detection and recognition such as smart cards, surveillance systems, information security, biometrics, access control, law enforcement [3]. There are many algorithms that have been defined to do face detection and recognition given a number of image conditions. The major issue that has grabbed a lot of attention in recent years is that of partially occluded faces. The model has to majorly concentrate on non-occluded facial features and ignore the occluded parts [4]. Face masks produce a similar challenge, as it covers a major part. After this pandemic outburst, a lot of implementations are being tried and tested. This paper aims to give a comparative study of the techniques used to tackle previous face occlusion problems and how it can be incorporated for the face mask.

In this paper, we also aim to include a study of computer vision algorithms to monitor other precautionary steps. The first facet is to detect if a safe distance of 6 feet is maintained among individuals. Secondly, calculating body temperature using thermal imaging. Finally, calculating blood oxygen concentration using camera-based pulse oximetry.

The rest of this paper is organized as follows. Section 2 describes in detail face detection and recognition, the technologies widely used followed by technologies used for the same on an occluded face. Section 3 describes the Safe Distance Detection techniques, Section 4 about body-temperature calculation, Section 5 about camera-based pulse oximetry. Section 6 gives the comparison of different techniques used in a tabular form, and the Inference obtained from the comparison. The conclusion is given in Section 7.

## 2 Face Detection and Recognition

Face Detection, a pattern recognition problem by origin, is a technique that identifies the presence of a human face(s) in a given digital image. A major application of face detection is face recognition, but it's just one of the many applications. Face Detection is also applied in auto-focus cameras that focus on the faces of people present in the frame. Face Recognition being a major application, has gained a lot of popularity recently, that it has even replaced signatures and biometrics such as fingerprint and pupil pattern for authentication. Most of the face detection algorithms often begin

by searching for human eyes since they are very prominent and are easy to detect. After detecting eyes, the system goes for other facial features like the mouth, nose, eyebrows, and iris. Once the system gets a hunch that there's a face in the given image in this way, it can then apply additional tests to validate if the detected part is a face or not. Face recognition, a technology that goes ahead of face detection and recognizes whose face is present in the given input digital image from a training dataset and thus it can also be described as biometric technology. As a first step, the face detection algorithm detects, extracts, crops, resizes, and usually converts to grayscale the face detected on the given digital image. The face recognition algorithm works upon this image passed on from the face detection algorithm and finds features that best describe the face and compares it to facial images in the dataset that was used to train it. Though facial recognition is not completely accurate, it can identify with good accuracy when there is a high chance that a person's face matches someone in the database [5].

There are different scenarios posed to this challenge. A normal and desirable scenario would be a controlled environment where the lighting is appropriate. In such scenarios, a simple edge detection technique can do the job. Another challenge is when extracting images from videos. A few of the common features that are used are skin texture and eye-blink.

## 2.1 Techniques

Face Detection methods are divided into four categories by Yan, Kriegman, and Ahuja. Many algorithms fall into multiple categories since the divisions are closely related to each other. The categories are [6]:

**Knowledge-Based Methods.** Knowledge-based methods are rule-based methods where the rules are formed by the researcher through his/her personal knowledge. We all have the knowledge that a human face will have two symmetric eyes, a nose, and a mouth, and we also know that there's an approximate ratio of the distance between these features. Our knowledge on human faces also includes the point that the eyes will be the darkest part of the face since it's a cavity. A knowledge-based method is where we apply the above-mentioned knowledge we have as rules and identify which part of the digital input image follows these rules. The rules are subjective to the researcher doing the implementation and there's no fixed set of rules here. When too many rules are used, or if the rules are too strict, a lot of faces that do not pass all those rules will not be detected. On the other hand, if the rules are too general, a lot of false positives happen. Another issue in this method is that, if the orientation, pose, or angle of the face changes from the forward position, there's very little chance that this method detects the face due to the fixed set of rules and ratio of the distance between the features. In the late 90 s, Han, Liao, Yu, and Chen developed an algorithm that first identifies eye-analogue pixels. Once this is found, the remaining area pixels can be removed from the image. Now they are left with a set of eye-analogue segments, which are checked using a set of rules to determine

if they are actual eyes or not. After getting the eye-analogues that are confirmed to be actual eyes, a surrounding rectangular area is calculated whose four vertices are determined by a set of functions. This rectangle represents a potential face which is then normalized to a fixed size and orientation and is verified using a backpropagation neural network. The final step was the application of a cost function. They obtained a 94% accuracy even in images with multiple faces [6].

**Feature-Invariant Methods.** As we saw earlier, Knowledge-based methods fail when the orientation, pose, or angle of the face is varied. But we can detect a face in all such extreme conditions. Thus, there should be some features or properties to a human face that is invariant to all these conditions and variabilities. Hence, this method tries to detect the individual features of the face first and then infer from that to check whether a face is detected or not by building a statistical model based on the relationship between the individual features detected such as eyebrows, nose, mouth, and eyes. The different elements that are detected individually here are 1. Facial features like eyes, nose, mouth, and eyebrows, 2. Skin color, and 3. Skin texture. The drawback of this method is that any kind of occlusion, noise, or extreme lightings can corrupt the facial features.

**Template Matching Methods.** Here, a template, which is deformable, is formed using the edge contours of a basic face shape. Each face image is looked upon as a function and is compared to the common template formed. The correlation values calculated from the common template on the features in the input image such as eyes, mouth, nose, and face contour are used to infer the existence of the face. Though this method is simple to implement, it fails when the pose, angle, or orientation of the face is varied. To tackle this and achieve the shape and scale invariance, multi-resolution, multi-scale, sub-templates, and deformable templates have been proposed.

**Appearance Based Methods.** Here, the template is learned by the machine based on the given training face images. Thus, this method is the one that majorly used algorithms like machine learning. The model hence formed is used to classify face and non-face images. In the model, the features of face learned by the machine are in the form of distribution models or discriminant functions. The various tools or methods used here are: a. Eigenface-Based: Nowadays, when proposing any new face recognition algorithm, the results are checked against that of eigenfaces approach, since this algorithm, considered by many, is the first-ever algorithm to be implemented for face recognition, and any new algorithm should pass the basic accuracy rate that this eigenfaces approach is providing. This approach represents a face as a coordinate system; b. Neural Networks: Neural networks were used by researchers to perform binary classification to classify face and non-face patterns at the beginning. The challenge here was to represent the digital images that didn't have face patterns. Another approach is to discover a discriminant function to classify patterns using distance measures. Few other methodologies have endeavored to locate a clear boundary between non-face and face images using a constrained generative model; c. Support Vector Machines: By modifying the perception of the output, and formulating a representation of facial images that is resonant with binary classification, SVM is adapted to be used for face recognition, it returns a binary value. In the training process, the problem space finds the dissimilarities between two facial input images.

And a few others include Naive Bayes Classifiers, Sparse Network of Winnows, Information-Theoretical Approach, Hidden Markov Model, Distribution-Based.

## 2.2 Occluded Face Detection and Recognition Techniques

Face detection and recognition for occluded faces was an area not much explored until very recently. It never posed a great challenge or there was no pressing need until recently. But this pandemic situation has created a pressing need to detect and recognize occluded faces since everyone wears a mask these days. Or the technology is at least needed to monitor whether everyone is wearing masks or not. A lot of researchers concentrated on variations in pose, orientation, and illumination for so long. But the current situation has forced researchers to concentrate on another major challenge—occluded face detection and recognition. Since the occlusion can be anywhere and in any size, there can't be a single training dataset or a single model that covers all the variations. Thus occluded face detection and recognition is still a challenging area for researchers. An occluded face recognition(OFR) system is of three units, each plays a crucial role in design decision: cross-occlusion strategy, feature extraction, and comparison strategy. Here, the later two parts are of general face recognition too, while the cross-occlusion strategy is only for OFR. The algorithms will work with any image which is more than  $32 \times 32$  pixels [7].

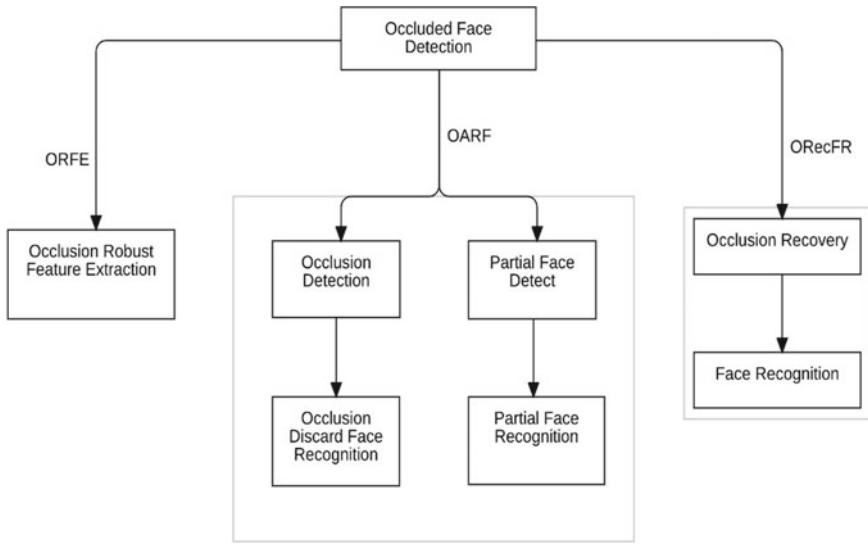
The different types of occlusions are:

1. Real occlusions: sunglasses, eyeglasses, hair, scarves, masks, hat
2. Partial faces: only part of the face in the visible area or a non-frontal pose
3. Synthetic occlusions: any unusual decorative jewels or other items covering parts of the face
4. Occluding rectangle: black or white rectangle occlusions
5. Occluding unrelated images: face occluded with unrelated images such as bottle or non-squared image (Fig. 1).

OFR can be broadly classified into the following three categories [7]: 1. Occlusion Robust Feature Extraction (ORFE) category: Here the algorithms find a feature space that is not affected much by occlusions. Generally, the cross-occlusion strategy uses patch-based, engineered, and learning-based features; 2. Occlusion Aware Face



**Fig.1** Different types of occlusion. (Real occlusions partial faces synthetic occluding occlusions rectangle unrelated images)



**Fig. 2** Three approaches to recognizing faces under occlusions

**Recognition (OAFR) category:** The OAFR category knows the location of the occlusion specifically. Usually, occlusion-discard is implemented as a cross-occlusion strategy. Resultantly, only non-occluded parts of the face qualify for face recognition. Partial face recognition techniques also come under OAFR because they do not consider the occluded parts of the face in the process of recognition, assuming that visible parts are ready in the beginning; 3. Occlusion.

**Recovery based Face Recognition(ORecFR) category:** The ORecFR classification means to recover an occlusion-free face from the given blocked face picture to meet the necessities of a traditional face recognition framework. All in all, it accepts occlusion recovery as the cross-occlusion procedure (Fig. 2).

### 3 Safe Distance Detection Techniques

Social Distancing or Physical Distancing is keeping a safe distance between two individuals in a public space. Social distancing ensures there is no spread of any kind of disease due to closeness or contact. As per Government guidelines, there must be at least 6 feet distance between two individuals [8–10]. This module presents the Safe Distance Detection Techniques that can be used to monitor if every pair of humans are following a safe distance between them. The technique has two major steps: 1. Detection of humans, 2. Detecting Social Distancing Violations.

### 3.1 Detection of Humans

To check if a crowd is maintaining social distance, we must first detect and find the position of people in the frame. This can be done with the help of Computer Vision, with basic object detection techniques by considering features like head, legs, arms, or any other prominent feature [11]. Human Detection faces a number of challenges as the input is taken from an outdoor environment. Humans can vary widely because of the clothes, objects they carry, and sizes. In addition to that, people may be in different poses and sometimes occluded due to objects around them and other people nearby. To deal with the mentioned problems, view-based, and part-based pedestrian-detection approaches are found to be efficient. These approaches naturally handle the partial occlusion since they do not require the whole body of the person to be visible. Moreover, this pedestrian detection is reformed into the detection of various human body parts and views. The different techniques to perform this human (or pedestrian) detection are:

**Histogram of Oriented Gradient Algorithm.** This algorithm basically analyzes the neighboring pixels of every single pixel directly, to compare it with others in terms of darkness. The algorithm uses arrows depending on the direction where image pixels get darker. The process is repeated for every pixel of the image, which later is replaced by arrows, namely gradients. These gradients show the flow of light from light to dark. These gradients are used to obtain histograms in which the angle and magnitude of the gradient are used. Finally, various discriminators are applied to all the histograms obtained, which can define if a person is there in the frame [12].

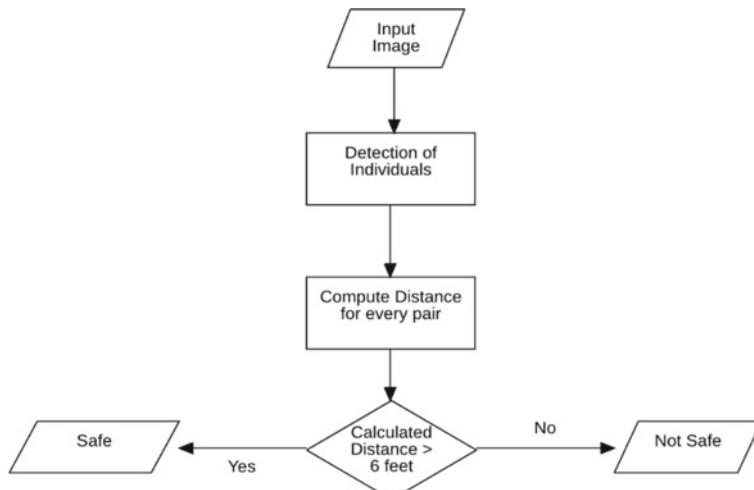
**Deep Sort Model.** Deep sort models use both the appearance of the person and their position to track. The Kalman filter is used to predict likely positions of the next box, to find the position of the individual. The appearance of a person is used in a deep learning model that generates embeddings. In Computer Vision, one of the basic tasks involved in detecting a moving person is to give an ID to the tracked person and assign the same ID again when the person appears in the next frame. No ID is repeated, even if a particular person with a certain ID has left the room. Every time a new person enters, a new ID is allocated. After detecting the person there is a rectangular boundary assigned around him/her. Here, Tracking is a bit difficult, as two persons can be very similar in appearance or features and this may lead to switching IDs between people. Another issue encountered is that many people can be occluded because of a vehicle or any other object. In such a case, the occluded person might be assigned a new ID every time in and out of occlusion. Deep learning techniques have made more efficient multi-object detection possible in the last few years with the current state of the art accuracy of 62.0 [13].

**Classifier-Ensemble-Based Approaches.** In cases where a person is occluded by an object or a vehicle, a part-based search can be done, wherein specific parts of the person being searched and matched together to classify under the human category. The parts taken into consideration can be either the upper half and lower half of the body, or hands, head, or legs separately [14]. Classifier runs over all the parts, and a score is given for each part. If a particular part is occluded, the matching score is

low. Similarly, all the scores are combined to get a final score, which must cross a given threshold score to result in a positive output.

### 3.2 Detecting Social Distancing Violations

We want to calculate the distance between two individuals. We can do this easily if we can get bounding boxes for each individual, and then we can directly calculate the distance between the boxes. But the problem we face in this process is to find the appropriate coordinates to make the box/rectangle. We can overcome this by using the bottom-most coordinate around the person, this will help us eliminate the height factor of the rectangle, which would vary from person to person. Also, we can use this because the distances differ when people are on different planes, and at different distances from the camera. Another way could be using the centroid of the coordinates around the person. A simple Euclidean Distance formula is enough to calculate the distance between the two points found [15]. Further, we need to check for all possibilities, as the other person/coordinate can be in any direction. This can be done by calculating distance among all pairs possible. As a result, the closest people can be found based on the threshold distance we give, for example, 6 feet here. On identifying the distance and comparing the distance between each and every other person, we can check if the distance is less than the threshold value. If so, then two people are too close to each other and violate the social distancing norms. So safe = 0 is denoted in the data frame for both the persons who are found close. This variable safe can be denoted for visualization purposes in the video stream [16] (Fig. 3).



**Fig. 3** Social distancing violation detection process

## 4 Body-Temperature Calculation

Body temperature plays an important role in defining the wellbeing of a person, a number of diseases are directly related to body temperature. Using the body temperature, appropriate findings can help in proper monitoring and deciding cause and treatment for humans. The traditional method includes touching the forehead to guess temperature, which later got replaced with devices like thermometers that took at least 5 min to give the body temperatures. Such methods are unreliable as well as inefficient when we consider a crowd. For immediate, efficient, and on-spot temperature calculation. There are a number of available techniques to determine the temperature of a person using a thermal camera. Presently, Infrared thermography (IRT) is being used for detecting body temperatures of travellers during mass traveling [17]. This is done using thermal cameras that have sensors with a longer wavelength. Infrared cameras detect heat radiated by an element and convert it into electrical signals. These signals can be further used to generate heat pixels and hence deduce temperature around a particular body. Also, FLIR cameras are good at detecting the individual temperature in a crowded place at a fast pace.

### 4.1 Techniques

**Image Processing.** In this method, on obtaining input from IR and FLIR camera input, video is converted to frame, after which a frame selection is done, either manually or automatically. The next step is to do Face Detection using a fast algorithm like Haar. On detection, a cascade detector is created that locates the position of the face. So, the video frame is read and the detector is run over the same. After Edge detection is done, the face is partitioned into two halves. The sum of pixels on the left and right side is found after dividing and the difference between both is found. Finally, the difference is compared with the threshold value, the frames that qualify the threshold value are further processed. Temperature calculation is done using the below-mentioned formulae 1, 2 [18].

$$\text{The temperature of Frame} = \frac{255 * (\text{sum of pixels of video frame})}{\text{No. of pixels in frame}} \quad (1)$$

$$\text{The temperature of Face} = \frac{255 * (\text{sum of pixels of face area})}{\text{No. of pixels in face}} \quad (2)$$

**Thermal Based using ROI Locating.** A thermal Camera is installed that gives input as image streams over time. Here, appropriate and high accuracy face detection is really important. Different Face Detection algorithms like the Image projection method, the Haar-cascade method, ML-based method are used on the input to detect and track the human face in the camera frame. If any face is found,

it gives back the positions of the faces as (3)

$$\text{Face} = \text{Rect}(x, y, w, h) \quad (3)$$

where  $x, y, w, h$  are the coordinates for the rectangle. Even though the detection can even help us track the correct position of the face, its processing time remains too long. To tackle this, the Kernel Correlation Filter tracker can be used that tracks the face region in an image. On completion of the detection and tracking process, we can now find the person's face in subsequent frames. For the measurement of body temperature, we make use of the face facial features to get the forehead area as our Region of Interest (ROI) [19]. Using some basic temperature and ROI formulas the temperature can be calculated for the person in the frame.

## 5 Camera-Based Pulse Oximetry

We all know that one of the most important components of the human body is hemoglobin. But this hemoglobin is not soluble in blood. To make it soluble in blood oxygen is needed. This oxygen mixes with the haemoglobin (HB) and forms a component called oxyhemoglobin ( $\text{HbO}_2$ ). This  $\text{SpO}_2$  level is a ratio of the oxyhemoglobin present in the blood to the total hemoglobin content of the blood. A normal healthy person would have this  $\text{SpO}_2$  level between 95–99, and anything lesser than this can cause hypoxemia or organ compromisation, thus should be immediately treated. The whole principle behind doing this calculation is that HB and  $\text{HbO}_2$  have different optical absorption spectra. This is what creates the difference between oxygenated and deoxygenated blood.

Photoplethysmography (PPG), which is needed to find out the oxygen concentration in blood, is a non-invasive low-cost optical technique used to detect a volumetric change in peripheral blood circulation. Recent studies show that a normal camera or even a mobile phone's camera is capable of doing this photoplethysmography. For doing this, a strong signal on the green channel should be extracted from a video recording of a particular region of a human's face [20]. There will be a variation in that signal due to variations in the volume of subcutaneous blood vessels. In [21], the system works even if the subject is slightly moving. This was made possible by authors using a technique of blind source separation to extract a heart-related signal. Multi-wavelength measurement is also required to find out the oxygen saturation.

For the photodetector, a photodiode or discrete phototransistors are used in a typical pulse oximeter. A camera also captures and measures the light that comes from the subject along these lines. Its pixels assume the job of the photodetectors and the broad-spectrum ambient light joined with the color filters in the camera gives the estimation at various wavelengths [22–25].

## 6 Findings

See Table 1.

The below comparison Table 1 compares the different techniques used for monitoring the covid precaution Standard Operating Procedures (SOP) which are: 1. Face Mask Detection, 2. Safe Distance Detection, 3. Body Temperature Calculation and 4. Blood Oxygen Concentration Calculation. The technique being used for occluded face recognition is chosen according to the situation, scenario, and requirements. For a quick on-the-go analysis system, heavy models like CNN, R-CNN, and Fast R-CNN don't seem to be suiting since it slows down the process of detection. One of the major drawbacks of many techniques is that it fails when the face is in a non-frontal orientation, or the mask is very differently designed than those present in the dataset. Another challenge faced by most of the researchers is that there's no proper dataset of face images of the same people with and without masks. Moreover, the type and way of wearing a mask differ greatly from person to person. This also poses a great challenge. The Safe Distance Detection, in which Detection of humans is the major part, researchers face similar challenges of occlusion. Most of them have tried to tackle this using part-based tracking, but this does not give high accuracy when it comes to various forms of occlusion. Separate feature detectors need to be implemented to get good results. New generation infrared cameras clearly have helped the body temperature calculation. Now, using Computer vision techniques like ROI Locating and image processing, thermal images can be deduced which gives us parameters for temperature calculation. Similarly, widely used mobile cameras are being used and experimented with for performing camera-based pulse oximetry. Though there are a lot of developments that need to be done in this area such as developing an apt dataset to calibrate the camera, this non-invasive and remote method of calculating a blood oxygen concentration without having to clip apparatus to the patient's body is a great advancement and an exemplary area to work upon.

## 7 Conclusion

Automating these Covid Precaution norms monitoring processes has become a pressing need recently seeing the amount of manpower needed to do this manually. Bringing in all four important features that need to be monitored in a single project proves to be the most effective surveillance system at this hour. Thus, a lot of researchers have turned towards this challenge to provide a state-of-the-art technological solution. And for such researchers, this paper on technical survey various methods may become very useful to implement this fulfilled project. We call this a fulfilled project since all four conditions that need to be monitored are included in this system. The four conditions are face mask detection, safe distance detection, thermal imaging for body temperature calculation, and camera-based pulse oximetry. As can be seen in the findings section, the best algorithm for each condition should

**Table 1** Comparison of computer vision techniques used for SOP procedures

S.No	Feature	Year	Technique used	Dataset used	Limitations
1	Face mask recognition	2020	GAN based network using two discriminators	Paired synthetic dataset using publicly available CelebA dataset	When masks are very different from those in the dataset the map module does not produce a reasonable segmentation map of the mask object
2	Face mask recognition	2020	Multi-task cascaded convolutional neural network (MTCNN)	Masked face dataset (MFD), IIIT Delhi Disguise Ver. 1 Face Database, AR Face Database	Fails when the mask is in the same color as human skin, or if the mask is very different than those present in the dataset
3	Face mask recognition	2005	Self-organizing map (SOM), Soft nearest neighbor (soft-NN) ensemble method	Synthetic database; AR and FERET dataset used for testing	The proposed method assumes that the system can differentiate between occlusion of faces. But, in a general case, this is a difficult problem as part of the face may be occluded in any way
4	Face mask recognition	2011	Super resolution method using nonlinear mappings on coherent features	FERET expression database, UMIST Database, ORL Database	The model is useful only for High resolution pictures, which are not available in most cases
5	Safe distance detection	2014	Histogram of oriented gradient (HOG) algorithm	MIT pedestrian, city scene database	The method does not consider occluded human cases
6	Safe distance detection	2014	Heterogeneous features and ensemble of multiview-pose parts	Haar Cascade Dataset	It is inefficient for occlusion greater than 30%

(continued)

**Table 1** (continued)

S.No	Feature	Year	Technique used	Dataset used	Limitations
7	Safe distance detection	2015	Boosted multi-task model	Caltech, TUD-Brussels, and INRIA	The model needs to consider all possible detector features
8	Body temperature calculation	2019	Thermal camera based detection using ROI locating	Custom made Dataset	The method cannot handle the occlusion of ROI locations
9	Body temperature calculation	2017	Image processing based temperature calculation	Haar Cascade Dataset	Surrounding temperatures and humidity can cause errors
10	Oxygen concentration calculation	2014	Remote camera based pulse oximetry	-	A lot of work still needs to be done before commercializing this algorithm such as creating a database to accurately calibrate the camera-based pulse oximeter

be chosen based on the scenario and the requirements. If a perfect combination of algorithms is found to perform this full project, it can help in a great way to reduce the social spread of covid and in that case, it can be used for any other epidemic or pandemic situation too. With regard to future work, there is scope for further research to handle occlusion-based challenges for face-mask detection and human detection. In addition, we further plan to implement this model in real-time after choosing the appropriate algorithms that can cumulatively be efficient and at a faster pace as compared to currently available multi-featured systems.

## References

1. Everything You Ever Wanted To Know About Computer Vision (2021). Available at <https://towardsdatascience.com/everything-you-ever-wanted-to-know-about-computer-vision-heres-a-look-why-it-s-so-awesome-e8a58dfb641e>
2. Enhancing camera surveillance using computer vision: a research note (2018). Available at <https://www.emerald.com/insight/content/doi/https://doi.org/10.1108/PIJPSM-11-20160158/full.html>
3. D.N. Parmar, B.B. Mehta., Face recognition methods and applications. IJCTA (2013)
4. H. Jia, A.M. Martinez, Face recognition with occlusions in the training and testing sets (2008)
5. Face Detection Versus Face Recognition: What is the Difference (1998). Available at <https://www.facefirst.com/blog/face-detection-vs-face-recognition/>
6. P.F. de Carrera, Face Recognition Algorithms 16 (2010)

7. D. Zeng, R. Veldhuis, L. Spreeuwiers, A survey of face recognition techniques under occlusion (2020). [arXiv:2006.11366v1](https://arxiv.org/abs/2006.11366v1) [cs.CV] 19 June 2020
8. S. Ramesh, C. Yaashuwanth, B.A. Muthukrishnan, Machine learning approach for secure communication in wireless video sensor networks against denial-of-service attacks. *Int. J. Commun. Syst.* **33**(12) (2019). <https://doi.org/10.1002/dac.4073>
9. T.N. Nguyen, B.-H. Liu, S.-Y. Wang, On new approaches of maximum weighted target coverage and sensor connectivity: hardness and approximation. *IEEE Trans. Netw. Sci. Eng.* **7**(3), 1736–1751 (2020). <https://doi.org/10.1109/TNSE.2019.2952369>
10. A. Rosebrock, OpenCV social distancing detector on June 1 (2020). Available at <https://www.pyimagesearch.com/2020/06/01/opencv-socialdistancing-detector/>
11. W. Liu, B. Yu, C. Duan, L. Chai, H. Yuan, H. Zhao, A pedestrian-detection method based on heterogeneous features and ensemble of multi-view-pose parts. *IEEE Trans. Intell. Transport. Syst.* 1–12 (2014)
12. C.Q. Lai, S.S. Teoh, A review on pedestrian detection techniques based on histogram of Oriented gradient feature. in *2014 IEEE Student Conference on Research and Development* (2014)
13. Y. Jiao, H. Yao, C. Xu, PEN: pose-embedding network for pedestrian detection. *IEEE Trans. Circuits Syst. Video Technol.* 1–1 (2020)
14. C. Zhu, Y. Peng, A boosted multi-task model for pedestrian detection with occlusion handling. *IEEE Trans. Image Process.* **24**(12), 5619–5629 (2015). <https://doi.org/10.1109/tip.2015.248337>
15. A. Rosebrock, Measuring distance between objects in an image with OpenCV” on April 4, (2016). Available at <https://www.pyimagesearch.com/2016/04/04/measuring-distancebetween-objects-in-an-image-with-opencv/>
16. A. Sharma, A.R. Yadav, Image processing based body temperature estimation using thermal video sequence (2017)
17. J.-W. Lin, M.-H. Lu, Y.-H. Lin, A thermal camera based continuous body temperature measurement system. in *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)* (2019). <https://doi.org/10.1109/iccvw.2019.00208>
18. W. Verkruyse, L.O. Svaasand, J.S. Nelson, Remote plethysmographic imaging using ambient light. *Opt. Expr.* **16**(26), 21434–21445 (2008)
19. M.-Z. Poh, D.J. McDuff, R.W. Picard, Non-contact, automated cardiac pulse measurements using video imaging and blind source separation. *Opt. Expr.* **18**(10), 10762–10774 (2010)
20. U.S. Freitas, Remote camera-based pulse oximetry. in *eTELEMED 2014: The Sixth International Conference on eHealth, Telemedicine, and Social Medicine* (2014). ISBN: 978–1–61208–327–8
21. H.A. Rowley, S. Baluja, T. Kanade, Neural network-based face detection. *IEEE Trans. Pattern Anal. Machine Intell.* **20**(1), 23–38 (1998)
22. K.-K. Sung, Learning and example selection for object and pattern detection”. PhD thesis, (Massachusetts Institute of Technology, 1996)
23. F. Raphael, B. Olivier, C. Daniel, A constrained generative model applied to face detection. *Neural Process. Lett.* **5**(2), 11–19 (1997)
24. H. Schneiderman, T. Kenade, Probabilistic modeling of local appearance and spatial relationships for object recognition. in *Proceedings IEEE Conference Computer Vision and Pattern Recognition* (1998), pp. 45–51
25. P. Dwivedi, Using AI to detect social distancing violations (2020). Available at <https://medium.com/swlh/using-ai-to-detect-social-distancingviolations-4707301844be>

# Trust-Based Node Selection Architecture for Software-Defined Networks



Aditya Kumar and C. Fancy

**Abstract** A feature that encourages the disclosure of an equipment segment naturally has detonated quickly in the ongoing years. With the presence of circulated computing, various environment and business norms are encountering anticipated changes and may have the alternative to clear out their IT framework support measures. Persistent execution and high openness essentials have impelled telecom associations to get the ground-breaking thoughts of the cloud model: software-defined network (SDN). Software-defined networking development is an approach to manage network the board that grants dynamic, consequently capable association arrangement so the executives area steering for network accessibility and quality of service (QoS) the board of organization assets are the main concerns of the present associations, and these necessities drove towards programming characterized organizing. Programming-defined networking is a building approach that decouples the control plane from the information place. In this paper, we endeavour to bring up and portray the benefits of using trust score-based node determination engineering for SDN. We also present the different inconveniences confronting SDN, from flexibility to consistent quality and security concerns, and conversation about existing reactions for these difficulties. And also, we depict the advantage of utilizing trust score-based node selection architecture for SDN.

**Keywords** SDN · OpenFlow · Cryptography · Control plane · Data plane · MD5 · TLS

---

A. Kumar (✉) · C. Fancy  
SRM Institute of Science and Technology, Chennai, India  
e-mail: [ak9448@srmist.edu.in](mailto:ak9448@srmist.edu.in)

C. Fancy  
e-mail: [fancyc@srmist.edu.in](mailto:fancyc@srmist.edu.in)

## 1 Introduction

The modern networking has developed into an opulent monster that is trying to oversee and which battles to manage to the needs of some of the present conditions. SDN speaks to a substitution way to deal with computer networking that endeavours to manage these shortcomings of the current worldview. SDN generally novel gratitude to program the switches used in current information organizations. SDN's transition to an exceptionally adaptable and incorporated organization control design is most appropriate to the amazingly huge organizations common in the present super scale information centres. Programming-defined networking is a design approach that upgrades and streamlines network tasks by more intently restricting the communication (i.e. provisioning, informing, and disturbing) among applications and organization administrations and gadgets, regardless of whether they be genuine or virtualized. It frequently is accomplished by utilizing some degree of legitimately unified organization control, which is normally acknowledged as a SDN regulator, which at that point coordinates, intercedes, and encourages correspondence between applications wishing to associate with network components and network elements wishing to pass on data to those applications [1].

Traditional frameworks organization is set up in fixed-work network devices, like a switch or router. These contraptions each have certain limits that function admirably together and backing the association. If the association's abilities are executed as hardware works, by then its speed is consistently upheld. Versatility is a monotonous deterrent for traditional associations. Scarcely any application programming interface (APIs) are revealed for provisioning and most trading gear, and writing computer programs is selective. Customary associations routinely work honourably with selective provisioning programming; anyway this item cannot be promptly changed differing. Standard frameworks organization contains the going with characteristics.

Programming portrayed sorting out has the potential gain of creating a framework that strengthens data concentrated applications, as enormous data and virtualization. Enormous data and virtual machines are genuinely joined. Ingram Micro contends that “virtualization appropriation is being driven by huge information, and SDN gives the way to oversee virtual machines and huge information network traffic”.

## 2 Literature Review

There are numerous studies in the field of software-defined networking. Each of this research work provides advantages of software-defined networks and the challenges plus how they are addressed with this concept plus its scalability and reliability.

Kreutz et al. [2] characterize administrator to program their organizations which prompts adaptability. But others can program the organizations which prompts assault in SDN. Creator portrays a few danger vectors that may empower the adventure of

SDN weaknesses which are fashioned or counterfeit traffic streams, misusing weaknesses in sending gadgets, assaulting control correspondence, abusing weaknesses in regulator, administrator station, absence of confided in assets for criminological and remediation, and between the regulator and applications. In any case, they have not checked if these vectors are practical against genuine SDN regulator usage.

Lee et al. [1] characterized organization and arising advances; security weakness appraisal is a significant cycle that must be led against any framework. This system, the assault surface of network, should be completely explored and also surveyed relieve conceivable threat penetrates against SDN. Roused by which need, also uncover many assault situations that influence SDN to assault SDN, as well as test the assault situations against all three famous SDN regulators which are Floodlight, Open Daylight (ODL), and Open Networking Operating System (ONOS), and every regulator has diverse execution. Furthermore, we talk about the conceivable safeguard components against such application-began assaults. Two guard instruments that could ensure such assaults consent checking and static or dynamic examination.

Delta [3] discovering robotizing and normalizing the weakness recognizable proof cycle in SDNs built up a security evaluation structure, delta, that instantiates distributed SDN assaults in different test conditions. Fluff testing calculation empowers the administrator to lead inside and out testing of the information input dealing with rationale of a scope of OpenFlow segment interfaces like other examination work, it additionally has a few constraints. To begin with, some testing cases require introducing a predetermined specialist (i.e. application agent) to a SDN regulator. For counteraction, it attempted to limit the measure of human communication, and our structure can be worked with basic designs. In any case, a few cases, for example, adding new assault situations, require manual alterations to certain pieces of the system. For this situation, we can comprehend an assault situation through the log data; however, this may require another approach to deal with SDN control streams or messages.

Phan et al. [4] proposed RMOF model that permits numerous regulators in various OpenFlow subdomains of an organization partner with one another and an answer for steering cross-space bundles dependent on the model. Also, a steering arrangement is proposed dependent on the model to accomplish higher organization execution. Plan to stretch out the model to manage the other systems administration issues that need the coordinated effort of different regulators to explain, for example, load balance over numerous organizations.

Lusani Mamushiane [5] characterized the prerequisite like number of regulators required and where should they put to fulfil client explicit necessity and imperatives and having an externalized control plane which rise scalability, adaptation to non-critical failure and performance of the controller. So, this paper is centre around regulator arrangement to limit proliferation inertness of control traffic and CapEx associated with introducing another regulator. To improve the quantity of regulator, they utilized outline examination and whole statics. Also, to decide the ideal location to put the regulator, they use parcel around monoids (PAM) bunching calculation.

Knight et al. [6] provided an investigation of organization geography that has pulled in a lot of consideration in the most recent decade; however, it has been

hampered by an absence of precise information. Existing techniques for estimating geography have imperfections, and contentions about the significance of these have dominated all the more fascinating inquiries concerning network structure. The Internet Topology Zoo [7] is a store of organization information made from the data that network administrators unveil. As such, it is the most exact enormous scope assortment of organization geographies accessible and incorporates meta-information that could not have been estimated. With this information, we can respond to inquiries concerning network structure with more assurance than any other time in recent memory—we show its capacity through a fundamental investigation of the PoP-level geography of more than 140 organizations [8]. We locate a wide scope of organization plans not adjusting all in all to any undeniable model. This paper portrays another informational index—the Internet Topology Zoo—in light of manual record of public organization maps. It contains 232 organizations as of now, is as yet developing. We have just utilized this dataset to perform measurable examination of PoP-level organization geographies.

Mishra et al. [9] give a relative examination of different AI techniques for network interruption discovery was done. The assessments were carried on benchmark dataset (KDDCUP99) for interference acknowledgement. We have performed two arrangements of tests; one complete dataset having 51 highlights and start decreased one with just 21 components credits [10]. Tests indicated that order calculations do not depend all the 41 highlights; in this manner, the strategy could undoubtedly improve results with an apparent reduction in assets required by chipping away at the equivalent dataset with decreased number of qualities.

### 3 Project Description

#### 3.1 SDN and Its Component

SDN programmability generates map that is dynamic, reasonable, financially savvy, what is more, adaptable, accepting for the critical-information move limit, variable in nature to the current application built. That is why this designing dissembles the association control and sending limits enabling the organization control to end up being clearly plan, and the fundamental foundation be disconnected to specific application and association system. The OpenFlow shows clearly fundamental part to developing SDN plans [4].

SDN is a media communications networks engineering that gives the guarantee of critical enhancements in the organization execution. Utilizing programming characterized networks, it is conceivable to make the organization more powerful, reasonable, savvy, and versatile. The key behind programming characterized organizing is that the SDN designs decouple network rules and send out capacities. This empowers organization control to end up being clearly programmable [11–13]. Subsequently, basic organization structure can disconnect from implementations for organization

admins. As more prominent degrees of proficiency are needed alongside more noteworthy adaptability to fulfil the changing needs after some time in light of various use designs every day, week and with uncommon functions and such, it is important to use all the components inside the information network as adequately as could be expected under the circumstances. Programming characterized organizing empowers assets to be arranged to meet the capacities required at some random time and to guarantee that traffic can stream in the most ideal way consistently.

RFC7426 follows a methodology focused on network gadgets. Organization gadgets are made out of assets, straightforward, and complex, with network gadgets being mind boggling assets themselves, in this way permitting recursive definition and reusability. Organization gadgets can be actualized in programming as well as equipment. This term asset start utilized conventionally, independent real example/usage asset, which can be directly and indirectly. So, conventional utilization which is RFC7426 recognizes accompanying five SDN planes.

### **3.2 Traditional Network v/s SDN Comparison**

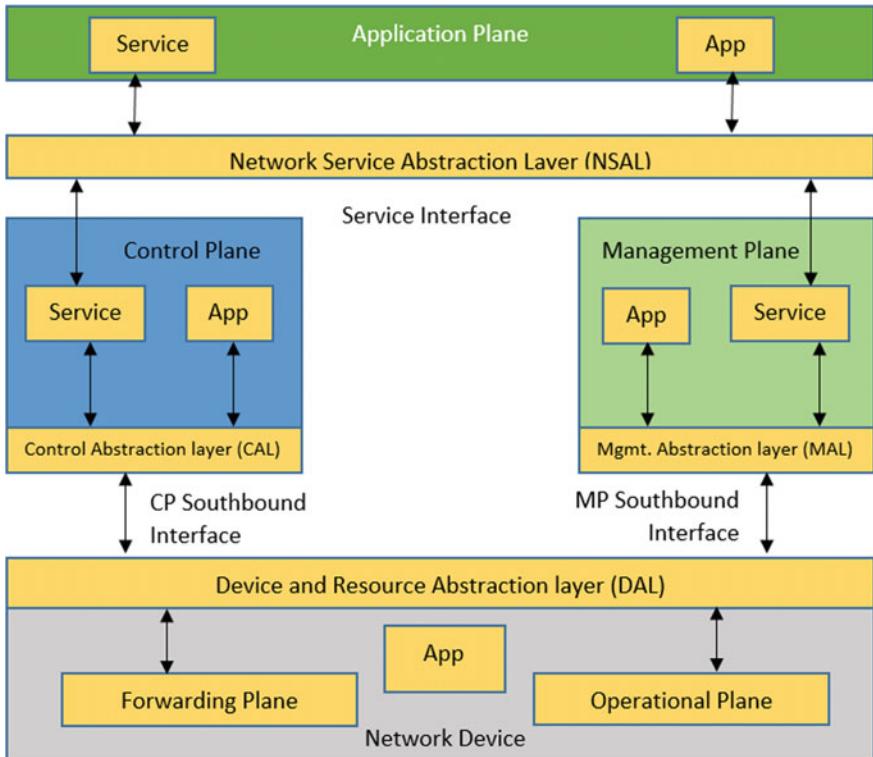
By setting up a focal control point for managing security and strategy data for your venture, the SDN regulator rapidly turns into a help for your IT office.

**Lower operating costs.** A few advantages to SDN, for example, having a proficient organization, worker use upgrades, and improved virtualization control, can dually help cut working expenses. Since numerous standard organization issues can be robotized and concentrated, SDN can likewise help diminish working expenses and develop authoritative reserve funds.

**Hardware savings and reduced capital expenditures.** SDN reception resuscitates more established organization gadgets and improves the way towards advancing commoditized equipment. By adhering to the guidelines from the SDN regulator, more established equipment can be repurposed while less exorbitant equipment can be conveyed to ideal impact. This cycle permits new gadgets to become authentic “white box” switches that have insight centred at the SDN regulator.

**Cloud abstraction.** Utilizing SDN to extract cloud assets streamlines the way towards bringing together cloud assets. SDN regulators can deal with all the systems administration parts that involve the huge server farm stages.

**Consistent and timely content delivery.** One major advantage of SDN is the capacity to control information traffic. It is simpler to have nature of administration for Voice over Internet Protocol (VoIP) and media transmissions in the event that you can coordinate and computerize information traffic. SDN likewise assists with steaming better recordings since SDN supports network responsiveness and, along these lines, makes an improved client experience (UX) [3] (Fig. 1 and Table 1).



**Fig. 1** Layered architecture RFC7426

**Table 1** Traditional network versus SDN

S. No	Characteristics	Traditional	SDN
1	Dynamics	Turns out to be exceptionally perplexing in multidevice and mobile environment	Rapidly adjusts to changing business needs
2	Policies and security	Requires device level changes	Policy applications becomes simplified
3	Scalability	Becomes problematic after a point due to network complexity	Easily scalable
4	Device control	Incomplete control	Vendor-agnostic
5	Resource utilization	Less	High
6	Maintenance	High	Less
7	Controller availability	Irrelevant	Important
8	Visibility	Lack of visibility	Centralized configuration gives more visibility

### **3.3 SDN Planes**

#### **3.3.1 Forwarding Plane**

Responsible for taking care of information packets dependent on the directions got over lower plane. Activities sending regular incorporate, yet far from restricted to, sending, release, and evolving data packets. Instances sending assets are allocation, unit, and so forth; the sending level is likewise generally alluded for the “information plane” or “information way”.

#### **3.3.2 Operational Plane**

Conduct manage for dealing with the in action condition of organization device like switch and routers, e.g. regardless of if the gadget works dynamic or rather dormant, and quantity of interfaces accessible, the level of each interface, etc. Instances of the operational plane assets are interface, storage, etc.

#### **3.3.3 Control Plane**

Supervise for deciding determinations the packets can be sent via at least less than two organization gadgets and forward such choices right enable the organization gadgets for performance. The control plane primary occupation fine-follow the sending tables that dwell in the sending plane, in light of the organization geography or outside help demands.

#### **3.3.4 Management Plane**

It manages for observing, arranging, also keeping up networking devices, e.g. settling on choices with respect to the condition of an organization gadget. The administration plane might be utilized to arrange the sending plane; however, it allows inconsistently and via a discount to meet than the control plane.

#### **3.3.5 Application Plane**

This plane deals with the applications and organizations which describe complete network direct abide. Application that is truly essentially uphold activity of the sending plane, (e.g. directing cycles reside into control plane) is not seen as a component of the application plane. These planes intimated higher than normal unit connected through ports [14, 15]. All interfaces might use various forms reckoning on whether or not the connected planes reside on a similar device or on completely

various devices. In event that the few planes region unit planned so they do not need to live inside a similar gadget; at that point, the interface can just appear as a convention [16–18]. On the off chance, the planes territory unit gathered on the comparable gadget; at that point, the interface possibly well be upheld by means of AN open/restrictive convention, open/exclusive programming bundle between measure correspondence API, or OS part framework calls [19]. RFC7426 centres around the north/south correspondence between elements in various plane; however, it does not excluding presence of connection at intervals anybody deploy plan.

### ***3.4 OpenFlow Specification***

The OpenFlow particular has been advancing for various years. The non-benefit Internet association, OpenFlow organization was made in year 2008 for securing from advance and hold up OpenFlow. It is without a doubt an alternate class of advancements remembered for the SDN Big Tent. It is a significant device and impetus for advancement and characterizes some portion of the information plane's activities alongside correspondences convention linking to programmability information plane and the SDN control plane [20]. It does not clarify the regulator's own activities. An OpenFlow framework comprises of an OpenFlow regulator that conveys to at least one OpenFlows. This convention characterizes few particular information and message designs traded between regulator (control plane) and gadget (information plane). The OpenFlow conduct determines when the gadget ought to respond in different circumstances and how it ought to react to orders from the regulator [12]. OpenFlow engineering comprises of different parts. This paper depicts key segments.

#### **3.4.1 OpenFlow Switch**

An OpenFlow switch involves in any event one stream tables and a device occasion table, which perform package questions and sending, and at any rate one OpenFlow channels to an exterior controller device [21]. The switch talks with the controller, and the controller manages the switch through the OpenFlow switch show. Utilizing the OpenFlow switch show, the regulator can add, update, and erase stream sections in stream tables, both responsively (because of bundles) and proactively. Every stream data in that switch have a large stream segments; every stream entrance includes coordinate data fields, counter, and a ton of rules to applicable to organizing packs [13].

#### **3.4.2 OpenFlow Protocol**

OpenFlow show portrays the correspondence between an OpenFlow controller and an OpenFlow device [22]. This show is what most strikingly recognizes OpenFlow

development at its essence; the show incorporates a huge load of messages that are sent from the regulator to the switch and a differentiating course of action of messages that are sent the other way. And OpenFlow show has grown inside and out with each interpretation of OpenFlow [23].

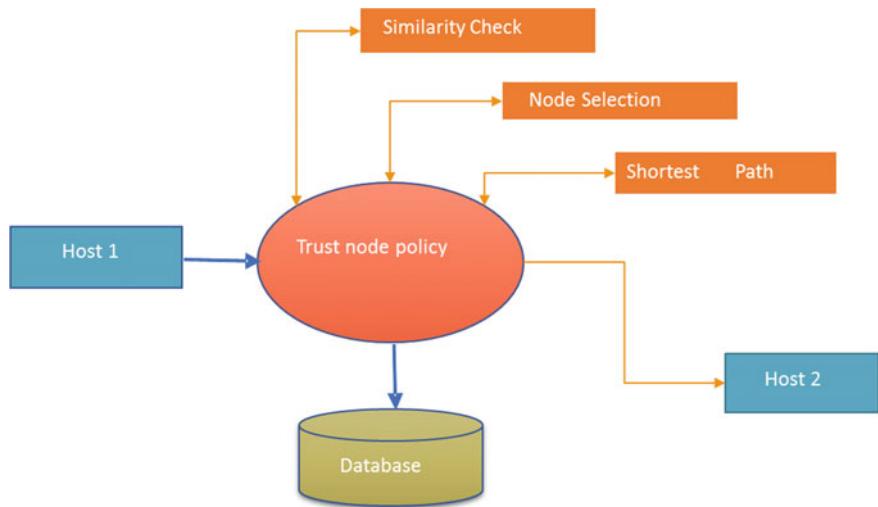
### 3.4.3 OpenFlow Channel

This channel interface associates all OpenFlow Logical Switch an OpenFlow regulator. From all of these interfaces, regulator designs also compromise with the switch and bring functions from the data layer device, and let out packets from data layer device [24]. The control channel of the switch may uphold a solitary OpenFlow channel with a solitary regulator, or various OpenFlow channels empowering different regulators to share the board of the switch. Between the information way and the OpenFlow channel, the interface is execution explicit; anyway, all OpenFlow channel messages must be arranged by the OpenFlow switch convention. The OpenFlow channel is typically encoded utilizing TLS; however it might be run straightforwardly over TCP [25].

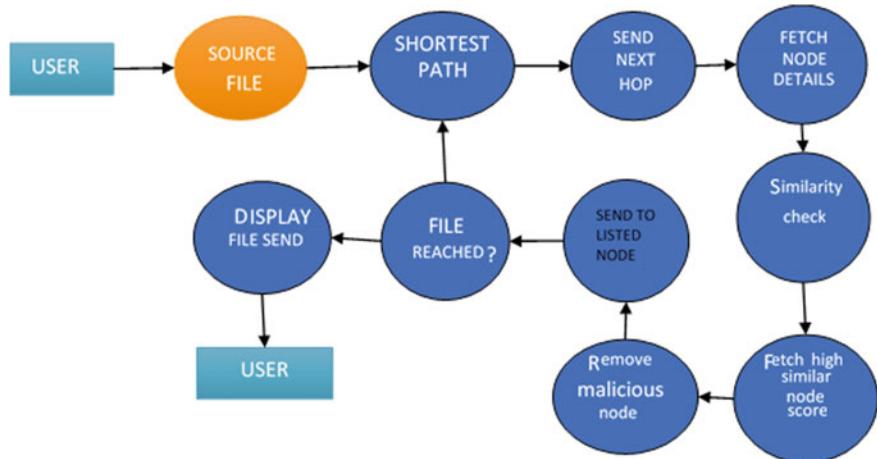
## 4 Proposed Work

Software-defined networking (SDN) decouples the conventional shut organization into information plane, control plane, and application plane, which empowers intelligently incorporated control and the executives of the entire organization. With this new plan rule, the organization could carry on more deftly and can undoubtedly adjust to the requirements of various associations. Additionally, the concentrated engineering permits significant data to be gathered from the organization and thus used to improve and adjust their approaches powerfully (Fig. 2).

We have demonstrated the specification of fine-grained similarity-based node selection policies on a variety of attributes, setting data, for example, area and steering data, and administrations got to in SDN. To find the similarity between the two nodes, this system uses cosine similarity algorithm [26]. If the cosine value of two nodes attributes is high, then the both the nodes are similar; else, it is not similar nodes. Security model in this system will work when sending the packet from source to destination through the shortest path; if in that path, any nodes is misbehaved, then this system automatically find that node and it will change the path, and then misbehaved node security trust score will be decreased. If a node successfully sends the packets, then security trust score will be increased, which helps the system to identify the misbehaved nodes and good behaved nodes in next routing time (Fig. 3).



**Fig. 2** Proposed system architecture



**Fig. 3** Proposed system algorithm

## 4.1 Algorithm Used for the Proposed Model

### 4.1.1 Node Similarity

This methodology is used to find the similarity between the two nodes using cosine similarity algorithm. If the cosine value of two nodes attributes like place, designation, and language, this is useful to find that where the nodes related to the same terms. And updating the similarity score in database.

#### • Cosine Algorithm Similarity calculation

$$\begin{aligned}
 d1 &= (5, 0, 3, 0, 2, 0, 0, 2, 0, 0) \\
 d2 &= (3, 0, 2, 0, 1, 1, 0, 1, 0, 1) \\
 d1 \cdot d2 &= 5 * 3 + 0 * 0 + 3 * 2 + 0 * 0 + 2 * 1 \\
 &\quad + 0 * 1 + 0 * 1 + 2 * 1 + 0 * 0 + 0 * 1 = 25 \\
 \|d1\| &= (5 * 5 + 0 * 0 + 3 * 3 + 0 * 0 + 2 * 2 + 0 * 0 + 0 * 0 + 2 * 2 + 0 * 0 + 0 * 0)^{0.5} \\
 \mathbf{0.5} &= (42) \mathbf{0.5} = 6.481 \\
 \|d2\| &= (3 * 3 + 0 * 0 + 2 * 2 + 0 * 0 + 1 * 1 + 1 * 1 + 0 * 0 + 1 * 1 + 0 * 0 + 1 * 1)^{0.5} \\
 \mathbf{0.5} &= (17) \mathbf{0.5} = 4.12 \\
 \cos(d1, d2) &= 0.94
 \end{aligned}$$

We are not intrigued by the words themselves however. We are intrigued distinctly in those two vertical vectors of tallies. For example, there are two occurrences of “me” in every content. We will choose how close these two writings are to one another by ascertaining one capacity of those two vectors, to be specific the cosine of the point between them.

### 4.1.2 Find the Neighbour Node

In this model, we are finding the neighbour node using distance calculation process; for each node, we are finding the neighbour and then finding the shortest path between the nodes.

$$\text{Math} \cdot \sqrt((x2 - x1) * (x2 - x1) + (y2 - y1) * (y2 - y1));$$

### 4.1.3 AES Algorithm for Encryption

The AES encryption algorithm will encrypt the text files; images to block cypher consist of a block length of 128 bits that uses the same encryption key to perform several rounds of encryption.

#### 4.1.4 MD5 Algorithm for Data Packet

It is a TCP alternative that adds a MD5-based mark to each TCP parcel. It signs the source and objective IP locations, ports, and the payload. In that way, the information is both validated and honesty secured.

### 5 Implementation

A trust scoring system is proposed in this model to measure the trustworthiness of a network nodes based on a metric termed trust score. Trust score varies dynamically with respect to time and is based on node behaviour; if the node transfers the data packets properly, then its score will be increasing, and if the node is dropping the packets, then the score will be decreased. Initially, all nodes have score 50; the score range is (0 to 100). If the node did data transfers correctly, then score increases step by step; if the node become malicious and it keeps dropping data packets, then its score will decrease gradually.

When a network node has to select the next hop node to transfer the data packet, it will find the nearest similar nodes, and then it will check the trust score value; based on the trust score only, next hop node will be shortlisted, and data packet will send to that node. It helps the system to identify the misbehaved nodes and good behaved nodes in next routing time.

Security model in this system will work when sending the packet from source to destination through the shortest path.

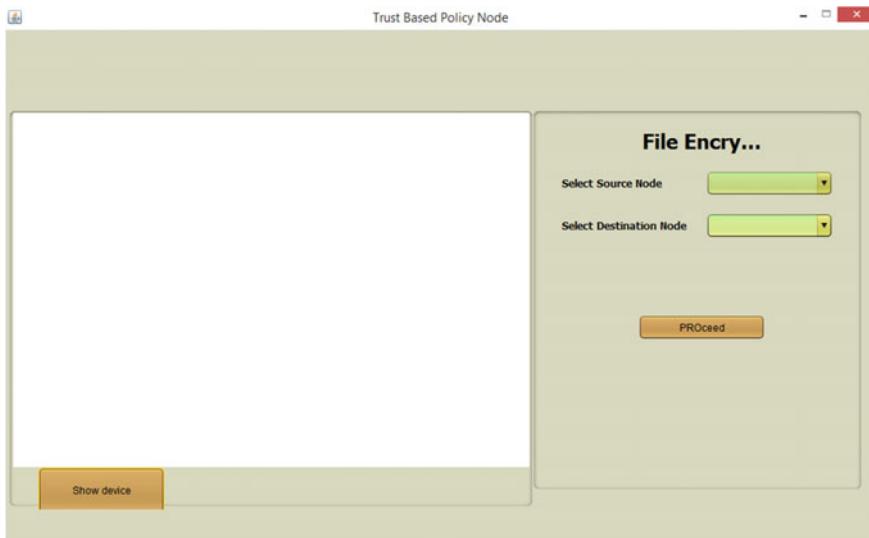
#### 5.1 Security and Result

Security in the proposed system is ensured by calculating the hash value of the packet using MD5 algorithm and storing it in the database, and then calculating the hash value of the packet again when it reaches to the node. This will identify if any packet is tampered with during transit.

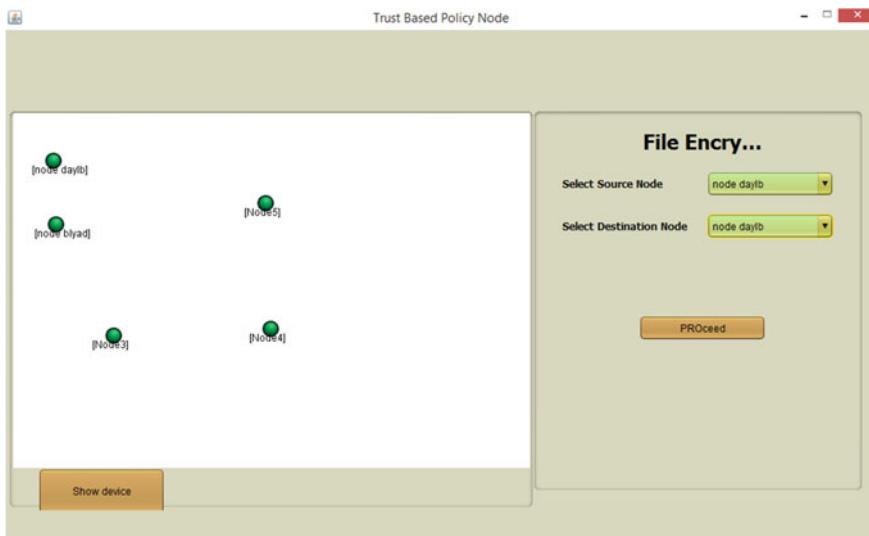
The results or output of this model will be shown on SWIM framework. Results will give information about the bad paths and defected nodes (Figs. 4, 5 and 6).

### 6 Conclusion

In this paper, we have tried to present the present the difference between traditional network and SDN, the OpenFlow specification. We have also covered the network function virtualization and did the comparison between SDN and NFV. This paper also describes the importance of policy-based model of SDN which uses ONOS,



**Fig. 4** Nodes display panel



**Fig. 5** Selection of path (trusted nodes)

	node_id	node_name	node_x	node_y	node_type	country	language	affiliation	position	topic_of_interest
1	node dayb	42	51	node	Germany	German	IBM	Researcher	Network management	
2	node 2	45	126	node	Japan	Japanese	University of Piraeus	Professor	Ad hoc mobile networks	
3	Node3	113	257	node	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)
*	(NULL)	(NULL)	(NULL)	node	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)	(NULL)

**Fig. 6** Details of nodes in SQL database

**Table 2** Hardware

System	Pentium IV 2.4 GHz
Hard Disk	500 GB
RAM	4 GB

**Table 3** Software

Operating system	Windows 8
Coding language	Ellipse
Database	SQL database
Database GUI	SQLYog
Ellipse tool	Eclipse lunar

it benefits and functions. In this paper, we have proposed a system with hardware and software specifications for identification of evaluation of trust of each node of distributed SDN architecture, allowing secure intra- and intra-secure inter-domain interactions and flows between various communications via various domains, end hosts. Our security layout uses an approach focused on policy terminology to define protection requirements at a fine granular level to regulate the flow of knowledge in an SDN multi-domain environment.

### 6.1 Hardware and Software Requirements

See Tables 2 and 3.

## References

1. R. Langner, Stuxnet: dissecting a cyber-warfare weapon. *IEEE Secur. Priv.* **9**(3), 49–51 (2011)
2. D. Kreutz et al., Towards secure and dependable software-defined networks. in *Proc of 2nd ACM SIGCOMM workshop on Hot topics in SDN*. (ACM, 2013), pp. 55–60
3. S. Lee et al., The smaller, the shrewder: a simple malicious application can kill an entire sdn environment. in *Proc 2016 ACM International Workshop on Security in Software Defined Networks and Network Function Virtualization*. (ACM, 2016), pp. 23–28
4. “Delta: A security assessment framework for software defined networks, in *Proc of NDSS*, vol. 17 (2017)
5. M.S. Kang et al., The crossfire attack. in *Security and Privacy (SP), 2013 IEEE Symposium on*. (IEEE, 2013), pp. 127–141
6. X.T. Phan et al., A collaborative model for routing in multi-domains open flow networks. in *Computing, Management and Telecommunicatons, International Conference on*. (IEEE, 2013), pp. 278–283
7. S. Knight et al., The internet topology zoo. *IEEE J. Sel. Areas Commun.* **29**(9), 1765–1775 (2011)

8. M.K. Shin, K.H. Nam, H.J. Kim, Software-defined networking (sdn): a reference architecture and open apis. in *ICT Convergence (ICTC) 2012 International Conference*, Oct (2012), pp. 360–361
9. P. Mishra et al., A detailed investigation and analysis of using machine learning techniques for intrusion detection. *IEEE Commun. Survey Tutorials*. Accepted, 2nd April (2018)
10. P. Sharma, S. Banerjee, S. Tandel, R. Aguiar, R. Amorim, D. Pinheiro, Enhancing network management frameworks with sdn-like control. in *Integrated Network Management (IM 2013) 2013 IFIP/IEEE International Symposium*, May (2013), pp. 688–691
11. M. Balaanand, N. Karthikeyan, S. Karthik, Designing a framework for communal software: based on the assessment using relation modelling. *Int. J. Parallel Prog.* **48**(2), 329–343 (2018). <https://doi.org/10.1007/s10766-018-0598-2>
12. D.V. Pham, G.L. Nguyen, T.N. Nguyen, C.V. Pham, A.V. Nguyen, Multi-topic misinformation blocking with budget constraint on online social networks. *IEEE Access* **8**, 78879–78889 (2020). <https://doi.org/10.1109/ACCESS.2020.2989140>
13. P. Goransson, C. Black, T. Culver, in *Software Defined Networks—A Comprehensive Approach*, 2nd edn. (2014)
14. P. Goransson, C. Black, T. Culver, in *Software Defined Networks—A Comprehensive Approach*, 2nd edn. (2016), pp. 89
15. P. Goransson, C. Black, T. Culver, in *Software Defined Networks—A Comprehensive Approach*, 2nd edn. (2016), pp. 93
16. R. Chayapathi, S.F. Hassan, P. Shah, in *Network Functions Virtualization with a touch of SDN* (2016), pp. 01, 05
17. Y. Zhang, Network function virtualization. in *Concepts and Applicability in 5G Networks* (2017), pp. 39–41
18. L.S et al., Athena: a framework for scalable anomaly detection in software-defined networks. in *47th IEEE/IFIP International Conference on Dependable Systems and Networks*, June (2017), pp. 249–260
19. D. Clark, Policy routing in internet protocols. Request for comment rfc-1102. (Network Information Center, 1989)
20. D. Estrin et al., Security issues in policy routing. in *Security and Privacy, Proceedings IEEE Symposium on* (IEEE, 1989), pp. 183–193
21. P. Thai et al., Decoupling policy from routing with software defined interdomain management: interdomain routing for sdn-based networks. in *Computer Communications and Networks (ICCCN), 2013 22nd International Conference on* (IEEE, 2013), pp. 1–6
22. J. Reich et al., Modular sdn programming with pyretic. Technical Reprot of USENIX (2013)
23. P. Goransson, C. Black, T. Culver, in *Software Defined Networks—A Comprehensive Approach*, 2nd edn. (2016), pp. 72
24. S. Scott-Hayward, S. O'Callaghan, G. Sezer, SDN security: a survey. in *IEEE Conference on Network Softwarization (NetSoft)*, November (2013), pp. 1–7
25. P. Goransson, C. Black, T. Culver, in *Software Defined Networks—A Comprehensive Approach*, 2nd edn. (2016), pp. 76
26. Virtual router redundancy protocol (VRRP), IETF RFC5798, March (2010)
27. T.D. Nadeau, K. Gray, in *Software Defined Networks: An Authoritative Review of Network Programmability Technologies* (2016)
28. T.D. Nadeau, K. Gray, in *Software Defined Networks: An Authoritative Review of Network Programmability Technologies* (2013), pp. 71–73

# A Review on Video Tampering Analysis and Digital Forensic



Pavithra Yallamandala and J. Godwin

**Abstract** Digital evidence collection and analysis have become an increasing tool to solve crimes and prepare courts' cases over the last two decades, undergoing major changes in the area of IT. Crime is a major problem every day, so that computer forensics are avoided and protected from crime. More information is created, stored and accessed with increasingly portable and powerful technology. Mobile systems may serve as large personal knowledge archives in a wallet still accessible through a hand or phrase. The advantage is obvious by having ample information in order to obtain judgments, but the collection and admissibility of digital proof should be balanced with the privacy concerns of law enforcement and other parties to criminal law. The need of validating the honesty of digital video content ranges from a person to associations, obstacles and security arrangements to law authorization/organizations'. With video and image changing, the change tools have made it simple to modify media content. Therefore, it is necessary to investigate viable methods for video falsification.

**Keywords** Video tampering · Object detection · Forensic video analysis · Image/Video enhancement · Forensics investigation · Machine learning · Deep learning

## 1 Introduction

Over the past decade or so, computer forensics digitally collecting evidence that emphasised disc imaging procedures has changed understandable procedures and methodologies. In certain surveillance networks, multimedia images are now considered to have a significant part in the investigative detection of crimes. Nearly six

---

P. Yallamandala (✉) · J. Godwin

Department of Information Technology, SRM Institute of Science and Technology,  
Kattankulathur, India

e-mail: [Yp3739@srmist.edu.in](mailto:Yp3739@srmist.edu.in)

J. Godwin

e-mail: [godwinj@srmist.edu.in](mailto:godwinj@srmist.edu.in)

million CCTV devices are now available in Britain, comprising 750,000 in “sensitive locations,” including banks, police stations, office buildings, and hospitals, and public places in the United Kingdom, including airports, shopping malls, hotels, and road junctions, aside from private surveillance. This provides vast quantities of photographs of photographic and video material. In 2015, 1 trillion photographs were also taken. In addition to the cost of capturing pictures, this significant increase in the number of images took place due to the increase in store media. As a result, massive digital pictures must be investigated on evidence or criminal scenes.

By incorporating facial reconnaissance services, several new CCTV systems are used to classify cyber offenders and perpetrators. Over the past two years, other services have been widely researched, such as motion identification, body and face recognition, cross-position recognition and gastronomic recognition. In some cases, human beings using the face, body, etc. are difficult to identify. In certain cases (poor conditions of viewing). Although many techniques of image treatment have been developed in recent decades, most of them don't benefit from facial, body, etc.

The forensic investigation has a number of parameters, such as video quality. Because the picture quality largely depends on research into video-related forensic analysis, poor quality significantly reduces the sense of credibility in the investigative process and therefore would not appear in court. The prime objective of the forensic surveillance video is to detect strong evidence at different levels. The massive enormous quantity of electronic audio / visual material that cheap and compact devices such as digital cameras and mobile phones create every day is combined with the large increase in video surveillance. All this material does not only function for leisure, but also for activities in all parts of the globe. Within the context of police or forensic science, intelligence, politics and journalism, the knowledge given in visual photographs and videos is focused on a variety of significant and consequential decisions. For example, surveillance videos of a crime during court trials can when appropriate, be admitted as ‘visual evidence’ and the substance of the crime must be truly portrayed.

Ten years before, digital videos could have been unfailing, but the cost-effective and easy to use low-cost video editing software such as Adobe Premiere, Photoshop, Cinelerre and Lightworks, as well as the development of fake techniques, made it clear that this is no longer the case. In this digital era, we are increasingly reliant for trustworthy evidence of events on multimedia content, especially digital images & videos. Even beginners can now alter Abstract content. However, there have been major concerns regarding the confidence of several sophisticated, but user-friendly content editing software. In recent years, forensics of visual media have therefore become a necessary field of research, mainly in the development of tools and techniques to determine if digital media is accurate or not. This field of research has shown enormous growth and innovation in the last two decades. This research presents a broad and examinational library reviewing published literature, mainly forging, video capture and phylogenetic identification, and forensics for video counters. The article analyses further the research breaches identified in the literature, gives valuable insights into contemporary research and offers guidelines for developers and future researchers to investigate new approaches in the field of video forensics. This study

is primarily aimed at producing a contour that is suitable both for those who work in digital video forensics and for those who are not yet digital videos and who do not differentiate them from authentic content. There are several types of fake1, but they are usually all in one of two categories: interframe falsification or intra frame fake.

(a) ***Inter-frame forgeries***

These are the sort of falsification which in some way affects the frame sequence in a video. These forgeries usually involve deleting or inserting a set of pictures into or from the video sequence. Frame reproduction or replication is also an interframe falsification that copies and inserts a set of frames in a temporal location in the same video. Such forgers may also be known as the ‘copy-paste forgings’ interframe. Time splicing, which interplays frames of two or more separate videos to produce a new video, is another form of interframe forgery.

(b) ***Intra-frame forgeries***

The actual contents of each frame are changed in an intraframe falsification. Examples of interframe forgeries are copy-paste and upscale-crop. In this study, we will present the review on the mechanisms and investigations of forensic analysis which have been dealt with in one way or another. First, we will examine investigations relating to the collection and extraction of important data from the collection of digital information from the criminal scene. Then we talked about the current classification algorithm and tried to find out which option could be better. We'll then focus on the CNN classification algorithm for the detection of forensic crime identification on the basis of digitally collected data.

## 2 Literature Survey

The paper [1] provides an analysis system for mobile forensics that recognises the user's mobile phone camera and its source with pictures taken from its patter noise sensor. Furthermore, they demonstrate the best way to connect the real user through the Mobile Camera Fingerprint (MCF) to their social networks account. Their strategy can combine isolated forensic mobile methods with digital forensic methods.

The increasing availability of recording devices calls for an effective video analysis method. In [2], forensic analysis methods are introduced for video formatting and anti-forensic techniques. The proposed methods can forensically evaluate the filtering and quantization parameters that are crucial to video editing anti-forensics. As video footage is available more and better, an enormous amount of video files can be analysed automatically. This paper [3] proposes the effective method of forensic identification and reconstruction of cached data on the online video stream, including YouTube, Twitter, Facebook, and WeChat.

In paper [4] the creators present a Real-Time framework based on the Faster R-CNN (regional Convolutionary Neural Network), which naturally distinguishes the

objects in the indoor world. The creators applies it to an ImageNet subset of 12 Item classes and Karina datasets to test the adequacy of the proposed framework. Their work was normally accurate at 74.33% and their performance was 0.12 s in Nvidia-TitanX GPU for each picture to identify objects.

The writer presents risk-sensitive proof processing technique for the selective retrieval of objects during the initial computer forensics collection process in the article presented in [5]. On dead networks with normal communication to target computer systems, this technique may be carried out, e.g. IDE converters or client/server applications for disc redirection, which allows limited artefact retrieval into a forensic network. The greatest benefit of this approach can be detected and removed from a dead system, thus allowing routine processes to proceed with the original systems, while recognising and extracting selective objects of significance from that system. This methodology includes the search and filtering of preacquisition evidence to minimise the collection of irrelevant data, which occurs following acquisition during laboratory examinations in traditional bit stream methodology.

A variety of video enhancing strategies have been developed over the last decade for video detection systems, smart road systems, safety control systems, etc. Tzanidou et al., for instance, have proposed a baggage detection with colours in [6] for low-quality video footage. This analysis enables the moving direction of human-like templates to be identified and aligned with the best matched examples.

Chuang et al. also developed a history-based ratio tracking model in [7] which is a colour histogram variable for colour detection. From the point of view of forensics, the main objective is to obtain information from low quality videos as much as possible. This section focuses on techniques which can help to find additional information from enhanced videos. The techniques based on Histogram Equalization (HE) can greatly improve the chances of finding additional information from video / images of low quality.

Due to the many years of the development and analysis of a wide range of image forensic methods, several research studies [8] have been carried out. However, only a couple of papers [9–11] assessed video forensic innovations. These papers have some of their own fine qualities, but in some ways they are inadequate. In [9], only two techniques for forgery detection were analysed [12] and only a few forensic video techniques were appreciated (the rest of the paper focuses on image manipulation. Notable and contemporary contributions have not been discussing or analysed. As it is expected that an effective survey in any particular area will be a thorough repository of all relevant and significant research and innovation, they may eventually be unidentified if certain important technology to detect videos are not quoted. Moreover, without such crucial references, the current state of affairs of the digital video forensics domain would become difficult to understand fully.

A recent study [13] looks at the techniques of passive video forgery and provides an informative study of the fundamentals in the field. However, several significant works of merit have also been overlooked. There was some limited description of the techniques discussed in the survey, the nature of test data employed by the respective authors, or of the qualitative or quantitative performances of the techniques proposed. The absence of this information can adversely affect the subject's understanding.

The research paper in [13] lacks a detailed examination, as well as the video anti-forensics domain and oen issues to be addressed in the near future. An analysis of improvements in forensically educated knowledge is given in a notable study [14], in which writers have studied different social variables and conduct patterns in forensic fields, such as forensic unit, embedded fingerprints and watermarking, and environmental signatures, for example, the frequency of electrical networks (ENF). The survey offered an incredibly brief description of the identification of manipulators and anti-forensic areas, but almost exclusively included digital pictures, with little forensic characteristics to identify forgeries. This thesis describes the most substantial contributions, including a wide variety of other research fields like video updating, removal and recordings, in the sense of the passive blind deceptive identification and anti-forensics.

We describe, review in depth the literature available to us and illustrate the positive aspects and shortcomings of each of the approaches discussed. False or video identification forensics. We are still addressing transparent topics and certain other long-term priorities that need urgent attention. In this discussion, we are looking to establish a perspective on the area of video production, which helps video scientists and practitioners to develop new research challenges.

Its key purpose is to provide the experts and professionals interested with interactive video forensics, as well as researchers and enthusiasts who are new to the area and still unstuffed, with detailed scientific aspects of video forensics. A simple and simplistic interpretation of the subject matter was important for this reason, which is why a detailed and nuanced scientific explanation of each particular approach was avoided.

The authors classified digital evidence in paper [15]. The vast array of digital devices and processes for extraction provide an adequate potential for retrieval. The author notes briefly the most common results from the processing of digital evidence. The whole statement is not true, but focuses on the key fields of data, all of which offer an explanation of how digital evidence can influence criminal justice and the future problems posed by departments as it is gathered, processed and used. Following are some of the resources from which the Internet, computers, portable electronics, etc. can be captured.

The authors note in paper [15] that there are critical forensic cloud problems, including physical inaccessibility, data from many people interwoven, and chain-of-custody issues where the location of the server is unknown. This issue author tested a variety of approaches for the isolation of Cloud extraction data with limited success comparable to the attempts of Cusack and Son at social network pages, but concluded that the best possible combination of the technologies at present was and that future experiments were needed.

### 3 Related Work

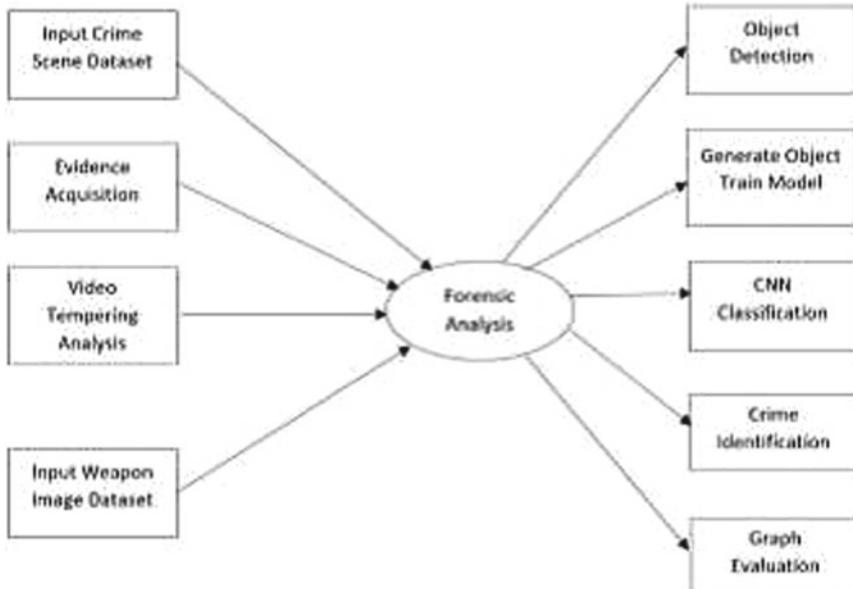
After Fig. 1 shows the architecture of the proposed system. The project will be conducted in subsequent phases.

**Step 1:** Video capture and format transition in forensic format this stage focuses on video capture, screen and analogue video imports with old format, playback time and multiple video processing time. Because of the diversity of capture equipment, various digital videos such as avi, mp4, etc. are not displayed in standard formats. These videos should not be diminished, but should be converted into standard formats.

**Step 2:** Enhanced video groups must be improved in order for more features to be provided and for hidden information to be discovered. Data can be obtained from different cameras in different areas, however available in the same incident.

**Step 3:** Extensive video forensic analysis This stage should be able to reduce various noise types in video, video editing, extraction, and filtering products, including extensive video analysis, including noisy audio recordings, source identification, and enhancement. The editing software can simultaneously preview multiple filters and apply them to combine them.

**Step 4:** Tampering Detection Denote the image that was manipulated as  $W'$ .  $W'$  is also divided into 8 blocks, 8 blocks that do not overlap. Afterwards, the 64 bits are extracted from each  $W'$  block's 1 LSB plane and separated in 48 recovery bits and 16 authentication bits by the same key. Feed into the Hash function 448 bits of the 7 MSB's and 48 retrieval bits extracted of 1 LSB and get the 16 recalculated



**Fig. 1** Architecture for proposed system

authentication bits for every suspicious block. If the authentication bits are different from the 16 authentication bits extracted, the authentication bits are marked as twisted in the 8 €block or reversed in the 8 €block. Note that reversed blocks can remove intact recovery bits and recalculate accurately the compression bits of reversed blocks too.

**Step 5:** Advanced-Recovery of evidence and object sensing by combining existing data in social networks, 3D geometries, AI, technology for machine-learning to extract further potential Proof.

**Step 6:** Use of the algorithm for learning machines (CNN)—In many cases, the colour of a particular object in your video is not stable due to several factors, such as the change of light, video quality, other similar coloured objects, etc. Fortunately, machine learning developers can provide much better identification.

## 4 Conclusion

This article describes a set of information on the types of tamper attacks that a video can experience and a detailed source of resources for passive-blind techniques used to detect such attacks. The area of anti-forensics and anti-forensics has also been discussed. In addition to analysing the most important pitfalls and realistic benefits of each strategy, we have addressed the limits that need to be resolved in the longer term and certain open questions that require urgent attention. In this work, we suggested a method that would increase the nature of the video as well as the facts. We suggested a video-based network investigation analysis structure. We have proposed, in particular, a technique to reverse further evidence. It is too useful for harmful acts or rapid reactions when exercising or practicing crime. We will build the links between the existing evidence and the detected evidence items later on. We used CNN algorithm here to classify forensic data.

## References

1. S. Li, Q. Sun, X. Xu, Forensic analysis of digital images over smart devices and online social networks. in *2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems (HPCC/Smart City/DSS)*, June (2018), pp. 1015–1021
2. T. Gloe, A. Fischer, M. Kirchner, “Forensic analysis of video file formats”. digital investigation. in *Proceedings of the First Annual DFRWS Europe*, vol. 11, (2014), pp. S68–S76
3. G. Horsman, Reconstructing streamed video content: a case study on youtube and facebook live stream content in the chrome web browser cache. *Digit. Investig.* **26**, S30–S37 (2018)
4. S. Saikia, E. Fidalgo, E. Alegre, L. Fernández-Robles, Object detection for crime scene evidence analysis using deep learning, in *Image Analysis and Processing—ICIAP 2017*. ed. by S. Battiat, G. Gallo, R. Schettini, F. Stanco (Springer International Publishing, Cham, 2017), pp. 14–24

5. E.E. Kenneally, C.L.T. Brown, Risk sensitive digital evidence collection. *Digital Investig.* **J. 2**(2), (Elsevier, 2005)
6. G. Tzanidou, I. Zafar, E.A. Edirisinghe, Carried object detection in videos using color information. *IEEE Trans. Inf. Forensics Secur.* **8**(10), 1620–1631 (2013)
7. C. Chuang, J. Hsieh, L. Tsai, S. Chen, K. Fan, Carried object detection using ratio histogram and its application to suspicious event analysis. *IEEE Trans. Circuits Syst. Video Technol.* **19**(6), 911–916 (2009)
8. H. Farid, Digital doctoring: how to tell the real from fake. *Significance* **3**(4), 162–166 (2006)
9. R.S. Ram, S.A. Prakash, M. Balaanand, C.B. Sivaparthipan, Colour and orientation of pixel based video retrieval using IHBM similarity measure. *Multimedia Tools Appl.* **79**(15–16), 10199–10214 (2019). <https://doi.org/10.1007/s11042-019-07805-9>
10. T.N. Nguyen, B. Liu, N.P. Nguyen, J. Chou, Cyber security of smart grid: attacks and defenses. in *ICC 2020–2020 IEEE International Conference on Communications* (ICC), Dublin, Ireland (2020), pp. 1–6. <https://doi.org/10.1109/ICC40277.2020.9148850>
11. A. Rocha, W. Scheirer, T. Boult, S. Goldenstein, Vision of the unseen: current trends and challenges in digital image and video forensics. *ACM Comput. Surv.* **43**(4), 26 (2011)
12. V. Joshi, S. Jain, Tampering detection in digital video e a review of temporal fingerprints based techniques. in *Proceedings of 2nd International Conference on Computing for sustainable global development*, New Delhi, India (2015), pp. 1121–1124
13. K.K. Sitara, B.M. Mehtre, Digital video tampering detection: an overview of passive techniques. *Digit. Investig.* **18**, 8–22 (2016)
14. M.C. Stamm, M. Wu, K.J.R. Liu, Information forensics: an overview of the first decade. *Access IEEE.* **1**, 167–200 (2013)
15. National Institute of Standards and Technology, Guidelines on mobile device forensics (draft), Special Publication 800–101. U.S. Department of Commerce, Gaithersburg, MD (2013)
16. S.E. Goodison, R.C. Davis, B.A. Jackson, in *Digital evidence and the U.S. Criminal Justice System: Identifying Technology and Other Needs to More Effectively Acquire and Utilize Digital Evidence*. (Santa Monica, CA: RAND Corporation, 2015)

# Efficient Sensitive File Encryption Strategy with Access Control and Integrity Auditing



D. Udhaya Mugil and S. Metilda Florence

**Abstract** Cloud storage performs a significant role in data sharing and data storage. The sharing of data is accessible over the internet. Protection has gotten to be a noticeable issue once the applying of the cloud data unit significantly developing in cloud computing in cloud computing. The benefits of the execution of these rising technologies have moved forward or changed administration models and progress application exhibitions from a different viewpoint. In any case, the exceptional developing volume of data sizes had to come about in numerous challenges in hones. The execution time of the data encryption is one of all the firm issues all through preparing and transmissions. A few current applications forsake data encryption to realize a versatile execution level companioning with protection issues. In this framework, we concentrate on efficient data protection. For that, a novel sensitive encryption approach is proposed with a sensitive file encryption strategy (SFES). Our proposed approach points to scramble the records which have more delicate data and utilize a sensitive file classification strategy. This methodology is intended to expand the security assurance scope by employing a sensitive encryption procedure that diminishes the time and distributed cloud storage space.

**Keywords** Sensitive file · Encryption process · Decision making · Team frequency · Weightage calculation

## 1 Introduction

Lately, cloud computing has been a developing innovation in IT industries. Due to the rapid increase in sharing the data in applications like iCloud, Google Drive, Dropbox, and so on. Users can exchange their data files to the cloud and share it

---

D. Udhaya Mugil (✉) · S. Metilda Florence

Department of Information Technology, SRM Institute of Science and Technology,  
Kattankulathur, India

e-mail: [Um9252@srmist.edu.in](mailto:Um9252@srmist.edu.in)

S. Metilda Florence

e-mail: [metildam@srmist.edu.in](mailto:metildam@srmist.edu.in)

with others. Be that as it may, the data stored within the cloud may be hacked and our information may spill out. To confirm whether the data is kept efficiently within the cloud, numerous further data astuteness reviewing schemes are proposed and to secure file all files which re-established in the cloud are encrypted and stored within the cloud space. This will take more execution time and storage space. Nonetheless, the shared data stored in the distributed cloud storage may contain a few touchy information.

Sensitive data has got to be protected against unauthorized admittance to safeguard the privacy or security of a private or association. So it is necessary to achieve a remote data integrity auditing scheme on the condition that the delicate data of shared information is ensured by hiding information utilizing the sanitizer process [1–3]. The sanitization will receive the recently uploaded data from the temporary storage and perform the team frequency process to calculate the weightage of the sensitive information based on the keywords in the file it decides whether to encrypt the data or not.

We concentrate on privacy and proposed a new concept sensitive file encryption strategy (SFES) with access control and integrity auditing technique for secure cloud storage administration. The main aim is to efficiently protect the file with more sensitive information stored in the cloud. For further use, the data stored within the cloud had encrypted and protected the delicate information of the data file utilizing the proposed conspire [4–6]. We design to maximize the security assurance scope by employing a sensitive file encryption procedure. Besides, our scheme on a sensitive file encryption strategy which reduces the time and cloud storage space.

## 1.1 Scope

The main outlook of the task is to secure the data file put away in the cloud storage area. If all the files put away in the cloud is encrypted, then it takes more time and space. Hence, the file is sensitive then it will encrypt the file and upload into the cloud and also the auditor user performs integrity testing on the uploaded files.

## 2 Related Work

Various research on works had as of now carried out in the area of secured cloud storage. Each of the work research provides an understanding of different reviewing planes and examination of secure cloud storage service.

Ateniese et al. [7] provides the PDP (provable data possession) model for remote data checking that allows us to confirm data ownership without having access to the genuine data file and bolsters enormous data sets in a broadly distributed storage system.

Ari Juels et al. [8] the objective of a POR (Proofs of retrievability) is to accomplish these checks without downloading the files on their own and the data file is retrievable inside specific time constrain. It achieves that users want to verify that data are not deleted or modified before retrieval.

Wang et al. [9] analysis of the matter of providing concurrent public audibility and information flow for farther knowledge principle sign up cloud computing. The two fundamental objectives are to achieve efficient data dynamics and move forward Merkle hash tree development for block tag authentication and uphold the gainful treatment of numerous evaluating assignments.

el-Khameesy and Rahman [10] an efficient and adaptable security strategy method with information backing and data alteration like block update, delete, and append. It accomplishes the integration of capacity accuracy confirmation and information flaw restriction.

Wang et al. [11] utilizes the homomorphic direct authenticator and arbitrary concealing to ensure that the TPA (third-party auditing) does not get any data file put away on the cloud. It eliminates the user's fear of their data spillage. Encourage, it amplifies the privacy-preserving public auditing protocol to play out various reviewing tasks efficiently to the cluster way.

Yu et al. [12] provides key exposure resilience with a proficient examining scheme to diminish harm in the client's key exposure. To overhaul the secret key for the client, it had binary tree structure and pre-order traversal strategy [13]. It develops a novel authenticator for forwarding security and blockless verifiability.

Worku et al. [14] alleviate the data integrity problem using the blinding technique. It provides the right for TPA to audit multiple users concurrently at the same time. This improved by minimizing intensive operations like binary mapping. It is not fully dynamic, some operations are not supported.

Luo et al. [15] allow the group together with intermediaries and convert the signature from revoked users. It progressed polynomial based verification labels make it infeasible for pernicious assaults.

Fu et al. [16] the issue of retrieving the complete data from the cloud for public verifiers is overpowered. This conspires ensures that a bunch of users can follow information changes through a binary tree and recoup the most recent data file block when the accurate data file block is harmed.

Yu et al. [17] utilize of distinct structure between the user key for overhauling and actual signing. That key overhaul to refresh the secret key utilization in one time period. It tremendously diminishes the harm of key disclosure for identity auditing based signatures.

### 3 Project Description

With the hazardous development of data file storage, it's an imperative burden for users to upload the erect entirety of information provincially. Hence, a lot of management and people might want to store their data files in the cloud storage system

[18–21]. Be that as it may, the file that contains data is stored within the cloud may well be adulterated or misplaced because of human blunder. Some information may leak out. On the user side, the computational burden is reduced, by introducing a third-party auditing scheme (TPA) to systematically confirm the astuteness of the data within the cloud for sake of the users. The previous plans all depend on public key infrastructure (PKI), it actuates impressive outlays from the arduous certificate administration.

Essentially, the data blocks of the users are comparing individual touchy information of the data file. It develops the resembling signatures for that user. To confirm the process of integrity and to guarantee the sanitized file authenticity is performed. The sanitizer receives the hidden data file and its analogous signatures from the user sides. The user sanitized these data blocks using the sanitizer process, and thus it is correlating to the administration's touchy information, it converts the sanitized data file into an exact once using the user signatures information. That the sanitizer stores this file with sensitive information along with its signatures to the storage system. It is utilized to identify the astuteness of the sanitized data file with hidden information of any integrity auditing. The auditing scheme wishes to know the astuteness of the file that is stored on the cloud storage service. It responses to the TPA process.

At last, the integrity of the TPA verifies the sensitized file whether the auditing proof checks are correct or not. The existing remote data integrity auditing schemes of all cannot reinforce sensitive information hiding with data shared on the cloud. To secure files, all files which are kept within the cloud perform an encryption process, and the encrypted data file is restored on the cloud storage system. This process will take more execution time and storage space.

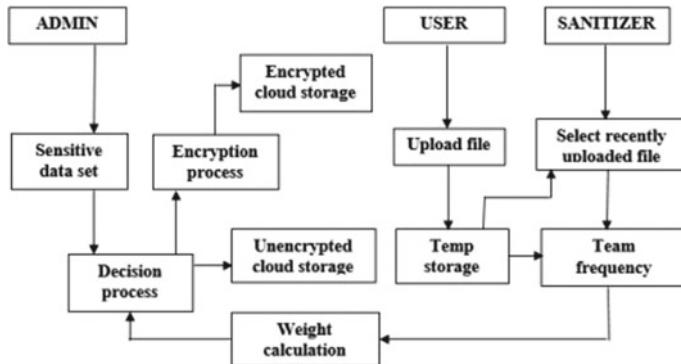
## 4 Methodology

The project's main idea is to expand the security protection extension by employing a sensitive file encryption strategy (SFES) to encrypt the files which have more sensitive data. It will reduce time and cloud space. Hence, we are proposing a novel sensitive data encryption approach, and using sensitive file classification technique it classifies the file with more sensitive data. The algorithm used is a natural language processing technique, term frequency, and sensitive file encryption algorithms, cryptography technique, and integrity test using hashing.

The SFES proposed architecture includes sorts of distinctive entities, as shown in Fig. 1.

**Cloud:** The cloud provides efficient cloud storage services. In these cloud services, user can store their data and share it with other users through the web.

**User:** The user is an individual or administrative, they can upload an enormous of the file to the cloud storage, and it stored in temporary storage for sanitizer process.



**Fig. 1** Proposed architecture process

**Sanitizer:** The sanitizer selects the recently uploaded file from the temporary cloud storage, and its process pre-processing technique to extract the keyword for weightage calculation for shortlisting the sensitive data. After the decision-making process, it checks whether to encrypt the data or store it as it is to the cloud storage.

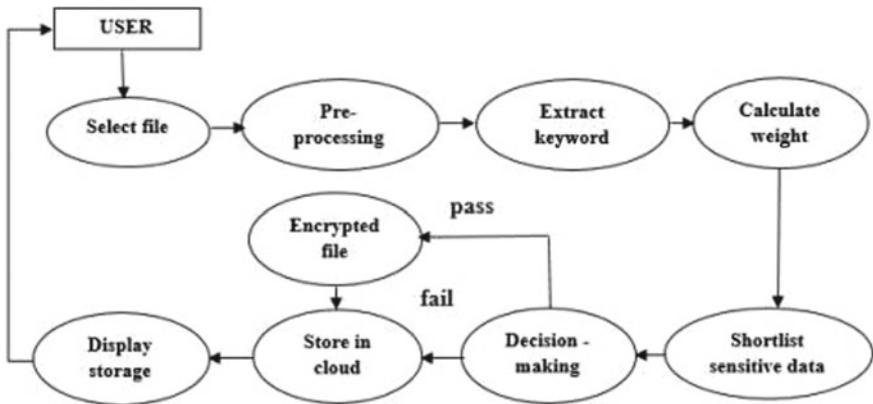
**Admin:** Admin will only access and monitor all the process. The sensitive data set is uploaded to calculate the weightage of sensitive data and process according that.

**Weightage calculation:** In this process, it calculates the weightage of the data file using the previous method term frequency. If the weightage of sensitive data reaches the given threshold value it processes to the next phase that is the decision-making process.

**Decision-making process:** This process will make the decision whether to encrypt the file or store it in the cloud. Using the term frequency weightage calculation method further makes the decision of the file stored in the cloud.

## 5 Decomposition of the System

The project main scope is to classify the data with a more sensitive file of the users and encrypt it according to the cloud by sensitive file encryption strategy (SFES) (Fig. 2).



**Fig. 2** Sanitization process

### 5.1 Data Pre-Processing

It is one of the foremost data processing steps that bargains with data planning and transformation of the dataset and seeks at an equivalent time to create knowledge disclosure more efficiently. The process of data pre-process as shown in Fig. 3.

#### 5.1.1 Special Character Removal

Unique characters and symbols are generally non-alphanumeric characters or maybe every so often numeric characters, which upload to the more noise in unstructured content. Normally, basic standard articulations can be utilized to eliminate them.

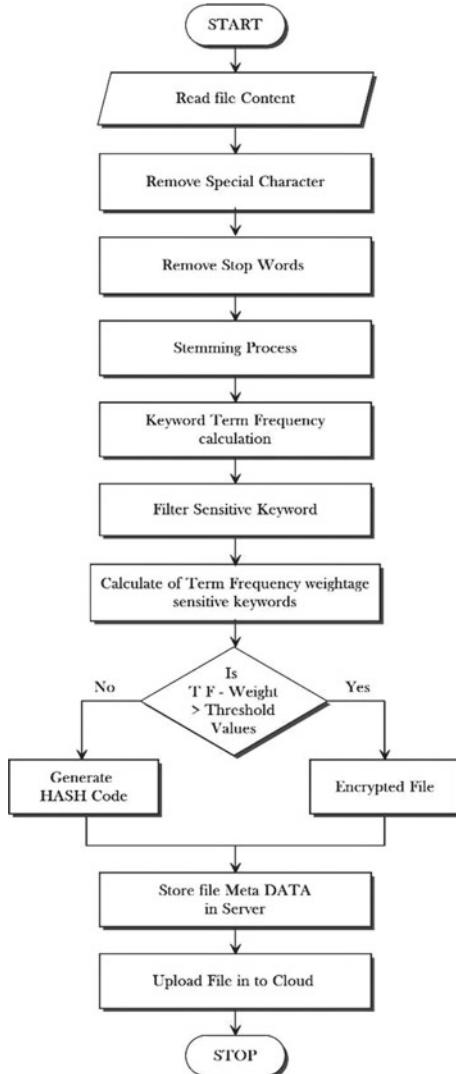
#### 5.1.2 Stop Word Removal

Stop words are a bunch of words in a language that are commonly used. Instances of stop words are a,” “are,” “the is” and so on in English. The instinct behind the use of stop words is that by removing uninformed words from text, all things are equal by tuning in on the important ones. This should be possible by keeping all words from your prevent word list from being dissected.

#### 5.1.3 Stemming Process

Stemming is the method that reduces the words repetition in the data file (for example inconvenienced, inconveniences) this can be combined as single word (for example

**Fig. 3** Flowchart of data processing method



inconvenience). This situation may not be a genuine common word, however, an accepted type of the first word.

## 5.2 Term Frequency-Inverse Document Frequency

The TF-IDF (term frequency-inverse document frequency) has two attributes weight involves term frequency and inverse document frequency. Term frequency is the

number of events of the term within the file. Document frequency alludes to the number of files that contain the term, and it denotes the number of files.

$$idft = \log(N/idft)$$

The total file number is denoted as  $N$ . The weighting process  $tf\text{-}idf$  allocate to term  $t$ , file weight is stated in  $tf - idft, f = tft, f * idft$ .

### **5.3 Cryptography Technique**

Using the cryptography technique, we are going to secure our outsourcing data. When the decision-making technique decided the file consists of data, the file should be encrypted using cryptography techniques and stored on cloud storage. We are using cryptography technique as AES (advanced encryption standard) algorithm for encryption and decryption process.

### **5.4 Integrity Checking Technique**

Integrity checking is that the method of scrutiny this state of stored data and/or programs to an antecedently recorded state to detect any changes. When the file is uploaded on the cloud storage, it should be verified to just check whether the file is corrupted or not. So integrity is used to verify a file using the hashing technique. We are using SHA (secure hash algorithm) for hashing algorithm to keep data secured.

## **6 Conclusion**

In our conspiracy, the data saved in the cloud can be shared efficiently and accessed by different users relying on the prerequisite that data file having sensitive information is encrypted. Besides, the data without sensitive information is stored as it is on the cloud. By using a sensitive file encryption strategy to the existing scheme, efficiently reduces the time and clouds storage space.

## **7 Future Enhancement**

This paper is not necessary to encrypt and store all the files on the cloud storage because it takes more execution time and storage space, hence the proposed system introduced sensitive file encryption strategy (SFES) which will classify the file

whether it is a sensitive file or not, using natural language processing techniques and if the file is sensitive then only it will encrypt the file and upload into the cloud storage else will be stored as it is. Also, this system has an auditor user who can able to cloud to conduct the integrity testing on the uploaded file and take necessary action.

## References

1. K. Ren, C. Wang, Q. Wang, Security challenges for the public cloud. *IEEE Internet Computing* (2012)
2. C. Guan, K. Ren, F. Zhang, F. Kerschbaum, J. Yu, Symmetric-key based proofs of retrievability supporting public verification (2015)
3. W. Shen, J. Yu, H. Xia, H. Zhang, X. Lu, R. Hao, Light-weight and privacy-preserving secure cloud auditing scheme for group users via the third party medium (2017)
4. S. Zhu, V. Saravanan, B. Muthu, Achieving data security and privacy across healthcare applications using cyber security mechanisms. *Electron. Libr.* **38**(5/6), 979–995 (2020). <https://doi.org/10.1108/el-07-2020-0219>
5. T.N. Nguyen, B. Liu, N.P. Nguyen, J. Chou, Cyber security of smart grid: attacks and defenses, in *ICC 2020—2020 IEEE International Conference on Communications (ICC)* (Dublin, Ireland, 2020), pp. 1–6. doi: <https://doi.org/10.1109/ICC40277.2020.9148850>
6. H. Wang, D. He, S. Tang, Identity-based proxy oriented data uploading and remote data integrity checking in public cloud. *IEEE Trans. Inf. Foren. Secur.* (2016)
7. G. Ateniese, R. Burns, R. Curtmola, J. Herring, D. Song, Provable data possession at untrusted stores, in *Proceedings of the 14th ACM Conference on Computer and Communications Security*
8. A. Juel, B.S. Kaliski Jr, Pors: proofs of retrievability for large files, in *Proceedings of the 14th ACM Conference on Computer and Communications Security*
9. Q. Wang, C. Wang, K. Ren, Enabling public auditability and data dynamics for storage security in cloud computing (2011)
10. el-Khameesy N, Rahman HA, A proposed model for enhancing data storage security in cloud computing systems
11. C. Wang, S.S.M. Chow, Q. Wang, Privacy-preserving public auditing for secure cloud storage. *IEEE Trans. Comput.* (2013)
12. J. Yu, K. Ren, C. Wang, V. Varadharajan, Enabling cloud storage auditing with key-exposure resistance. *IEEE Trans. Inf. Foren. Secur.* (2015)
13. Y. Yu, M.H. Au, G. Ateniese, X. Huang, W. Susilo, Y. Dai, G. Min, Identity-based remote data integrity checking with perfect data privacy preserving for cloud storage. *IEEE Trans. Inf. Forensics Secur.* **12**(4), 767–778 (2017)
14. S.G. Worku, C. Xu, J. Zhao, X. He, Secure and efficient privacy preserving public auditing scheme for cloud storage (2014)
15. Y. Luo, M. Xu, D. Wang, Efficient integrity auditing for shared data in the cloud with secure user revocation (2015)
16. A. Fu, S. Yu, Y. Zhang, Npp: a new privacy-aware public auditing scheme for cloud data sharing with group users (2017). doi:<https://doi.org/10.1109/TB DATA.2017.2701347>
17. J. Yua, R. Hao, H. Xia, Intrusion-resilient identity-based signatures: concrete scheme in the standard model and generic construction (2018)
18. J. Hur, D. Koo, Y. Shin, K. Kang, Secure data deduplication with dynamic ownership management in cloud storage. *IEEE Trans. Knowl. Data Eng.* **28**(11), 3113–3125 (2016)
19. Y. Li, Y. Yu, G. Min, W. Susilo, J. Ni, K.K.R. Choo, Fuzzy identity based data integrity auditing for reliable cloud storage systems. *IEEE Trans. Dependable Sec. Comput.* (2017)

20. H. Wang, Proxy provable data possession in public clouds. *IEEE Trans. Serv. Comput.* **6**(4), 551–559 (2013)
21. H. Wang, D. He, J. Yu, Z. Wang, Incentive and unconditionally anonymous identity-based public provable data possession (2016)
22. J. Yu, H. Wang, Strong key-exposure resilient auditing for secure cloud storage. *IEEE Trans. Inf. Forensics Secur.* **12**(8), 1931–1940 (2017)

# Zone Based Crop Forecast at Tamil Nadu Using Artificial Neural Network



S. Nithiya, S. Srividhya, and G. Parimala

**Abstract** Yield expectation is a significant issue in horticulture. This research work shows the tendency of artificial neural network technology to be utilized for the crop forecast season based on Rabi, Kharif, Summer. Government is keen on knowing how much yield is going to anticipate on district wise. Previously, yield expectation was performed by survey report and manual calculation of field workers. The yield forecast is a significant issue that remaining parts to be understood dependent on reliable dataset. Artificial neural network is one of the forecasting algorithms in recent trends. Main contribution of this classification algorithms is used for obtain the better result. This research is discussed with 5 years data of various district in Tamil Nadu from 2008 to 2013. This research paper is used to mainly concentrate on yield prediction of entire Tamil Nadu state and also district wise yield prediction among major crops of Tamil Nadu. The experimental results show that artificial neural network has the higher accuracy rate.

**Keywords** Weight · Bias · Rabi · Kharif

## 1 Introduction

As an exhaustive issue, food security is made out of a few perspectives, including creating adequate food and penning food gracefully in the market, and so on. As the establishment of figuring approaches to guarantee food security, the forecast of harvest yield, which is pivotal to ensure the well-balance among flexibly and request, likewise raises a lot of consideration [1]. One of the second highest populations in

---

S. Nithiya (✉) · S. Srividhya · G. Parimala  
SRMIST, Kattankulathur, India  
e-mail: [nithiyas@srmist.edu.in](mailto:nithiyas@srmist.edu.in)

S. Srividhya  
e-mail: [srividhs1@srmist.edu.in](mailto:srividhs1@srmist.edu.in)

G. Parimala  
e-mail: [parimalg@srmist.edu.in](mailto:parimalg@srmist.edu.in)

the world is India. India is recognized as the second biggest populated nation in the Universe. As much of growth in the population requires the food demand in the field of agriculture. Food scarcity as major good security thread in the world it is very fundamental to improve the food creation. So as to satisfy the developing needs of populace, agribusiness must tussle to predict the yield production by different weather season and area wise crop production. India has endured extreme dry droughts in 1965 and 1966 [2]. The Green insurgency occurred in 1976 has recuperated India from these dry droughts and made India as independent in the food creation [3]. In any case, presently, the food creation rate is required to increment in four overlap with the current possibilities. Subsequently, it is basic to build the yield rate with the current assets of water and land [4].

This research work is carried out by 31 districts of Tamil Nadu used to find out crop prediction based on different season. The major crops of Tamil Nadu are Rice, Sugarcane, Maize, Ragi, Tapioca, cashew, groundnut identified by their yield rate with respect to their production and cultivated area.

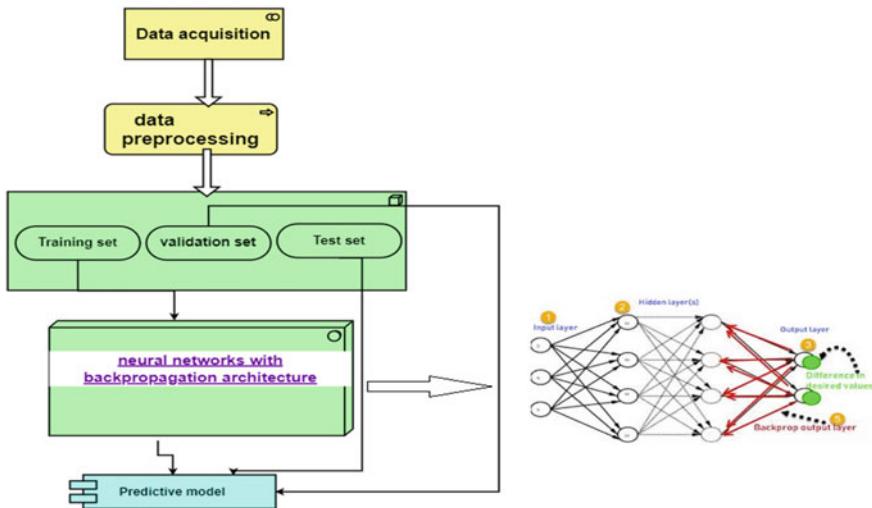
## 2 Dataset and Preprocessing

Dataset is collected for this paper is from data.gov.in which is under the administration of government of India contains all agriculture related datasets. This dataset is collected over the year of 1997 to 2013. In my research work, I am using it from 2007 to 2013. This dataset consists of crop related information's which consist of crop yielding district, year, season, cultivated area(ha) and crop production (tones). Season wise, Tamil Nadu state crop information is used for crop prediction. Crop is cultivated in three ways known as Rabi season, Kharif season, whole year [5–7]. Usually, the kharif time cultivating season is from July to October during the south-west storm, and the Rabi season for cultivation is from October to March (winter).

## 3 Preprocessing

Figure 1 showing the various types of data categorized into numerical and categorical data. It is easy to train the model when using numerical dataset. Pre-processing is the first stage in the training phase. In pre-processing stage, missing values are filled by zeros are by some min and max values [8]. Present study data's are pre-processed by converting categorical data into numerical data.

Figure 1 gives the detail about conversion of categorical data into numerical data. In this research work, State, District, Crop information are in categorical type which is converted into numerical type. E.g., Rice—numerical type is 0.



**Fig. 1** Pre-processed dataset

## 4 System Architecture

Figure 2 shows that various steps in neural network model which starts with data acquisition, pre-processing of data, transforming of data by splitting into training and test model. Predict the expected output of the model.

### 4.1 Data Acquisition

Look into the available and reliable data. Select the needed data and remove the missing and irrelevant datas.

	STATE	DIS	YEAR	SES	CROP	AREA	PRO
0	0	0	2009	0	16	25978	80462.0
1	0	0	2009	0	18	404	649.0
2	0	0	2009	2	0	160	122.0
3	0	0	2009	2	1	644	1082.0
4	0	0	2009	2	2	134	5761.0
..	...	...	...	...	...	...	...
102	0	0	2013	2	6	14	5.0
103	0	0	2013	2	8	212	58.0
104	0	0	2013	2	18	187	491.0
105	0	0	2013	2	19	9875	1112304.0
106	0	0	2013	2	22	189	5655.0

**Fig. 2** System architecture

## 4.2 Data Pre-Processing

Pre-processing of dataset step includes formatting of data to appropriate file format next step is cleaning of data by removal of missing values and final step is sampling of data.

## 4.3 Transforming Data

Scaling, decomposition and aggregation are some of the data transforming techniques after the pre-processing step.

# 5 Proposed Method

## 5.1 Neural Network

Neural network is the simplified network model which is like neuron in the animal brain. This network model consists of input layer, hidden layer, output layer.

*Input layer:* It brings with the initialized input feature to next level of layer.

*Hidden layer:* This layer is placed in between of input layers and output layers. This consists of neuron which is used to set weighted input and produce the output based on the activation functions. Weight is assigned to each hidden layer of the neuron.

*Output layer:* the last layer is output layer which produces the expected output. When the input layer transfers the input feature to the next layer the input feature is multiplied with weight value.

$$y = \sum (\text{Weight} * \text{input}) + \text{bias} \quad (1)$$

In Eq. (1), neuron is the basic unit of a network. It receives certain number of input and bias with default value 1. When the input arrives, it is multiplied by a weight. Suppose if the neuron consists of 4 input then it has 4 weight values. The weight value will adjust during the training time

$$z = x_1 * w_1 + x_2 * w_2 + x_3 * w_3 + x_4 * w_4 + b * 1 \quad (2)$$

$$\hat{Y} = \text{aout} = \text{Sigmoid}(Z) \quad (3)$$

$$\text{Sigmoid}(Z) = \frac{1}{1 + e} - Z \quad (4)$$

Forward propagation is a procedure of taking care of info esteems to the neural system and getting an output which, we call anticipated worth [9]. At the point when we feed the information esteems to the neural system's first layer, it abandons any tasks. Second layer takes esteems from first layer and applies multiplication, addition and activation activities and passes this incentive to the following layer. Same procedure rehashes for ensuing layers lastly, we get a predicting value from the last layer.

## 6 Neural Network Processing Steps

**Step1:** We have our input feature in first 6 column State, district, crop name, area, production year and season. Last 6 column consist of predictive feature of major crops in Tamil Nadu.

**Step2:** Splitting the dataset into input feature ( $x$ ) and predict feature as( $y$ ).

**Step 3:** Scaling the input feature between 0's and 1's using min\_max scaler function.

**Step 4:** Splitting the dataset for training set and test set.

Xtrain have 6 input features and 70% of data.

Xval have 6 input features and 15% of data.

Xtest have 6 input features and 15% of data.

Ytrain have 1 output features and 70% of data.

Y\_val have1 output feature and 15% of data.

Y\_test have1 output feature and 15% of data.

**Step 5:** Building the model: It consists of sequential model with 2 hidden layers with 36 neuron and ReLU as the activation function and output layer consist of sigmoid as the activation function.

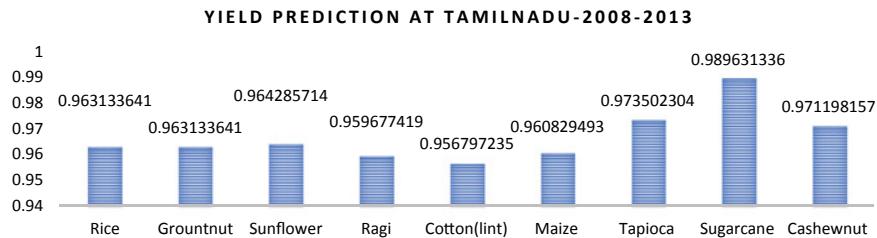
**Step 6:** Adam optimizer used for optimization algorithm for the network model. In loss function, we are using binary cross entropy function.

**Step 7:** Fitting the parameter as xtrain and ytrain into fit model. After that we are using batch size as 32 and epoch is 100.

**Step 8:** After 100 iteration, we can evaluate loss and accuracy of a model by plotting the graph.

## 7 Implementation

This Fig. 3 shows that maximum yield prediction at Tamil Nadu in all season. This research work was done by 31 districts includes all the season in the Tamil Nadu. The major objective of this figure to identify the major yielding crop in Tamil Nadu [10]



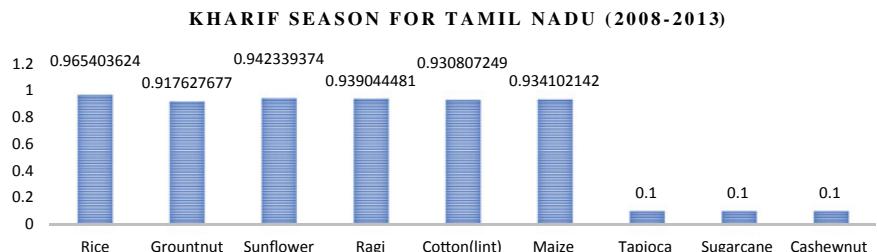
**Fig. 3** Yield at Tamil Nadu

in the year of 2008 to 2013. The highest yield at Tamil Nadu is Sugarcane, Tapioca and cashew nut.

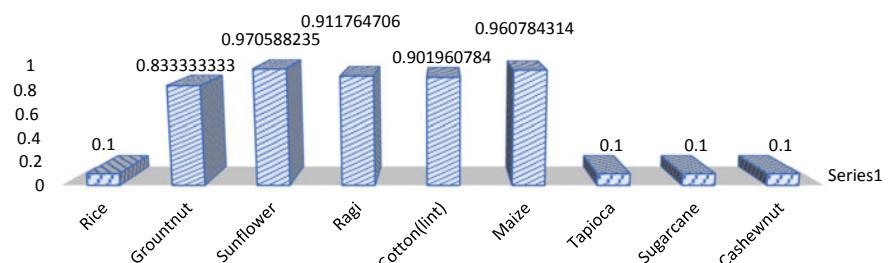
Figure 4 shows that yield in kharif season at Tamil Nadu. Kharif season start with rainy season from June to November. Major cultivation at this season is rice and sunflower. Crops which need more water is prefer to cultivate in this time.

Figure 5 shows Rabi season yield prediction. Rabi season started from winter season, and it is harvested in the spring. In Rabi season, short duration crops are cultivated. Cultivation and harvest occurred in October and November month duration. Major crops which is identified in this season is sunflower.

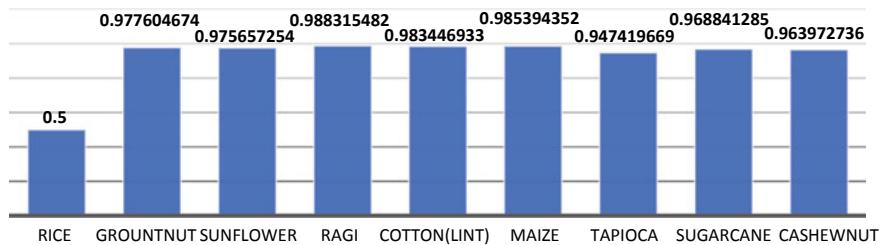
Figure 6 shows Summer Season crops. Crops are cultivated and harvest in January and February duration. Highest yield at Summer ragi, cotton. The reason for highest



**Fig. 4** Kharif season yield at Tamil Nadu



**Fig. 5** Rabi season for Tamil Nadu 2008–2013

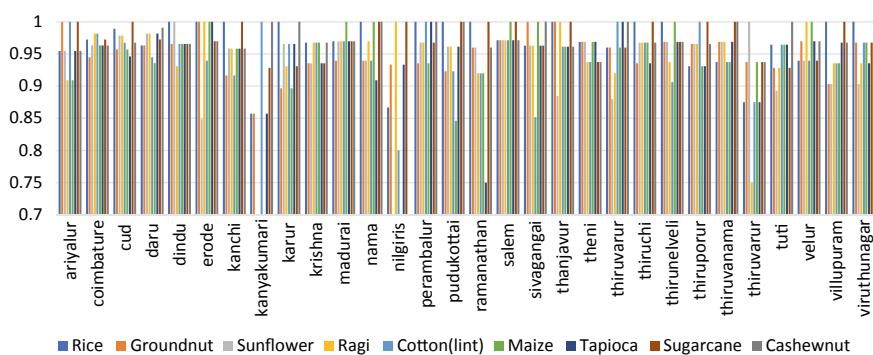
**Fig. 6** Summer season crops**Table 1** Major crop in whole Tamil Nadu

Season	Years	Major crop
All district	2008–2013	Sugarcane
Kharif season	2008–2013	Rice
Rabi	2008–2013	Maize
Summer	2008–2013	Ragi

yielding of this crop is less amount water needed for it. Highest yielding crop at rabi season is Mazie and Sunflower season cultivation is in Table 1 [11–13]. It gives the conclusion for 31 district crop cultivation. Major crop in Tamil Nadu is sugarcane. Many districts prefer to cultivate rice at kharif season and maize at rabi season and ragi in summer season.

Crop priority is varied from district among Tamil Nadu. This research work done among 30 district which is mention in Fig. 7.

Crop is segmented as major crop, medium crop and low crop based on the prediction accuracy. This information is used to clearly understand Fig. 7.

**Fig. 7** District wise crop prediction

## 8 Conclusion

Neural network is one of the best models to get highest accuracy of prediction output. In this research, focused on past year data which is used to focus zone wise major crop prediction among 9 types of crops. In future, recommendation and neural network-based crop yield simulation model will provide better solution for former to make the estimation in their own cultivation land.

## References

1. N. Gandhi, L.J. Armstrong, Assessing impact of seasonal rainfall on rice crop yield of Rajasthan, India using association rule mining, in *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)* (IEEE, 2016)
2. S. Thenmozhi, D. Uma, K. Vikram, Quantifying Yield Gap of Rice Production in Various Regions of Karnataka, in *2016 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM)* (IEEE, 2016)
3. B. Garg, S. Aggarwal, J. Sokhal, Crop yield forecasting using fuzzy logic and regression model. *Comput. Electr. Eng.* **67**, 383–403 (2018)
4. <http://www.slbctn.com/uploads/Agriculture.pdf>
5. M. Anbarasan, B. Muthu, C.B. Sivaparthipan, R. Sundarasekar, S. Kadry, S. Krishnamoorthy, A.A. Dasel, Detection of flood disaster system based on IoT, big data and convolutional deep neural network. *Comput. Commun.* **150**, 150–157 (2020). doi:<https://doi.org/10.1016/j.comcom.2019.11.022>
6. T.N. Nguyen, B.-H. Liu, S.-Y. Wang, On new approaches of maximum weighted target coverage and sensor connectivity: hardness and approximation. *IEEE Trans. Netw. Sci. Eng.* **7**(3), 1736–1751 (2020). doi:<https://doi.org/10.1109/TNSE.2019.2952369>
7. [http://www.arthapedia.in/index.php%3Ftitle%3DCropping\\_seasons\\_of\\_India-\\_Kharif\\_%2526\\_Rabi](http://www.arthapedia.in/index.php%3Ftitle%3DCropping_seasons_of_India-_Kharif_%2526_Rabi)
8. Y.E. Lu, et al., Vegetable price prediction based on pso-bp neural network, in *2015 8th International Conference on Intelligent Computation Technology and Automation (ICICTA)* (IEEE, 2015)
9. G. Camps-Valls, L. Gomez-Chova, J. Calpe-Maravilla, E. Soria-Olivas, J.D. Martin-Guerrero, J. Moreno, Support vector machines for crop classification using hyper spectral data. *Lect Notes Comp. Sci.* **2652**, 134–141 (2003)
10. A. Suresh, P. Ganesh Kumar, M. Ramalatha, Prediction of major crop yields of Tamil Nadu using K-means and Modified KNN, in *2018 3rd International Conference on Communication and Electronics Systems (ICCES)* (IEEE, 2018)
11. G.R. Batts, Effects of CO<sub>2</sub> and temperature on growth and yield of crops of winter wheat over four seasons. *Euro. J. Agron.* **7**, 43–52 (1997)
12. G. Ruß, Data mining of agricultural yield data: a comparison of regression models, in *Conference Proceedings, Advances in Data Mining—Applications and Theoretical Aspects*, P Perner ed. by, Lecture Notes in Artificial Intelligence 6171, (Berlin, Heidelberg, Springer, 2009), pp. 24–37
13. H. Jorquera, R. Perez, A. Cipriano, G. Acuna, Short term forecasting of air pollution episodes, in *Environmental modelling*, ed. by P. Zannetti (WIT Press, UK, 2001)

# Secure Cloud Data Storage and Retrieval System Using Regenerating Code



S. Yuvaraman and D. Saveetha

**Abstract** Since technology started evolving people started using all the latest technology to make them updated and use the technology for their day-to-day life usage. In the exponential growth of technology, cloud computing plays an important role all over the world to store user's data in it. Cloud storage provides higher storage space for less amount and can access it from anywhere, whenever you want to access the resources. In this paper, they have used a three-layer storage architecture to protect the data from unauthorized users and make the data available to the user's all the time without any modification in it. The three-layer architecture includes the user local machine, fog layer, and the cloud layer. We propose a system where it provides confidentiality, integrity, availability of the user's data and we are implementing new techniques like XOR-Combination to fragment and defragment the data and for creating regenerating blocks to retrieve the data successfully, hashing is used to check the integrity of the data while downloading it from the cloud server, block management is used to store the fragmented blocks in the different cloud server to make it secure and keep away from attackers.

**Keywords** Data block retrieval · XOR · FMSR · Block management · Hash

## 1 Introduction

One of the drastic or exponential growth in evolution is the cloud, which helps the users to store their data for lower cost and available all the time to access it from anywhere and whenever you want to access the resources. As the day passes, the data generated by the individuals increase, the data created by each and every individual for a second is 1.7 MB in 2020 and the data created in a day is roughly about 1.145

---

S. Yuvaraman (✉) · D. Saveetha

Department of Information Technology, SRM Institute of Science and Technology,  
Kattankulathur, Chennai, Tamil Nadu, India

e-mail: [ys1808@srmist.edu.in](mailto:ys1808@srmist.edu.in)

D. Saveetha

e-mail: [saveethd@srmist.edu.in](mailto:saveethd@srmist.edu.in)

trillion MB, which is extensively too high [1]. So, all the people around the world started migrating to cloud service providers as they provide sufficient storage for a lesser amount when compared to the physical storage devices which cost a bit high when compared to cloud storage [2]. As cloud storage increased day-by-day threats and attacks also increases to steal the data which are sensitive and confidential. To protect the data from the attackers, each and every organization tries to provide a security measure that will help the organization to stay away from the attackers [3]. Data storage in the cloud uses various techniques to protect it from unauthorized users, which includes cryptographic algorithms, and so on. The data in this project is going through a few steps before storing it in the cloud which has hashing, encryption, XOR-operation, and block regeneration which is an important part of this project to provide efficient data block retrieval.

## 2 Contribution

To store data in a secure environment to provide confidentiality, integrity, and availability with efficient data block retrieval from the cloud storage. Retrieval of the data blocks is more efficient than the traditional retrieval cloud storage system. Where it splits the entire data set to the cloud into four blocks  $A, B, C, D$ , respectively, and four cloud servers  $C_1, C_2, C_3, C_4$  correspondingly [4]. Where blocks  $A$  and  $B$  are stored in  $C_1$ , blocks  $C$  and  $D$  are stored in  $C_2$  ( $C_1$  and  $C_2$  are primary cloud servers). The data block stored in cloud servers  $C_1$  and  $C_2$  are XOR with each other creates a new data block to be stored in  $C_3$  and does the same process for storing the data in  $C_4$  ( $C_3$  and  $C_4$  are secondary cloud servers). When one or another node fails the remaining nodes will be able to retrieve the data from cloud storage to the user [5]. The file uploaded by the user will be split into four blocks and stored in primary nodes  $C_1$  and  $C_2$  and the remaining nodes  $C_3, C_4$  are the secondary nodes that help the retrieval process when the primary node fails. Secondary nodes are the ones that store the data in the combination of XOR format where it provides an efficient retrieval process to cloud storage. This can also be known as regenerating code, where it uses XOR to regenerate the primary block which is stored in the primary storage, and provides another four blocks of data which have to be stored in the secondary storage called the regenerated blocks or codes [6–8]. These blocks can withstand up to dual-block failure and are able to retrieve the data user requested.

## 3 Scope

The main scope of this project is to protect the data stored in the cloud from unauthorized access, modification, and destruction caused by the attackers or due to some technical errors with efficient data block retrieval [9–12]. It moves toward a scheme that undertakes preventive activities to store data in a twisted format and retrieve the

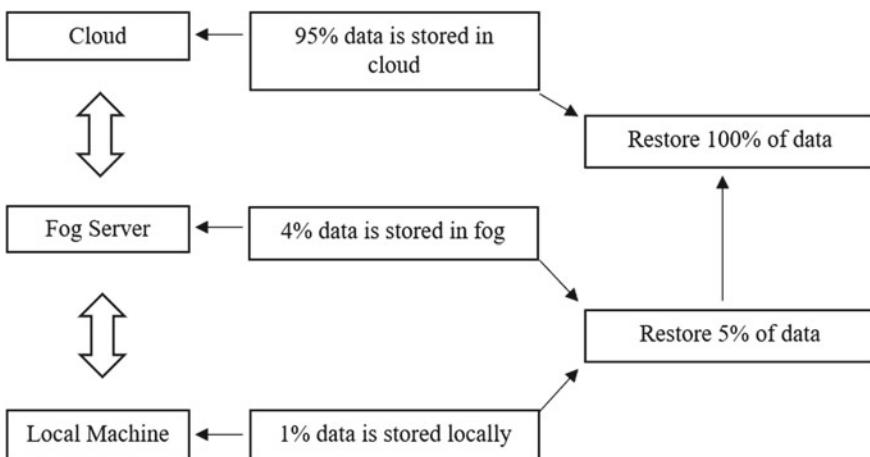
data efficiently from the cloud. Hence, using an efficient block retrieval storage, it is possible that this project can provide retrieve of data more efficiently even if some blocks of data are missing.

## 4 Existing System

The existing uses a three-layer storage framework to protect the user data based on fog computing, where the data is separated into three different parts and stored in three different layers which include local machine, fog, and cloud [13–16]. Data is divided into three parts using the Hash-Solomon code algorithm. A small part of the data is stored in the local machine and fog server to ensure the privacy of data. Cryptographic encryption algorithms are not used rather encoding, and decoding concept is used to fragment and store the data in different layers [17–20] (Fig. 1).

### 4.1 Drawbacks of Existing System

- If any one layer is compromised, the data cannot be retrieved.
- Security mechanisms like encryption and hashing algorithm are not used.



**Fig. 1** Architecture of existing system

**Table 1** Blocks stored in cloud servers

$C_1$	$C_2$	$C_3$	$C_4$
$A$	$C$	$A + C$	$A + D$
$B$	$D$	$B + D$	$(B + D) + C$

## 5 Proposed System

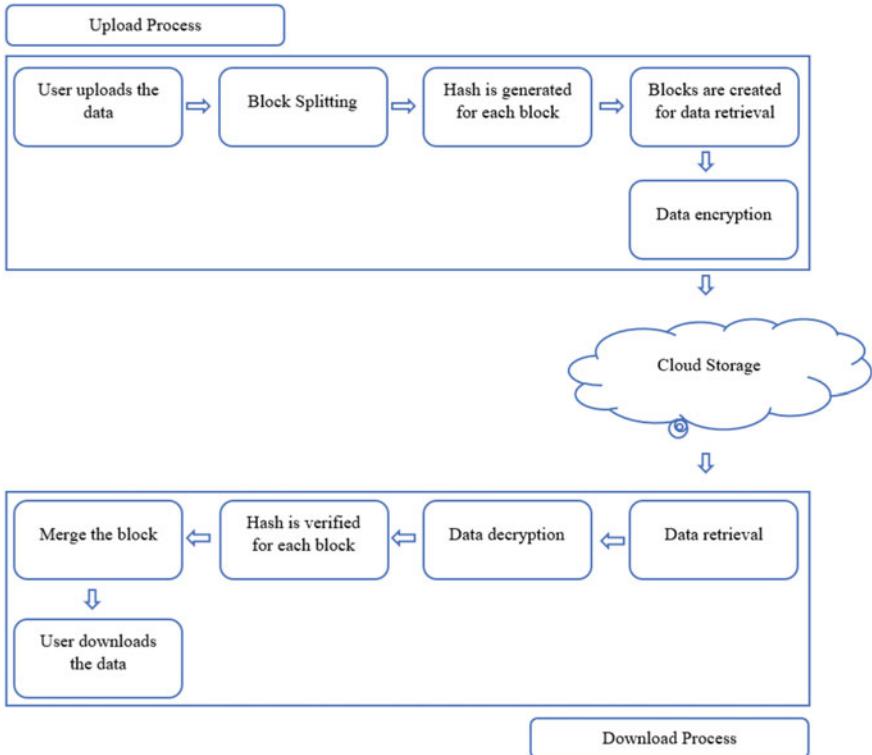
Proposed system provides a block regeneration technique that helps in block retrieval and protects the data from attackers or hackers. In this, we have block fragmentation and de-fragmentation, encryption and decryption, hashing, and XOR-operation to store the data in the cloud directly instead of using the fog layer which is used in the existing system. Proposed technique for data block retrieval uses FMSR to retrieve the data blocks efficiently. Where blocks  $A$  and  $B$  are stored in  $C_1$ , blocks  $C$  and  $D$  are stored in  $C_2$  ( $C_1$  and  $C_2$  are primary cloud servers). The data block stored in cloud servers  $C_1$  and  $C_2$  are XOR with each other creates a new data block to be stored in  $C_3$  and does the same process for storing the data in  $C_4$  ( $C_3$  and  $C_4$  are secondary cloud servers). When one or another node fails the remaining nodes will be able to retrieve the data from cloud storage to the user. The file uploaded by the user will be split into four blocks and stored in primary nodes  $C_1$  and  $C_2$  and the remaining nodes  $C_3$ ,  $C_4$  are the secondary nodes that help the retrieval process when the primary node fails.

### 5.1 Advantages of Proposed System

- Maintain privacy, prevent data loss, and modification detection.
- Efficient data block retrieval (Table 1).

### 5.2 System Design

- This is the system design where the data is uploaded by the user and it gets split, and hash is generated for all the split blocks.
- Blocks are regenerated again and creates another set of blocks to make sure the data is available for the user when on server failures.
- Finally, the all the blocks created and regenerated are encrypted using AES algorithm, and the data is stored in different cloud servers.
- When the user needs his data, the user will request the cloud for relevant data, then cloud searches for the relevant data blocks in the primary cloud servers.
- If the relevant data is found in the cloud server the cloud fetches the data and proceeds for data decryption and then the hash value is verified for each and every block to be retrieved.



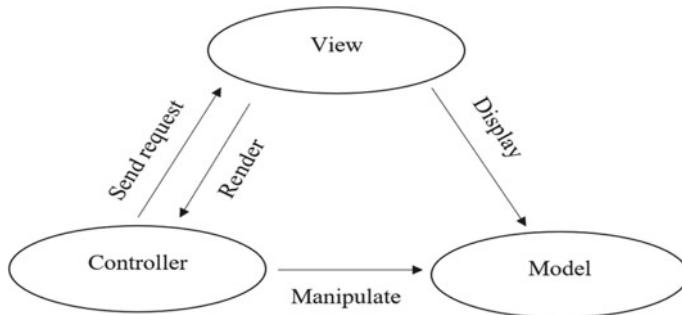
**Fig. 2** Dataflow architecture of proposed system

- If the hash value successfully matches then the data is merged and it will be available for the user to download (Fig. 2).

### 5.3 MVC Architecture

MVC or Model View Controller Architecture is a design pattern for web application development.

- **Model**—This is the layer where data handling is done.
- **View**—This is responsible for displaying the data to the user in the way of user interface.
- **Controller**—It handles the request and response which is taking place (Fig. 3).



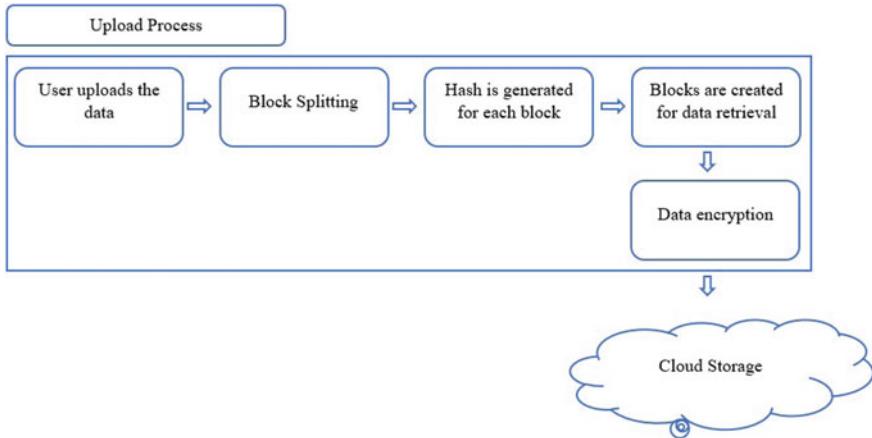
**Fig. 3** MVC architecture

#### 5.4 *Upload Procedure*

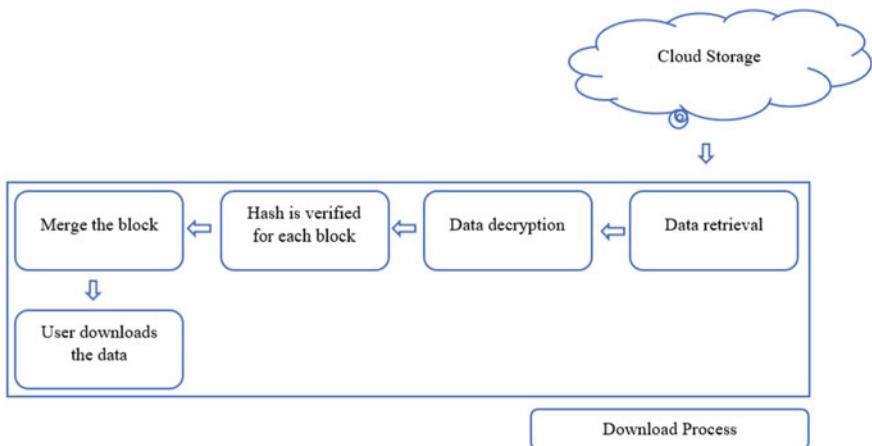
- User selects the file to be uploaded and stored in cloud storage.
- Before storing the file into the cloud, the data uploaded is split into blocks using XOR operation.
- All the blocks are generated with the hash value which is used for integrity checking.
- The hash value generated for each and every block is stored in the hash table for verifying when the data download takes places.
- After the hash value generation, the block which were split are regenerated using the same XOR-operation which was used to split the blocks in the beginning process.
- The blocks are regenerated in the process of combining two blocks of data using XOR, and the regenerated block is used for secondary storage.
- Finally, after the creation of regenerating blocks, encryption process is taken place and all the data blocks are encrypted.
- Encrypt all the blocks and store it in the cloud server using block management technique (Fig. 4).

#### 5.5 *Download Procedure*

- User selects the file which he has to download, and makes the request to cloud server to download and view the needed data.
- Cloud server receives the request and searches for the relevant blocks in the primary cloud storage. If the relevant data is found in the primary storage and if the primary is active then, download all the relevant blocks.
- Then, decrypt the blocks which has the relevant information. Once decryption is done, then verify the hash value of the blocks using the already stored hash.

**Fig. 4** Data upload process

- If the hash value remains the same and verified successfully then merge all the data blocks using XOR-combination and now user will be able to download and view his data.
- If the primary cloud storage servers are not active.
- Then, go for efficient data block retrieval which uses FMSR (Fig. 5).

**Fig. 5** Data Download Process

**Table 2** Efficient data block retrieval

Cloud 1	Cloud 2	Cloud 3	Cloud 4	A	B	C	D
T	F	T	F	A	B	$(A + C) + A$	$(B + D) + B$
F	T	T	F	$(A + C) + A$	$(B + D) + D$	C	D
T	F	F	T	A	B	$(B + D) [(B + D) + C]$	$(A + D) + A$
F	F	T	T	$(A + C) + C$	$(B + D) + D$	$(B + D) + [(B + D) + C]$	$(A + D) + A$

### 5.6 Efficient Data Block Retrieval

The important part for data retrieval is the efficient data block retrieval or regenerating code, where the data split into blocks will be able to be retrieved from the cloud storage even if any issue is caused and data is lost. With the help of primary and secondary storage blocks, it is made that can be retrieved even if two cloud storage fails due to issue (Table 2).

## 6 Conclusion

From this, the proposed work will provide confidentiality, integrity, availability, and an efficient way to retrieve the data blocks from the cloud storage. Thus, this technique will be more useful and effective to retrieve the data and stores the data in different cloud storage to protect it from unauthorized users.

## References

1. V.R.R. Atukuri, A novel approach: reliable and secure data storage and retrieval in a cloud (2017)
2. A. Ghani, Cloud storage architecture: research challenges and opportunities (2020)
3. V. Swathy FMSR-AES in multiple cloud storage system to provide secured fault tolerant (2015)
4. M.K. Saravana, Data integrity protection scheme for minimum storage regenerating codes in multiple cloud environment (2016)
5. Z. Diao, Study on data security policy based on cloud storage (2017)
6. R. Suriya, Empower data purity protection using FMSR code (2015)
7. W. Delishiya Moral, Improve the data retrieval time and security through fragmentation and replication in the cloud (2016)
8. L. Hu, N.-T. Nguyen, W. Tao, M.C. Leu, X.F. Liu, Md. Rakib Shahriar, S.M. Nahian Al Sunny, Modeling of cloud-based digital twins for smart manufacturing with MT connect, Procedia Manufact. **26**, 1193–1203 (2018). ISSN 2351–9789. <https://doi.org/10.1016/j.promfg.2018.07.155>

9. M. Balaanand, N. Karthikeyan, S. Karthik, Envisioning social media information for big data using big vision schemes in wireless environment. *Wirel. Pers. Commun.* **109**(2), 777–796 (2019). <https://doi.org/10.1007/s11277-019-06590-w>
10. Y. Yang, Efficient regular language search for secure cloud storage (2018)
11. Q.-A. Zeng, Secure and efficient product information retrieval in cloud computing (2017)
12. Z. Li, Secure cloud storage system based on ciphertext retrieval (2020)
13. R. Chen, Secure data storage and retrieval in cloud computing in cloud computing (2016)
14. S. Talwani, Comparison of various fault tolerance techniques for scientific workflows in cloud computing (2019)
15. S. Alraddadi Safaa Yousef, Enhancing an availability of cloud-based fault tolerance techniques (2018)
16. I. Chana, Fault tolerance techniques for scientific application in cloud (2017)
17. S.M. Dilip Kumar, Fault-tolerant scheduling for scientific workflows in cloud environments (2017)
18. N.L. Kodumru, Secure data storage in cloud using cryptographic algorithms (2018)
19. S. Prathiba, Survey of failures and fault tolerance in cloud (2017)
20. R. Arora, Secure user data in cloud computing using encryption algorithms (2013)

# Visual Question Answering System for Relational Networks



M. Fathima Najiya and N. Arivazhagan

**Abstract** Visual question answering (VQA) is a multi-modal research area combining deep learning, natural language processing, and computer vision. It is a multi-modal or cross-media problem where the computer system is built to answer the questions about an input image. The answers may be open-ended or multiple choice. For descriptive answers, the structure of the answer sentence generated is also verified for correctness. This project proposes to model a VQA system for relation network. The major goal of the relation network is to explore the spatial or logical relation among objects present in an image through the questions. It mainly concentrates on answering questions where the position or query is placed through relation to another object or region within the image. The proposed model aims on exploring the object-to-object-based or feature-to-feature-based relations in an image. The main applications of VQA include assistance for visually impaired people, enhancing e-learning capabilities, and improving image retrieval systems.

**Keywords** Image understanding · Visual question answering · Natural language processing · Relational networks

## 1 Introduction

Visual question answering systems combining natural language (NL) processing, computer vision, and image understanding are one of the challenging and complicated research areas in artificial intelligence. Since all VQA tasks require visual understanding, and semantic reasoning, it is also viewed as a key progress measure for machine intelligence. With the availability of large-scale standard datasets and swift

---

M. Fathima Najiya (✉) · N. Arivazhagan

Department of Information Technology, SRM University Chennai, Chennai, India

e-mail: [fn9326@srmist.edu.in](mailto:fn9326@srmist.edu.in)

N. Arivazhagan

e-mail: [arivazhn@srmist.edu.in](mailto:arivazhn@srmist.edu.in)

development in deep learning techniques, the research area has been of tremendous interest to machine learning experts.

A typical VQA problem has three basic components: feature extraction from the input text and visual scene, feature fusion between the modalities, and employing the joint representation of features for answer prediction. The main goal of single modality feature extraction is to obtain the most relevant information from the original input data and present it in a lower dimensionality space for ease of further processing. Feature fusion involves combining the features from multiple modalities (visual and textual) using different techniques. It plays a crucial part in the accuracy of the model developed.

Majority of the existing benchmark VQA datasets contain triplets consisting of an image, a natural language text question based on the image, and the correct answer. Some also provide extra information like image regions, image captions, or multiple-choice candidate answers. The VQA datasets presently available can be classified depending on three major components: question–answer format, type of images available, and the use of external knowledge. There are mainly three types of images: synthetic, clip-art, and natural images. The questions may be arbitrary and can contain subproblems in computer vision. Some of the common question types include object detection, object recognition, attribute classification, scene classification, and counting. The answer types may be yes/no, count, multiple-choice, one word (rarely two or three words), or descriptive type. However, there are a very few number of systems providing descriptive type answers.

## 2 Related Works

### 2.1 Literature Study

The paper [1] proposes to give answers to questions based on images by combining the input of natural language query with the output of visual scene analysis in a probabilistic structure using the Bayesian formulation. It follows a multi-world approach where many hidden ‘worlds’ and different interpretations of the scene are taken into account to answer the questions. The model has achieved several key parameters like it can handle human questions of high complexity and it is able to return answers with object count and class. The research also introduced a new testing mechanism WUPS for VQA. However, the model assumes that the facts in question–answer sets used for training the system are correct; this may lead to incorrect learning.

The paper [1] proposes using visual-semantic embeddings and neural networks, without the usual intermediary stages like object detection from image and segmentation of image, for predicting the answers to simple questions about images. The paper provides two major results—the model uses convolutional neural network (CNN) for image processing and long short-term memory (LSTM) as visual-semantic embeddings technique. The proposed model presented as LSTM\_VIS treats the image as a

word in the question. The paper also proposes synthesizing QA pairs from currently available image descriptions which helps in creating more datasets with an easily automatable question generation algorithm.

However, the implementation is limited to single word answers for classifier questions within a limited domain.

The paper [2] proposes the task of open-ended visual question answering. The paper classifies between different type of questions and answers possible and the computational algorithm for each. The study also found that answering accuracy of a model depends on the type and the length of question, and answer type can be mostly determined by the type of question-and-answer length which is majorly one word followed by a minute fraction of two- and three-word answers. The study also shows that a multiple-choice question answering task needs only an algorithm which will opt for the correct answer from a preset list of possible answers, whereas an open-answer problem mostly needs a free-form acknowledgement and hence employs a large set of dormant AI capabilities. It also states how in some cases, a prior knowledge about the objects in the image might be required for correct answering.

Most of the researches in VQA systems states about the bias present in most of the benchmark datasets available. The paper [3] proposes to unbias datasets for exploiting the image information for improving scores. It focuses more on the information from the image rather than relying on prior knowledge and question formulation. This paper addresses binary VQA on abstract scenes in three steps: tuple extraction, tuple and object alignment, and question answering. The initial parsing step condenses a yes/no question into a tuple of the form  $\langle P, R, S \rangle$ , where P points to the primary object, S to the secondary object, and R to relation. The extracted  $\langle PRS \rangle$  tuples are aligned to the objects in the image, and the score is calculated. The score function measures the compatibility between the text and the image features. However, this study restricts the model to using only clip-art images and the performance drops drastically with real scene images.

The paper [4] proposes reasoning over the natural language (NL) queries and visual images. Because of the NL question related to the image, the model iteratively updates the question representation after each time the question interacts with the image content. It has the ability to concentrate on image areas applicable to the questions. The model consists of multiple reasoning layers, exploiting the complex visual relations within VQA task. The suggested network is an end-to-end trainable model, by back-propagation, where the weights are initialized employing a pre-trained CNN and gated recurrent unit (GRU). The model contains a pooling step using an attention mechanism to understand the relevance between the representation of original questions and updated ones.

The network has four main components—image understanding, question encoding, reasoning, and answering layers. The image understanding layer is designed to model image content into semantic vectors. Question encoding layer encodes the NL question using recurrent neural network (RNN). The reasoning layer has question–image interaction and weighted pooling characteristics. However, the counting ability of the model is substantially weak.

The paper [5] presents stacked attention networks (SANs) for image QA. SAN utilizes the semantic depiction of an input question as an enquiry to examine for the areas within a visual scene that are relevant for the answer. SAN uses multi-layer attention mechanism to query the input image multiple times which then detects the appropriate image region and infer the answer progressively. The three major components of SAN are the image model which uses CNN, the question model which employs both CNN and LSTM, and the stacked attention model which generates vectors consisting of question and visual information with higher weight to visual regions relevant to the question. This paper proposes models with either one or two attention layers.

Artificial neural networks (ANN) are an effort to emulate the human brain actions. Attention mechanism is an ANN technique which attempts the same by selectively concentrating on the relevant things, while ignoring others. The paper [6] proposes a hierarchical co-attention model for VQA. Co-attention allows a model to at the same time attend to different regions of an image in addition to different fragments of the question. Image representation is employed as a guide to the question attention, and also, the question representations are employed to tutor the image attention. The hierarchical architecture jointly attends to the question and image at three levels: question level, phrase level and, word level via one-dimensional CNN. The co-attended features are recursively combined to produce the final result. The model employs both parallel and alternating co-attention mechanisms. The model faces two main issues—parallel co-attention is hard to train since it employs the dot product between image and text and compress two vectors into a single value, and alternating co-attention suffers from errors accumulated at each round.

The paper [7] proposes to keep check on the language biases innate for the datasets and boost up the purpose of image comprehension in VQA. To achieve this, a balanced and coherent VQA dataset with a minimized bias is created. For each question in the dataset, there are two complementary images that looks alike but has different responses to the question. This forces the VQA model on the visual information. It follows the hypothesis that if an input NL question Q has two different answers (A and A<sub>0</sub>) for two different images (I and I<sub>0</sub>, respectively), the sole technique to figure out the correct answer is through understanding the images. Also, the proposed dataset has complementary image I<sub>0</sub> which is very similar to the original image I in semantic features. Thus, the VQA model will have to perceive the fine variations in between the two given images to correctly predict the responses.

The model answers questions about images as well as ‘explaining’ the answer about a question–image pair by supplying ‘hard negatives’.

This research developed a unique interpretable model that additionally to answering all the questions relating the images, it also provides a specimen-based description by retrieving visual scenes that look very similar to the actual input image but has totally distinct answers to the provided question. However, the process of creating a balanced dataset is highly time-consuming, and identifying complementary images involves considerable human efforts.

The paper [8] proposes a study which provides a detailed study of existing datasets, algorithms, models, and evaluation metrics employed for VQA systems.

It compares between captioning systems and VQA. The paper also conducts an analysis of publicly available benchmark datasets for VQA such as VQA, DAQUAR, COCO-QA, FM-IQA, Visual7w, and Visual Genome. It also provides an analysis of commonly used algorithms like CNN, LSTM, BOW, GRU, and other variants.

It also analyses different models employed for solving VQA tasks such as Bayesian and question aware model, attention-based, bilinear pooling, compositional VQA, and multi-modal residual networks. The paper wraps its study by evaluating different evaluation metrics for the VQA systems such as WUPS, simple accuracy, consensus metric, manual evaluation, BLEU, METEOR, CIDEr, and ROUGE.

The paper [9] proposes a multi-level attention network that joins semantic and visual attentions in an overall structure to solve the problem of automatic VQA. Visual attention allows for deep image grasping through querying, whereas semantic attention reduces the domain gap between images and questions. A multi-level attention network helps in reducing the semantic gap through semantic attention as well as gain from deep spatial inference through visual attention. First, the semantic concepts are generated in the CNN, and input text query-related concepts are selected as semantic attention. The region-based mid-level responses from the CNN are then encoded to form a spatially embedded portrayal by bidirectional RNN. Also, the answer-related regions are selected as visual attention. Finally, the semantic attention, question embedding, and visual attention are jointly optimized by a softmax classifier to finally predict the answer.

The paper concludes that attention to high-level concepts can find crucial semantic information from image and helps in removing the noisy information unrelated to the question. Incorporating circumstantial information from the neighbouring regions into the target regions gives good results for the spatial inference in images. However, the context encoding schemes employed suffer from long-term dependency problems by bidirectional. It also fails to capture the interaction among objects.

Conditional random fields (CRFs) are a class in statistical modelling technique commonly used for pattern recognition and structured prediction. A CRF varies from a classifier in the sense that it takes context into account. The paper [10] proposes a structured visual attention mechanism, which will model attention with binary hidden variables and a grid-structured conditional random fields on the hidden variables. The reasoning capabilities of CRF are executed as recurrent layers in the ANN. The model converts the iterative inference algorithms, loopy belief propagation, and mean field as recurrent layers of a full-scope neural network.

The model is able to represent the semantic structure of the scene in line with the textual input question, which removes the issue of unstructured attention which will capture only the most important nouns in the questions. However, it inclines to get bottled with the important nouns in the question and fails to find the right region in smaller images. In these situations, the model tends to give a popular answer according to the question.

The paper [11–13] proposes a fresh idea of using explanations to improve the ensembling of multiple systems. It proposes to use stacking with auxiliary features (SWAF) as an efficient ensembling technique. It acquires the knowledge to merge the results of various models using the features of the present problem as a background

for building a VQA model. It proposes four categories: three wherein parameters can be deduced from a question–image pair and does not require any querying for the constituent models and the last where auxiliary features use prototype-specific explanations. SWAF combines outputs from multiple systems using their confidence scores and task-relevant features.

For stacking the visual question answering systems, unique question–answer pairs are generated throughout all of the systems outputs prior to passing them through the stacker. If the system generates an output, then the probability is estimated for that output, if not, the confidence is considered zero. If a QA pair is classified as correct, and there are other answers that are also classified as correct for the same question, the output with the highest confidence is chosen. For input text questions that does not have even one answer classified as correct, the answer with the lowest classifier confidence is chosen; thus, it is least likely to be incorrect. Integrating explanation and ensembling thus provides a twofold advantage. Firstly, the explanations are employed to improve the accuracy of the ensemble. Secondly, the explanations from the component systems are used to build a good explanation for the overall ensemble. It also has an advantage that the visual explanations for VQA represented as localization maps can be used to aid stacking with auxiliary features.

The paper [14] proposes a unique attention mechanism that considers the reciprocal relationships between two levels of visual details. It is motivated by the study that suggests that questions can relate to both object instances as well as their parts. The model employs a CNN architecture to obtain image features for local regions on the image grid and object. The bottom-up attention is then combined with the top-down information to focus on the most appropriate scene elements for a given question. Based on the feature encodings thus received, a hierarchical co-attention model is developed which learns the mutual relationships between given questions, objects, and object parts to predict the best response. This model hierarchically combines multi-modal information such as language, object-level features, and grid-level features, with the help of a very efficient tensor decomposition scheme. In addition, the co-attention mechanism improves the image scene study by combining local image grid and object-level visual hints.

The paper [15] proposes a new word-to-region attention network for VQA systems that involves two attention layers—image attention and word attention. The prototype consists of four main parts: question model, image model, classifier model, and word-to-region attention model. The question modelling prototype processes NL queries to produce word as well as integrated question representations. The image prototype produces image characteristics from image regions. The word-to-region attention prototype contains both image and word attention, wherein the word attention projects key terms in the questions which are employed to query important image regions. The classifier model integrates the textual and visual information to predict result. This paper demonstrates the major role played by keywords of a question in identifying appropriate image object regions. It also constructs a word map that highlights the most important words as well as extracting the most relevant lingual or textual information from the question. The model also allocates varying weights to the data sequence as per their importance.

However, the model fails in extracting the keywords and useful image regions when the questions are long, complex, and contain lots of nouns.

The paper [16] proposes to advance an entity graph and use a graph convolution net (GCN) for ‘reasoning’ about the right answer by considering all entities of a fact together for a given image  $I$ , question  $Q$ , and an external knowledge base of facts to predict answer A. A fact is represented  $f = (x, r, y)$ , wherein  $x$  is the visual concept of the image,  $y$  is any attribute or phrase, and  $r \in R$ . Every question  $Q$  is associated with a single fact that helps answer the question. To retrieve a set of fitting facts for a  $QI$  pair, the model uses a scoring-based approach computing the cosine similarity of the embeddings of the words in the fact base with the words from the input text question and the words of the visual notions detected in the input visual scene. A fact score is obtained by calculating the average of the TopK word similarity scores based on downstream accuracy. For a given set of candidate entities  $E$ , to ‘rationale’ about an answer, a GCN-based approach together with a multi-layer perceptron (MLP) is used. The GCN amalgamates the entity representations in multiple iterative steps and the final remodelled entity representation is used as input in MLP for predicting a binary label,  $\{1, 0\}$ , for each entity  $e \in E$ , designating if  $e$  is the answer or not. However, the fact retrieval model is not trainable end-to-end.

The paper [17] proposes a supervised learning technique by employing VGG\_19 and Bi-LSTM for drawing features from the input visual scene and problem queries. These values are then integrated and applied attention to the prediction answer. CNN VGG\_19 algorithms treat the feature vectors from the actual input images and also take the feature vectors from the second last layer. Bidirectional long short-term memory (Bi-LSTM) is employed to extract feature values of the problem statement. A Bi-LSTM considers the current output, prior state, and posterior conditions, whereas a unidirectional LSTM can only consider the directional semantics. Then it proceeds to do vector fusion between textual and visual feature vectors through multiplication. Next, the fused vector and problem sentence feature vectors are fed for attention correlation. Finally, the fully connected layer of CNN will predict the answer. The main drawback of the model was that it had issues with problem sentence classification and identifying the central words.

The survey paper [18] discusses the three basic components in VQA—feature extraction from images and text inputs, feature fusion between image and lingual channels, and using the learned joint representation for answer prediction. The paper also elaborates on image feature extraction using convolutional neural networks (CNNs), LeNet, AlexNet, GoogLeNet, VGG-Net, and ResNet. Some textual feature extraction using recurrent neural networks (RNNs) such as Bi-LSTM, GRU, LSTM, and convolutional neural networks (CNNs) techniques like SAN, CNNQA, and MAVQA, CRAN—a combination of CNN and RNN—are also discussed.

It also details different attention mechanisms—multi-hop and single-hop attention. An m-hop attention is equivalent to extending the single-hop attention m times iteratively. Both are further classified into textual attention, visual attention, and co-attention. It also provides good insight on different techniques used for fusion between two feature channels such as fusion based on simple vector operation, CNN-based fusion, fusion based on neural networks, fusion based on bilinear models,

LSTM-based fusion, and fusion between multiple feature channels and memory network. The study reveals that fine-grained and semantic feature extraction improves the overall execution and that the models using local visual features often outperform the ones that uses only global visual features. It also notes that combining or extending the training dataset with external corpus can help in additional boosting of accuracy.

The paper [19] proposes to impose a correlation among the non-attention and attention fragments as a restriction for attention learning. Initially, an attention-guided erasing model is employed to figure the non-attention and attention parts by using a distance margin in the feature embedding space. This model removes the majorly or the minimally superior data based on attention weights as important measures and generates positive or negative training sets. The metric learning loss administers a distance constraint. Then the classification loss is applied to the model output for optimizing the whole model.

This method does not introduce any additional model parameters or inference complexity. It can also be efficiently integrated with any attention-based models.

The paper [20] proposes to extend the capabilities of VQA to community question answering platforms. The proposed model performs text pre-processing for extracting features from the question text, which is comparatively bigger in size than VQA tasks, using CNN-based architecture. Image is processed using ResNet network which utilizes the final spatial representation in case of attention-based networks and uses the final flat embedding for other models. Images are mostly resized to a uniform size. The model uses global image weight for category classification and retrieval tasks.

This paper gives a notable contribution by proposing a method for devising the global image weight for better classification and retrieval.

The paper [21] broadly and crucially reviews the present level of VQA research in terms of methodologies, datasets, and evaluation metrics.

VQA solutions employs deep learning models such as CNN for image featurization, and RNN, GRU, and LSTM for question featurization. These feature extraction phases are termed as Phase I. The Phase II is where the prepared features are combined for proving an answer to the input question regarding the input image.

Image featurization employs pure CNN, CNN with last layer removed and employing normalization and dimensionality reduction which is used to present the image scene content as a numerical vector. Question featurization employs CNN, LSTM, and GRU to implement various featurization methods like: (1) Count-based methods depend on the global occurrence of counts from the collection for calculating the word depiction, (2) prediction-based methods study the word depiction using co-occurrence facts, and (3) hybrid methods. In visual question answering system, the question and the image are generally processed parallelly to acquire the vector depictions.

The joint depiction of question text and image is implemented through baseline fusion models. The collective neural network models train the networks for the functions with selected layers chosen for joint representation of question and image

features. Some examples are deep residual networks and encoder-decoder architecture. It also discusses joint attention models like word-to-region attention network (WRAN), question-guided attention map (QAM), and question-type-guided attention (QTA). This paper also analyses some popular and benchmark datasets for VQA tasks like DAQUAR, COCO, COCO-QA, VQA, and CLEVR. It identifies that datasets are mostly categorized on category of images, QA representation, and use of external corpus. Images can be clip-art, synthetic, or natural images. It also states that datasets should be wide-ranging enough to study the mutability within the questions and must support for an unprejudiced evaluation strategy for validating the VQA models. The datasets should also be minimally biased.

## 2.2 *Dataset Description*

This paper proposes to use CLEVR dataset to train and test the model.

CLEVR is a distinguished dataset in the computer vision research area that examines a vast range of visual reasoning abilities. Unlike other widely used datasets, the CLEVR dataset has extremely low bias. It also contains descriptive annotations about different reasoning strategies to be used for different question types.

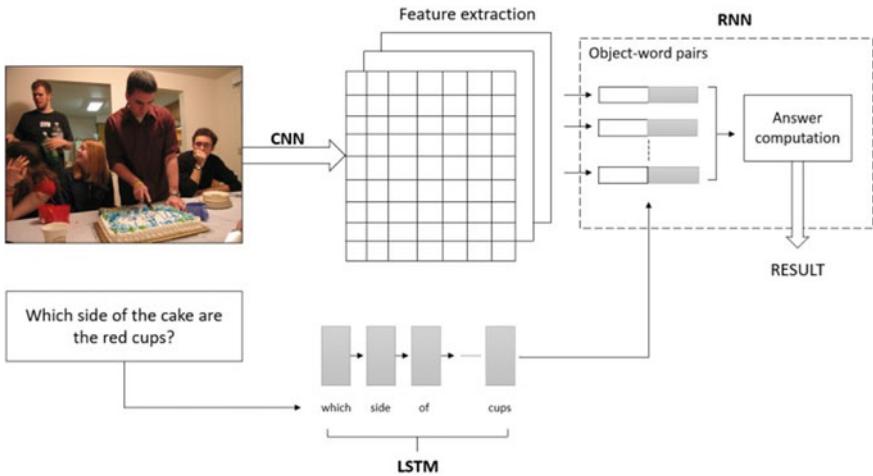
The CLEVR dataset contains a total of 100,000 images and 864,968 questions. The dataset also provides answers for all the questions in train and val sets. The dataset additionally provides annotations explaining the ground truths and relationships among various objects in the images. The questions in the CLEVR dataset analyse for various concepts in visual reasoning including counting, attribute identification, comparison, logical operations, and spatial relationships.

## 2.3 *Problem Statement*

This project proposes to model a VQA system for relation network. A relation network explores the spatial or logical relation among objects present in an image through the questions. It mainly concentrates on answering questions where the position or query is placed through relation to another object or region within the image. The proposed model will mainly concentrate on exploring the object-to-object-based or feature-to-feature-based relations in an image.

## 3 Proposed System

This project proposes to model a VQA system for relation network—questions are framed with terms which has a logical or spatial relation between them in the image. The model should be able to identify the key terms in the question and the key



**Fig. 1** Overview of the proposed model

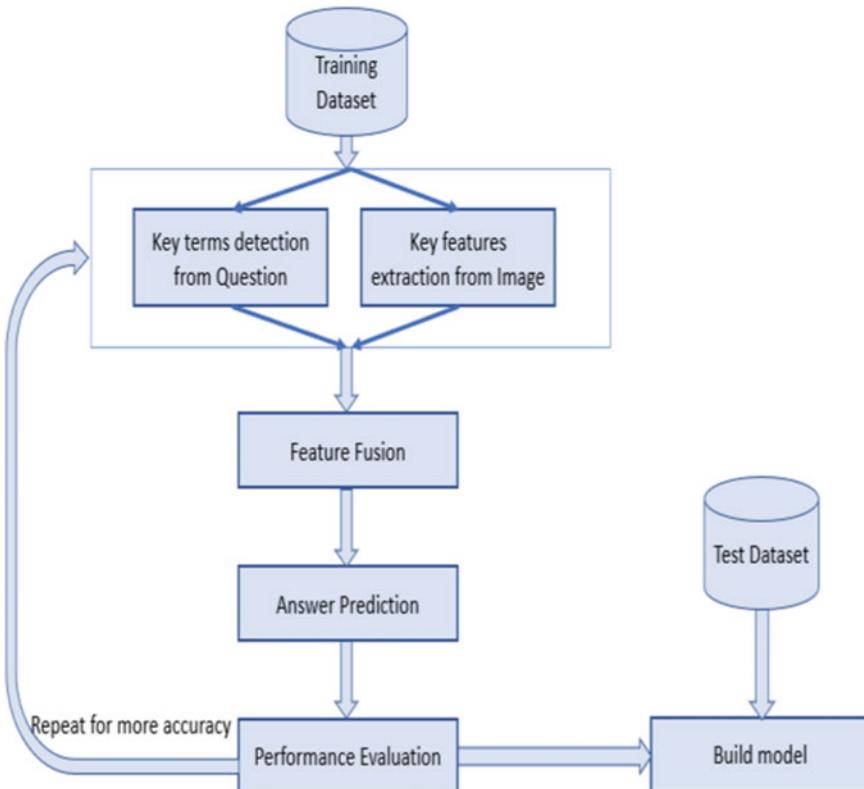
object/features from the image. Unlike the questions which can be answered through analysis of the natural language text, the proposed system requires the model to learn the image to give the correct answer (Fig. 1).

The proposed model extracts keywords from the text questions and key features from the input image. Extracting key features of text and image are in itself single modal problems. The next stage is for fusion of the extracted features. This is the stage where the problem becomes multi-modal. After the features are fused, the answers for the question are predicted using the model (Fig. 2).

## 4 Conclusion

VQA is one of the most prestigious research areas in natural language processing and computer vision that requires a system to extend beyond the capabilities of task-specific algorithms. This paper proposes a novel research in VQA systems by introducing relational networks domain into VQA.

The future work in VQA includes creating unbiased datasets with a wider variety of images and question–answer pairs. Task-specific datasets can be developed to improve accuracy for VQA systems. Another research area would be on how to minimize the human efforts for developing complementary images for images in a dataset. Another focus area with more research possibilities is to develop VQA systems to provide descriptive answers.



**Fig. 2** Architecture of the proposed model

## References

1. M. Ren, R. Kiros, R. Zemel, Exploring models and data for image question answering, in *Proceedings of the 28th International Conference on Neural Information Processing Systems—Volume 2 (NIPS’15)* (MIT Press, Cambridge, MA, USA, 2015), pp. 2953–2961
2. S. Antol et al., VQA: visual question answering, in *2015 IEEE International Conference on Computer Vision (ICCV)* (Santiago, 2015), pp. 2425–2433. doi:<https://doi.org/10.1109/ICCV.2015.279>
3. Z. Yang, X. He, J. Gao, L. Deng, A. Smola, Stacked attention networks for image question answering (2016), pp. 21–29. <https://doi.org/10.1109/CVPR.2016.10>
4. P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra, D. Parikh, Yin and Yang: balancing and answering binary visual questions, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (Las Vegas, NV, 2016), pp. 5014–5022. doi:<https://doi.org/10.1109/CVPR.2016.542>
5. J. Lu, J. Yang, D. Batra, D. Parikh, Hierarchical question-image co-attention for visual question answering, in *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS’16)* (Curran Associates Inc., Red Hook, NY, USA, 2016), pp. 289–297

6. R. Li, J. Jia, Visual question answering with question representation update (QRU). NIPS (2016)
7. Y. Goyal, T. Khot, A. Agrawal et al., Making the V in VQA matter: elevating the role of image understanding in visual question answering. Int J Comput Vis **127**, 398–414 (2019)
8. K. Kafle, C. Kanan, Visual question answering: datasets, algorithms, and future challenges. Comput. Vis. Image Underst. (2016). <https://doi.org/10.1016/j.cviu.2017.06.005>
9. D. Yu, J. Fu, T. Mei, Y. Rui, Multi-level attention networks for visual question answering (2017), pp. 4187–4195. <https://doi.org/10.1109/CVPR.2017.446>
10. C. Zhu, Y. Zhao, S. Huang, K. Tu, Y. Ma, Structured attentions for visual question answering. 1300–1309 (2017). <https://doi.org/10.1109/ICCV.2017.145>
11. M. Balaanand, N. Karthikeyan, S. Karthik, Designing a framework for communal software: based on the assessment using relation modelling. Int. J. Parallel Prog. **48**(2), 329–343 (2018). <https://doi.org/10.1007/s10766-018-0598-2>
12. N. Nguyen, B. Liu, V. Pham, C. Huang, Network under limited mobile devices: a new technique for mobile charging scheduling with multiple sinks. IEEE Syst. J. **12**(3), 2186–2196 (2018). <https://doi.org/10.1109JSYST.2016.2628043>
13. N.F. Rajani, R. Mooney, Stacking with auxiliary features for visual question answering. pp. 2217–2226 (2018). <https://doi.org/10.18653/v1/N18-1201>
14. M. Farazi, S. Khan, Reciprocal attention fusion for visual question answering. BMVC (2018)
15. L. Peng, Y. Yang, Y. Bin et al., Word-to-region attention network for visual question answering. Multimed. Tools Appl. **78**, 3843–3858 (2019)
16. M. Narasimhan, S. Lazebnik, A.G. Schwing, Out of the box: reasoning with graph convolution nets for factual visual question answering (2018)
17. C. Wang, J. Sun, X. Chen, Feature fusion attention visual question answering, in *Proceedings of the 2019 11th International Conference on Machine Learning and Computing ICMLC '19* (Association for Computing Machinery, New York, NY, USA, 2019), pp. 412–416
18. D. Zhang, R. Cao, S. Wu, Information fusion in visual question answering: a survey. Inf. Fusion, **52**, 268280 (2019). ISSN 1566 2535
19. F. Liu, J. Liu, R. Hong, H. Lu, Erasing-based attention learning for visual question answering, in *Proceedings of the 27th ACM International Conference on Multimedia (MM '19)* (Association for Computing Machinery, New York, NY, USA, 2019), pp. 1175–1183
20. A. Srivastava, H.W. Liu, S. Fujita, Adapting visual question answering models for enhancing multimodal community Q&A platforms, in *Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '19)* (Association for Computing Machinery, New York, NY, USA, 2019), pp. 1421–1430
21. S. Manmadhan, B.C. Kovoov, Visual question answering: a state-of-the-art review. Artif. Intell. Rev. **53**, 5705–5745 (2020)

# Crop Yield Prediction on Soybean Crop Applying Multi-layer Stacked Ensemble Learning Technique



S. Iniyam and R. Jebakumar

**Abstract** Due to increasing population and rapid industrialisation across countries lead to the enormous demand in food supply chain. Our proposed work results in crop yield prediction based various parameters affecting yield applying machine learning models. Comparative analysis also done with basic machine learning and advanced ensemble technique, in terms of yield accuracy with respect to various parameters such as environmental, soil and crop management factors. Our proposed multi-layer stacked ensemble model outperformed Decision tree regression (DTR), Multiple linear regression (MLR), Support vector regression (SVR), K-nearest neighbour (KNN), Random forest (RT) and Gradient boosting regression (GBR) with an accuracy score of 94.43% along with various accuracy parameter metrics like Mean absolute Error (MAE), Mean square error (MSE), Root mean square error (RMSE).

**Keywords** Multiple linear regression · Decision tree regression · Support vector machine · Random forest · Gradient boosting regression · Multi-layer stacked ensemble

## 1 Introduction

Agriculture still coined as predominant factor in countries economy. Increase in the production of crops leads to increase in GDP of a nation, which is an important economy indicator of a country. Nowadays, farmers across the globe facing lots of challenging issues in increasing crop production listed as (i) drastic climatic changes (ii) insufficient rainfall or excess rainfall (iii) lack of organic contents in the soil (iv) lack of crop management practises. Increase in crop production needs prior prediction of weather, rainfall and other important crop production factors. Crop yield

---

S. Iniyam (✉) · R. Jebakumar  
SRM Institute of Science and Technology, Kattankulathur, Chennai 603203, India  
e-mail: [iniyans@srmist.edu.in](mailto:iniyans@srmist.edu.in)

R. Jebakumar  
e-mail: [jebakumr@srmist.edu.in](mailto:jebakumr@srmist.edu.in)

prediction plays a vital role in decision making process, predicting whether particular crop will produce sufficient yield or not based on the parameters such as past crop yield history, environmental parameters, soil parameters and crop management practises. Our proposed work also engaged in crop yield prediction based on the above mentioned crop yield affecting parameters.

Rapid technological development across the globe results reduced human work and time. In agriculture also lots of technologies are involved in order to increase the production. In [1], Internet of things (IoT) helps in fixing various sensors and cameras in the agricultural fields are interconnected through network used to monitor crop water level, soil moisture, temperature, etc., results automation in various agricultural activities such as irrigation, monitoring of crops, disease detection and fertilisers management. In [2], cloud used as a storage part to store all the crop related data which was collected through the various sensors placed in the agricultural fields. Agricultural experts remotely can able to retrieve the crop data from the cloud and computation can be done for increased production. In [3], data mining used to mine the relevant crop data for crop classification, crop water management, crop disease management, crop fertiliser management, etc.

In [4], artificial intelligence and machine learning recently dominating the agriculture in terms of crop yield prediction or crop yield forecast, crop disease classification and detection, classification of crops, prediction of environmental parameters of crops, prediction of harvest with respect to rainfall expected. In [5–7], image processing also involved in early detection and classification of plant disease combined with various machine learnings classifiers. It easily classifies disease affected leaves from the healthy leaves images can be done through training and testing. In [8], remote sensing has also been widely in classification of crops, weed management, detection and classification plant disease, identification of water stress in plants, etc., from the satellite images of plant. Unmanned aerial vehicles (UAVs) and Drones are widely used for crop monitoring due to high resolution with low cost. In [9], there are several recent technologies strengthening the precision agriculture named as IoT, mobile devices, robotics, irrigation, sensors, weather modelling, etc.

The organisation of the paper is defined as follows: Section 2 describes about the related works of crop yield prediction applying machine learning, Sects. 3 and 4 mentioned about data availability and pre-processing, Sect. 5 discussed about methods and metrics used for proposed work, Sect. 6 deals about result and discussions of proposed work, and Sect. 7 discussed about conclusion.

## 2 Related Works

In [10], used combination of machine learning and deep learning approaches for crop yield estimation of corn and soybean crops. Convolutional Neural Network (CNN), Artificial Neural Network (ANN), Stacked-Sparse Auto Encoder (SSAE) engaged in crop estimation including 14 years of weather data and moderate resolution Imaging Spectroradiometer (MODIS) data. CNN as a result produced higher

accuracy with respect to accuracy parameter metrics of Root Mean Square Error (RMSE). In [11], with suitable crop modelling and remote sensing data, yield estimation of spring maize was done. Machine learning algorithms such as Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF) involved in classification and prediction of yield validating with RMSE and Mean Absolute Error (MAE). SVM produced higher accuracy with other models. In [12–14], RF engaged with the combination of forward feature selection and hyper parameter tuning for yield estimation of sugarcane crop getting higher prediction accuracy when compared to the prediction done by human experts manually as well as prediction based on last year yield results are considered as two base lines.

In [15], yield prediction of wheat implies on various climatic factors as a strong fundamental role in wheat production. Proposed statistical-based integrated climatic assessment indicator (ICAI) used to build statistical model along with meteorological factors. Two machine learning algorithms SVM and RF involved in yield prediction of wheat production. RF produced higher predictions compared to SVM. In [16], used multi-layered approach and machine learning for yield estimation of grain crop from multi-farm data sets. Observations based on pre-sowing, mid-season and late season are considered as the predictive ability factors of prediction modelling, RF outperformed other models. In [17], proposed Extreme Machine Learning (EML) model for analyse the soil fertility factors impact the coffee crop yield. Observation on organic components of soil such as pH, nitrogen, magnesium, zinc, etc., are considered as multiple predictor variables of yield. Proposed EML gives higher accuracy compared with multiple linear regression (MLR) and RF with RMSE and MAE as an accuracy parameter metrics.

In [18], proposed hierarchical machine learning for the prediction of yield estimation of different varieties of seeds belong to single crop. Weather forecast model also integrate with this predictive mechanism to classify the optimal weather parameters needed for different varieties of seeds to produce high yields. Prediction and classification accuracies are compared with Linear Regression (LR) and RF. In [19], proposed model using nonlinear approach involved various regression and classification algorithms for the weekly recommendation of irrigation plan based on the irrigation data collected from 21 sensors in the field for the duration of two years.

Models based on Gradient Boosted Regression Tress (GBRT) and Boosted Tree Classifiers (BTC) outperformed other regression and classification model. In [20], used RF for the estimation of crop production across global and regional counties of United States covering thirty years of crop yield data includes climatic and biological parameters. The performance MLR was fixed as benchmark, the proposed model compared in terms of accuracy parameter metrics RMSE and it produced higher prediction accuracy than MLR. In [21], used Supervised Kohonen Networks (SKNs) for the crop yield estimation of wheat crop based on the different layers of soil data and various characteristics of crop growth data in the form of satellite imagery. SKN outperformed other two ANN models as the benchmark.

### 3 Data Availability

The Soybean crop yield dataset was found publicly available is utilised for the proposed work includes yield performance data, soil data, weather data and crop management data, collected from the duration between the years 1980 and 2018 across 12 states of USA, which consist of 1045 countries. Yield performance data is a measure of yield values with respect to a year and a particular location. Bushels per acre is the unit of yield. Weather data includes daily observations of precipitation, maximum temperature, minimum temperature, vapour pressure, solar radiation and snow water equivalent. Soil data includes percentage of clay, bulk density of wet soil, bulk density of dry soil, available water content at upper limit of plant, available water content at lower limit of plant, percentage of organic matter in soil, percentage of sand, hydraulic conductivity and measurement of saturated water content at various depths. Management data includes weekly average of planted field's covering from April month of each year across counties in each state. In [22], from National Agricultural Statistics Service of the United States (USDA-NASS), the crop management and yield data were acquired. In [23], from Daymet: Daily Surface Weather Data on a 1-km Grid for North America, the weather data was acquired. In [24], from Gridded Soil Survey Geographic Database for the United States (gSSURGO), the soil data was acquired.

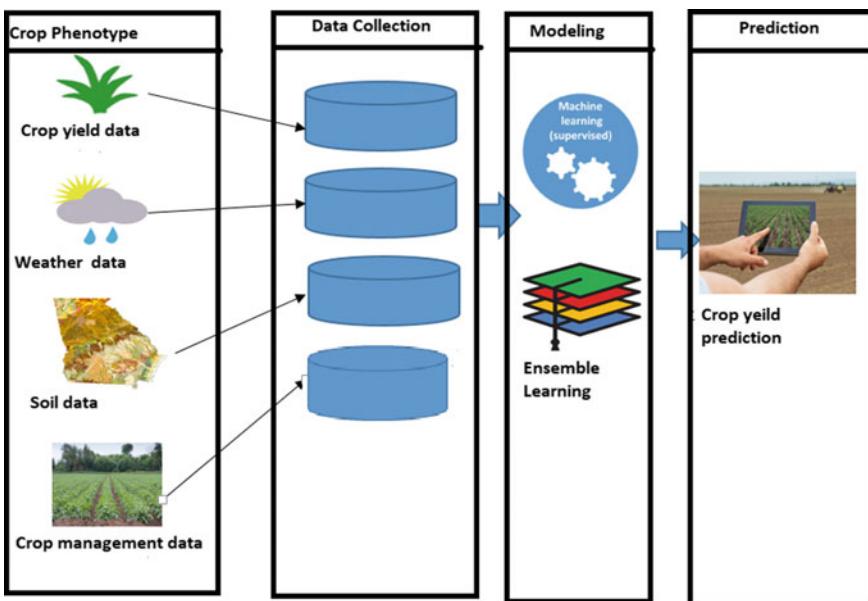
### 4 Pre-processing

The accuracy level of prediction or classification of any machine learning or deep learning models are completely rely on pre-processing of data. Sometimes, data sources may have irrelevant or noisy or missed data may drastically affects the learning ability of machine learning models. Pre-processing ensures the quality of data or more relevant data that can be extracted from data sources improves the learning rate. Daily observations of six weather data can be converted as weekly observations by averaging of week days. This results in decreases of 52 instances from 365 instances of each weather data leads significantly reduced the trainable parameters. Soil data having around 7% of missing values in various counties, missed values are replaced with mean value of same soil available in other counties. Crop management data also having around 6.5% of missing values are replaced with mean value of same variable of other counties. There is no missing values in the weather data. Other substitution methods such as median and mode were tried but mean produced higher accuracies.

## 5 Methodology and Metrics

Our proposed work, enhanced stacking ensemble machine learning based crop yield prediction or estimation underlying in the category of supervised machine learning, since the learning was carried out through various input and target variables. Crop yield can be predicted through the continuous learning from the weather data, soil data and crop management data were acted as the input variables and the crop yield value with respect to year wise and locations wise treated as output or target variable. There are two familiar types of supervised learning problem listed as regression and classification. Our proposed crop yield prediction model falls under regression, since it involved in the prediction process of crop yield.

Various machine learning regression algorithms are analysed, and the results are compared with proposed model. Figure 1 the architecture of crop yield prediction describes data collection from various sources such as crop yield, weather, soil and crop management data. Modelling done with various machine learning and ensemble learning techniques used for crop yield prediction.



**Fig. 1** Architecture diagram of crop yield prediction

## 5.1 Regression Accuracy Parameter Metrics

In machine learning, there are number of metrics were involved in measuring the accuracy of the learned model in terms of regression or classification were mentioned as follows.

**Percent Error.** This is generally termed as

$$\text{Percent Error} = \frac{\text{predicted value} - \text{original value}}{\text{original value}} \quad (1)$$

**Mean Absolute Error (MAE).** Represents the averaged absolute difference amongst original and predicted values.

$$\text{MAE} = \frac{1}{N} \sum_{j=1}^N |x_j - \hat{x}| \quad (2)$$

**Mean Square Error (MSE).** Represents the averaged square of absolute difference amongst original and predicted values.

$$\text{MSE} = \frac{1}{N} \sum_{j=1}^N (x_j - \hat{x})^2 \quad (3)$$

**Root Mean Square Error (RMSE).** Represents square root of MSE.

$$\text{RMSE} = \sqrt{\text{MSE}} \quad \text{or} \quad \sqrt{\frac{1}{N} \sum_{j=1}^N (x_j - \hat{x})^2} \quad (4)$$

**Coefficient of determination (R-Square).** Represents closest determination of coefficients to the original values.

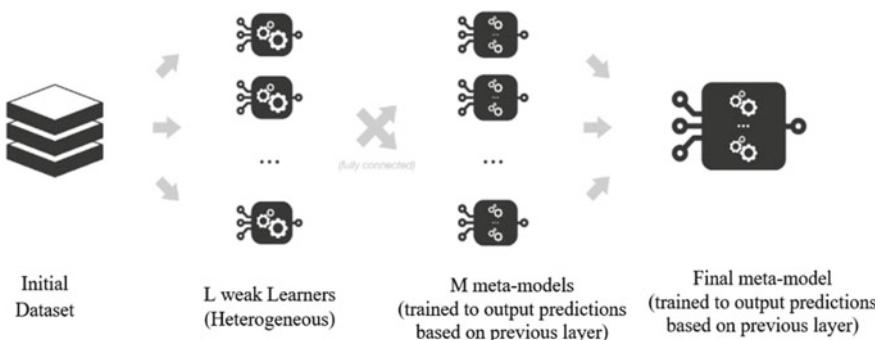
$$R^2 = 1 - \frac{\sum(x_j - \hat{x})^2}{\sum(x_j - \bar{x})^2} \quad (5)$$

From Eqs. (2), (3), (4) and (5) where,  $N$  is the total of target points in the dataset,  $x_j$  is the original value of  $x$ ,  $\hat{x}$  is the predicted value of  $x$ , and  $\bar{x}$  is the mean value of  $x$ .

## 5.2 Multilevel Stacked Ensemble Regression

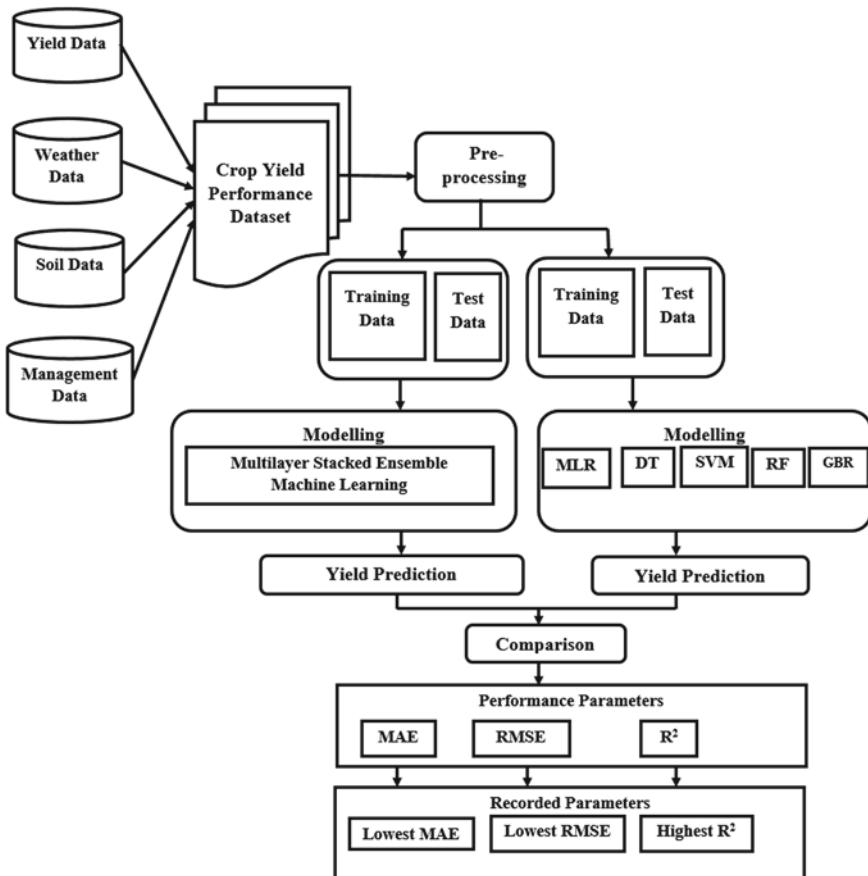
Our proposed model focused on crop yield prediction based on multiple stacked ensemble regression. It initially started to learn from one or more machine learning model, based on the prediction observations of those models, comparison are made between predicted and original values. Performance analysis of each models was done by evaluating the error metrics. Based on the error rate found on these, machine learning models were assumed as weak learners. Meta-model is the one which learns from the results of the weak learners. It combines all the results of the weak learners and complete analysis has been made with mistakes, identification of features were not fit to the model to produce accurate predictions are gathered as experiences from the weak learners. Combining all the prediction results and experiences from individual models or weak learners will act as the input for the meta-model, final predictions are made with the learnings from the previous experience and its own learning capabilities. Multi-level stacked ensemble consists of multiple layers of weak learners, final predictions by meta-model learned from results of previous layer weak learners as shown in Fig. 2. Our multi-layer stacked ensemble prediction model outperformed other bagging and boosting ensembles, since the prediction of stacked ensemble based on the combined learning of heterogeneous weak learners but predictions in bagging and boosting ensemble combined learning of homogeneous weak learners. If any homogeneous weak learner were not fit to data model for higher prediction accuracy that may traverse to the final prediction.

Proposed crop yield prediction also based on continuous learning from the week learners. Our proposed yield prediction model engaged several machine learning algorithms namely *K*-Nearest Neighbour regression (KNN), decision tree regression and support vector machine regression were individually involved in the prediction process are treated as the weak learners in the first layer. In the second layer, multiple linear regression and random forest regression are acted as meta-model learned from the previous experience or results of weak learners. Final prediction made by Gradient Boosting Regression learned from previous results in the second layer.



**Fig. 2** Multilevel stacking ensemble learning

In the multilevel stacking regression based crop yield prediction, the crop yield performance dataset includes yield data, weather data, soil data and crop management data collected from the various data sources. Pre-processing of data includes daily recorded weather observations are converted as weekly observations by cumulative average of week days, missed soil and management data are replaced with mean value of related data. Dataset after pre-processing is divided into two partitions namely training and testing. Our proposed multi-layer stacked ensemble involved in the crop yield predictions on training and testing data. Other machine learning algorithms and ensemble learning also involved in the process of prediction of crop yield on training and testing data. Comparison of both the predictions are carried out by evaluation of error metrics of regression. All the above process are represented in the form of work flow diagram as shown in Fig. 3.



**Fig. 3** Flow diagram for crop yield prediction

### Algorithm—Multilevel stacking based ensemble learning

**Given:** Set of observation  $A = \{a_i \in R^M\}$ , Prediction set  $B = \{b_i \in N\}$ , Training data set  $D = \{(a_i, b_i)\}$ .

**Input:**  $D = \{(a_i, b_i) | a_i \in A, b_i \in B\}$ .

**Output:** An multi-layer stacked ensemble prediction P.

1. **Step 1:** Learning of first-level predictions.

2. For  $t \leftarrow 1$  to  $T$  do.

3. Learn a base predictions  $p_t$  based on  $D$ .

4. **Step 2:** Construct new dataset from D.

5. For  $i \leftarrow 1$  to  $m$  do

6. Construct a newly extracted data set contains  $\{a_i^{new}, b_i\}$  where

$$a_i^{new} = \{p_j(a_i) \text{ for } j = 1 \text{ to } T\}$$

7. **Step 3:** Learning of second level predictions.

8 New predictions learning  $p^{new}$  based on the newly extracted dataset.

9. **Return**  $P(x) = p^{new}(p_1(a), p_2(a) \dots, p_T(x))$ .

## 6 Results and Discussions

The multi-level stacked ensemble based crop yield prediction model was trained and tested by a Python package called “keras” on the top and “Tensorflow” as the back end. The crop yield performance dataset includes data on yield, soil, weather and crop management data were collected from various data sources includes duration between the years 1980 and 2018 across 12 states of USA from various data sources. The weather data consist of six components as mentioned earlier are recorded on the daily basis for the duration of 38 years across 12 states. The daily observed weather component data are converted as week wise by accumulated average of week days. This drastically reduced the number of trainable parameters and made fit for the learning model. The soil data comprises 10 soil components are observed across the above-mentioned duration of years and states. Crop management practises data consist of 14 components in terms of week across the states of stipulated duration. The yield data measured in the unit of bushels per acre are fixed as target or output variable for crop yield prediction. The data on soil, weather and crop management data recorded location wise and state wise are fixed as input variables or independent variables for the prediction. There are 432 features or parameter were considered as input or independent variables used for training the proposed learning model. The crop yield performance data set having total count of 1,732,433 data used by the proposed learning model for prediction. The conceptual representation of crop yield performance data set as shown in Fig. 4. Visualisation of crop yield data points or variables univariately distributed across whole dataset as shown in Fig. 5. Figure 6 shows clearly demonstrated how the yield, year and few weather components variables are distributed across the dataset.

Loc_Id	Year	Yield	W11	W12	W13	W14	W15	W16	W17	...	P11	P12	P13	P14	P15	P16	S_surface1	S_surface2	S_surface3	
0	18	2018	190.3	0.011905	1.059524	0.821429	3.369048	0.214286	3.797619	0.476190	...	0	0	0	0	0	0	1.816667	0.818167	0.82475
1	18	2017	204.1	0.035714	1.287619	6.559524	0.619048	0.035714	1.984286	0.000000	...	5	0	0	0	0	0	1.816667	0.818167	0.82475
2	18	2016	215.6	0.000000	2.214286	0.142857	0.333333	2.083333	0.119048	0.547619	...	3	0	0	0	0	0	1.816667	0.818167	0.82475
3	18	2015	182.9	3.071429	0.678571	0.119048	0.416667	4.238095	0.023810	0.000000	...	3	0	0	0	0	0	1.816667	0.818167	0.82475
4	18	2014	194.1	2.250000	2.821429	0.880952	0.797619	1.297619	1.095238	0.476190	...	3	2	0	0	0	0	1.816667	0.818167	0.82475
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...		
3995	104	1990	129.6	0.742857	0.171429	4.685714	0.285714	8.057143	2.228571	6.342857	...	10	8	5	0	0	0	1.680000	0.652600	0.65260
3996	104	1989	132.0	2.857143	4.257143	1.428571	1.142857	2.514286	0.000000	6.171429	...	8	10	0	0	0	0	1.680000	0.652600	0.65260
3997	104	1988	79.1	0.142857	0.142857	5.028571	0.342857	14.142857	0.000000	1.800000	...	0	0	0	0	0	0	1.680000	0.652600	0.65260
3998	104	1987	135.2	0.285714	0.857143	2.142857	0.457143	0.342857	0.000000	0.742857	...	0	0	0	0	0	0	1.680000	0.652600	0.65260
3999	104	1986	135.8	0.000000	0.000000	4.171429	0.142857	5.257143	3.942857	2.800000	...	4	0	0	0	0	0	1.680000	0.652600	0.65260

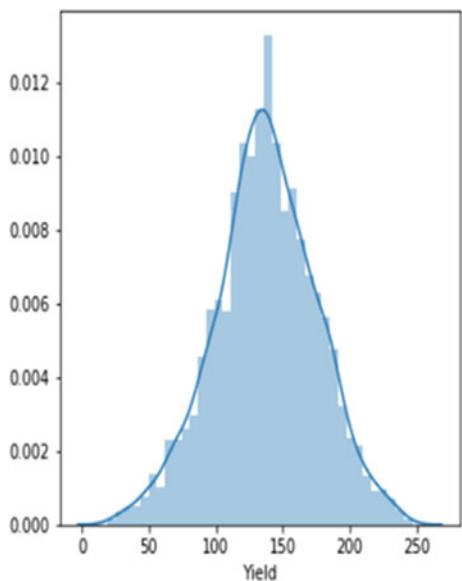
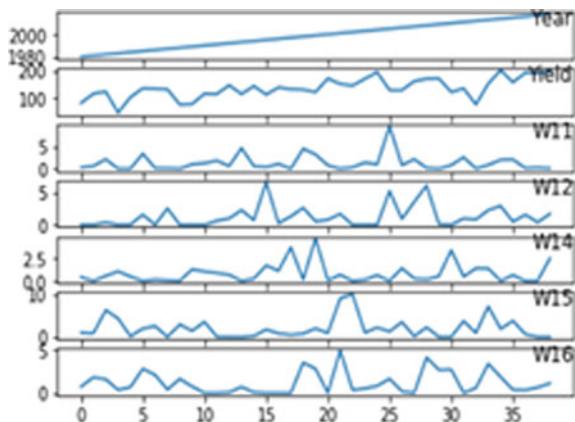
**Fig. 4** Conceptual representation of crop yield performance dataset**Fig. 5** Yield distribution summary

Table 1 shows that the prediction accuracy comparison between the proposed multi-layer stacked ensemble and other machine learning models used for crop yield prediction with evaluation of regression error metrics. Table 2 clearly depicts that differences in the actual crop yield and the predicted crop yield. Figure 7 shows the graphical representation of multi-layer stacking ensemble predictions on actual and predicted crop yield. Figure 8 shows the graphical representation of comparison of accuracies from different prediction models.

**Fig. 6** Distribution of year, yield and weather



**Table 1** Performance comparison table

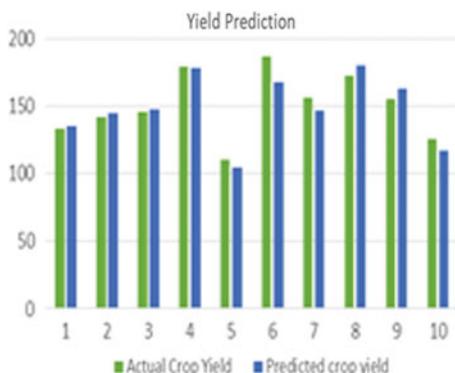
Techniques	Yield accuracy (%)	MAE	MSE	RMSE
Multiple linear regression	88.388	10.117	139.425	11.807
Decision tree regression	81.698	12.601	270.215	16.438
Support vector machine regression	84.861	11.321	250.425	15.824
K-nearest neighbour regression	78.291	14.821	352.636	18.778
Random forest tree regression	90.848	8.68	135.121	11.624
Gradient boosting regression	92.648	7.39	120.231	10.964
Multi-layer stacked ensemble	94.439	6.63	111.213	10.545

**Table 2** Multi-layer stacked ensemble predictions

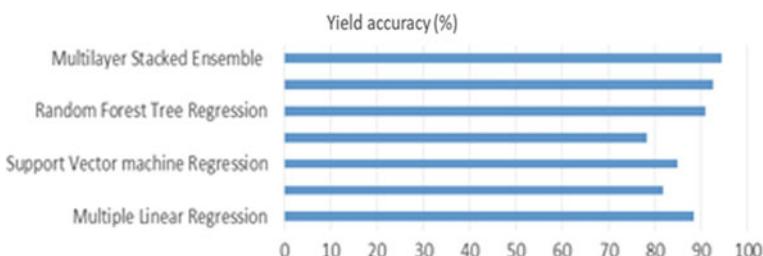
Actual crop yield	Predicted crop yield
133.0	134.80
142.0	144.68
146.0	147.74
179.0	178.05
110.0	104.14
186.5	168.00
...	...
156.0	146.49

## 7 Conclusion

Crop yield prediction plays major role in global food production. Accurate yield prediction helps the policy makers to make the timely decision about export and import of food products to strengthen national food needs. Farmers across the globe



**Fig. 7** Comparison between actual predicted yield



**Fig. 8** Comparing the accuracies of different prediction models

facing the problem of not getting the expected yield due to climatic changes, lack of soil organic contents, lack of water resources for crop needs, etc., our proposed work focused on prediction of crop yield not only based on single-crop yield value factor but it also includes the other crop yield affecting factors such as weather, soil contents and crop management parameters for crop yield prediction. In this research paper, machine learning techniques are involved in the crop yield prediction process. Proposed crop yield prediction model used multi-layer stacked ensemble machine learning for the yield prediction based on the results and experiences of the weak learners in the form of multi-layered approach. Comparative analysis also made with various machine learning regression such as MLR, DT, SVM and KNN, advanced ensemble learning algorithm such as RF and GBRT also compared. Proposed multi-level stacked ensemble outperform with the prediction accuracy of 94.439% having lowest MAE and RMSE values 6.63, 10.545, respectively.

## References

1. M. Torky, A.E. Hassanein, Integrating blockchain and the internet of things in precision agriculture: analysis, opportunities, and challenges. *Comput. Electron. Agricult.* **178**, 0168–1699 (2020)
2. Y. Ampatzidis, V. Partel, L. Costa, Agroview: cloud-based application to process, analyze and visualize UAV-collected data for precision agriculture applications utilizing artificial intelligence. *Comput. Electron. Agricult.* **174**, 105457 (2020)
3. V. Pandiyaraju, R. Logambigai, S. Ganapathy, A. Kannan, *Wirel. Person. Commun.* 1–17 (2020)
4. J. Jung, M. Maeda, A. Chang, M. Bhandari, A. Ashapure, J. Landivar-Bowles, The potential of remote sensing and artificial intelligence as tools to improve the resilience of agriculture production systems. *Curr. Opin. Biotechnol.* **70**, 15–22 (2020)
5. S. Zhang, W. Huang, H. Wang, Crop disease monitoring and recognizing system by soft computing and image processing models. *Multimedia Tools Appl* **79**(41), 30905–30916 (2020)
6. S. Iniyian, R. Jebakumar, P. Mangalraj, M. Mohit, A. Nanda, Plant disease identification and detection using support vector machines and artificial neural networks, in *Artificial Intelligence and Evolutionary Computations in Engineering Systems 2020*, ed. by S.S. Dash, C. Lakshmi, S. Das, B.K. Panigrahi, vol. 1056 (Springer, 2020), pp. 15–27. (AISC)
7. M. Sharif, R. Jebakumar, S. Iniyian, IoT based hybrid plant disease detection for yields enhancement. *Europ. J. Mol. Clin. Med.* **7**(8), 2134–2153 (2020)
8. J. Segarra, M.L. Buchaillot, J.L. Araus, S.C. Kefauver, Remote sensing for precision agriculture: sentinel-2 improved features and applications. *Agronomy* **10**(5), 641 (2020)
9. S. Verma, A. Bhatia, A. Chug, A.P. Singh, Recent advancements in multimedia big data computing for IOT applications in precision agriculture: opportunities, issues, and challenges, in *Multimedia Big Data Computing for IOT Applications*, pp. 391–416 (2020)
10. S. Ju, H. Lim, J. Heo, Machine learning approaches for crop yield prediction with MODIS and weather data, in *40th Asian Conference on Remote Sensing: Progress of Remote Sensing Technology for Smart Future*, ACRS, Daejeon, South Korea (2019)
11. I. Ahmad, A. Ullah, M.H. ur Rahman, J. Judge, Yield forecasting of spring maize using remote sensing and crop modeling in Faisalabad-Punjab Pakistan. *J. Indian Soc. Remote Sens.* **46**(10), 1701–1711 (2018)
12. G. Rajmohan, C.V. Chinnappan, A.D. William, S.C. Balakrishnan, B.A. Muthu, G. Manogaran, Revamping land coverage analysis using aerial satellite image mapping. *Trans. Emerg. Telecommun. Technol.* (2020). <https://doi.org/10.1002/ett.3927>
13. N.T. Le, J.-W. Wang, C.-C. Wang, T.N. Nguyen, Novel framework based on HOSVD for ski goggles defect detection and classification. *Sensors* **19**, 5538 (2019). <https://doi.org/10.3390/s19245538>
14. P. Charoen-Ung, P. Mittrapriyanuruk, Sugarcane yield grade prediction using random forest with forward feature selection and hyper-parameter tuning, in *IC2IT: International Conference on Computing and Information Technology* (2018), AISC, vol.769, ed by H. Unger, S. Sodsee, P. Meesad (Springer, Thailand, 2018), pp. 33–42. [https://doi.org/10.1007/978-3-319-93692-5\\_4](https://doi.org/10.1007/978-3-319-93692-5_4)
15. X. Xu, P. Gao, X. Zhu, W. Guo, J. Ding, C. Li, M. Zhu, X. Wu, Design of an integrated climatic assessment indicator (ICAI) for wheat production: a case study in Jiangsu Province, China. *Ecol. Indic.* **101**, 943–953 (2019)
16. P. Filippi, E.J. Jones, An approach to forecast grain crop yield using multi-layered, multi-farm data sets and machine learning. *Precision Agric.* **20**, 1015–1029 (2019)
17. L. Kouadio, R.C. Deo, J.F. Adamowski, Artificial intelligence approach for the prediction of Robusta coffee yield using soil fertility properties. *Comput. Electron. Agricult.* **155**, 324–338 (2018)
18. H. Zhong, X. Li, D.B. Lobell, S. Ermon, M.L. Brandeau, Hierarchical modeling of seed variety yields and decision making for future planting plans. *CoRR* abs/1711.05809 (2017)

19. A. Goldstein, L. Fink, A. Meitin, S. Bohadana, O. Lutenberg, G. Ravid, Applying machine learning on sensor data for irrigation recommendations: revealing the agronomist's tacit knowledge. *Precision Agricult.* **19**(3), 421–444 (2018)
20. J.H. Jeong, J.P. Resop, N.D. Mueller, D.H. Fleisher, K. Yun, E.E. Butler, D.J. Timlin, K.M. Shim, J.S. Gerber, V.R. Reddy, Random forests for global and regional crop yield predictions. *PLoS One* **11**(6), e0156571 (2016)
21. X.E. Pantazi, D. Moshou, T. Alexandridis, R.L. Whetton, A.M. Mouazen, Wheat yield prediction using machine learning and advanced sensing techniques. *Comput. Electron. Agricult.* **121**, 57–65 (2016)
22. USDA—National Agricultural Statistics Service Available at: <https://www.nass.usda.gov/>
23. P. Thornton, M. Thornton, B. Mayer, Y. Wei, R. Devarakonda, R.S. Vose, Daymet: daily surface weather data on a 1-km grid for North America, version 3. (ORNL Distributed Active Archive Center). <https://doi.org/10.3334/ORNLDAAAG/1328>
24. Soil Survey Staff. Gridded Soil Survey Geographic (gSSURGO) Database for the United States of America and the Territories, Commonwealths, and Island Nations served by the USDA-NRCS (United States Department of Agriculture, Natural Resources Conservation Service)

# A Comparative Analysis to Obtain Unique Device Fingerprinting



T. Sabhanayagam

**Abstract** Main focus of this paper is to obtain unique device fingerprint using JavaScript specifically Angular JS and Client JS techniques which together can make the comparative analysis more efficient and valuable. The term ‘Device Fingerprinting,’ is a phenomenon to collect information such as device name, java installed and its version, user agent, browser version, etc., of an individual computing device for the purpose of identification. Through this technique, even if the cookies are turned off we can obtain individual device fingerprints. Device fingerprinting is often immutable and tends to rapidly change, making it challenging to get a unique one. The majority of solutions available to obtain a unique device fingerprint are complex, and most solutions don’t have the sufficient efficiency to obtain the required. Thus, this comparative analysis to obtain unique device fingerprint using JavaScript techniques simplify this challenge and create an opportunity in analyzing and obtaining the data in much more efficient way..

**Keywords** Device fingerprinting · Unique identification · Unique fingerprint · PC fingerprinting

## 1 Introduction

Device fingerprinting uses JavaScript techniques which includes Angular JS and Client JS along with Spring MVC and Restful Web service as back-end techniques for the comparative analysis. Device fingerprinting is about collecting information or data, its wider usage in different applications and obtaining efficient fingerprint [1]. Device fingerprinting is a business initiative through which new applications use this unique fingerprint to ensure security, prevent fraudster, easy and user-friendly environment to users. Device fingerprinting enable organizations to accomplish several objectives:

---

T. Sabhanayagam (✉)

Department of Computing Technologies, SRM Institute of Science and Technology,  
Kattankulathur, Tamil Nadu, India

- Analysis done to detect anomaly beyond the traditional analysis to provide first unique fingerprint. Use cases using this first fingerprint will lead and support real-time applications to obtain unique fingerprint anytime anywhere.
- We can create velocity filters based on the fingerprint that will give a way to minimize the costs of fraud when details like names, personal card details and Internet Protocol addresses are changed.
- Analyze information related to transactions in order to have a secure transaction.
- Detect all types of illegally involved or wrongly involved accounts or subscriptions that are made by fraud experts or automated systems.
- Improve big business outcomes and manage risk of detect risk which comes from customers that share same device network, i.e., they are blacklisted.

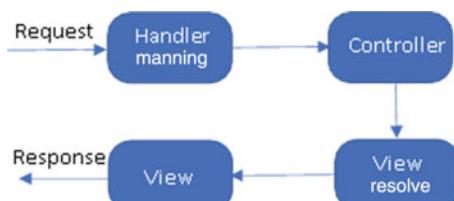
In short, device fingerprinting provides the capability for an organization or individual to ensure data security and individual identification to have efficient enterprise. Device fingerprinting uses techniques which are available to also work on mobile phones and extract data from an individual mobile phone [2]. Device fingerprinting means for performance and capacity because of the following listed below: Unique-ness—how confidently and specifically one lists out and differentiate a computer from the others on the web. It relies on the data that the fingerprint possesses [3–5]. Existing fingerprinting system is of two types; client based and server based, for client based type executable software are installed via these system. For server based type, measured remotely via a profiling server.

## 2 Spring MVC Framework

Spring MVC builds versatile around coupled web applications. Model-view controller setup separates business legitimization, introduction technique for considering and course present. Models are in charge of tending to the application information. Viewpoints condense reaction to client's assistance for model request [6]. Controllers are responsible for continuing demand from the client and coming back to the back-end affiliations. The dispatcher Servlet first gets the demand; actuate Handler Mapping and sums up the Controller related with the demand of the user. Controller approach the urgency by employing right association structures (Fig. 1).

Model and view challenge contain the model information and the view tag. View with assistance of model information will render the outcome back to the client.

**Fig. 1** Spring MVC framework



### 3 Restful Web Service

Restful web administrations are worked to work best on the Web. Illustrative State Transfer (REST) is an engineering style that determines limitations, for instance, the uniform interface that if connected to a web benefit incite alluring properties. REST structural style constrains an engineering to customer/server design intended to utilize stateless correspondence convention.

### 4 Characteristics of Device Fingerprinting

There are various variables you should consider, while executing a device fingerprinting innovation. Zero effect device Fingerprinting arrangement must have zero effect arrangement on both your client encounter and additionally your IT basics. A client to need to download programming or utilize an equipment token that will prompt disappointment and deserted framework. Establishment adaptability the gadget fingerprinting innovation can be executed as a web administration to diminish establishment, upkeep and adjustment costs [7].

An all-around characterized web-API will empower straightforward and financially savvy blend of device intelligence into your current association rules motors, hazard-based access control frameworks. Continuous settlement motor knowledge is just as profitable. This device arrangement fit for computing risk continuously as opposed to minutes, hours, days. Fluffy coordinating one way to deal with producing a device play out basically hash of estimated characteristics. The impediment of such an approach is, to the point that it takes just a single parameter to change, for instance swapping the program utilized from Internet Explorer to Firefox, and a completely new device is created [8–12]. Search device arrangement that requires a fluffy coordinating method to give, precise and constant device fingerprint. The present programs bolster advancements, for example, Flash and JavaScript that are equipped for social occasion broad device fingerprinting data. Be that as it may, such innovations are likewise ready to be promptly impaired by phonies and protection cognizant like network surfers. Have the capacity in order to separate a phony and significant client. Gadget fingerprint depth much the same as an ice sheet.

Any network servers examine organizations give essential program techno designs, for example, program compose, program dialect and the sorts of mixed media and dialects that are upheld [13–15]. Notwithstanding, this data additionally simple to control by an educated fraudster and is likewise ignorant concerning cautioning signals underneath which program will be let you know. Search for device fingerprinting approach looks past program and can perform operating system, protocol and connection fingerprinting continuously. Advantage of advancement can perceive a fraudster notwithstanding when the program traits change or when treats are erased, and would more be able to precisely aware of a high hazard intermediary being utilized. For these organizations, perceiving an arrival fake gadget isn't as helpful as

having the capacity to distinguish a misrepresentation endeavor the first-run through, progressively. Search for a device fingerprinting arrangement that can give first-time extortion insight, for example,

- Whether the Device is holing up behind an intermediary and precisely decide the hazard related with the intermediary.
- The True IP and not only the Proxy IP that the gadget is association from.
- The True Geo that the association began from, and not only the area of the intermediary utilized.
- Regardless of whether, a gadget has been bargained by malware and has a place with a botnet.

## 5 Device Fingerprinting Mechanism

Device fingerprinting is becoming an important component of identification in the age of new tools and applications. As the uniqueness, persistence, fit and modularity of unique identification of individual remote computing device grows, device fingerprinting will also be become more important to be used in different application including security.

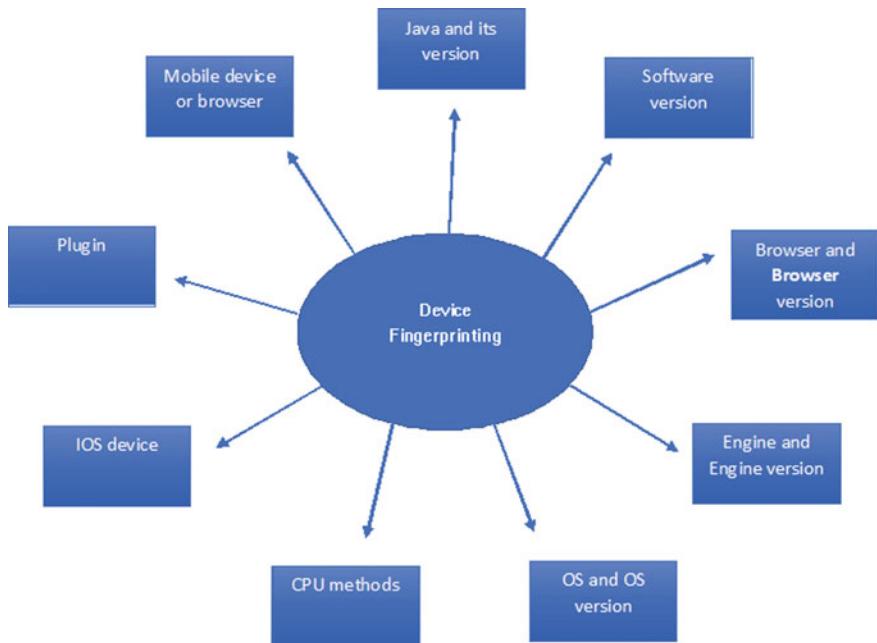
Dealing with unique fingerprint id obtained and integrating it to various systems will increase efficiency and usability if the system. The mechanism used consists of implementation of a client is system using spring MVC framework. Analyzing it to obtain unique fingerprint by testing it on different systems. Followed by implementation of angular js system using restful web services [16]. Interpreting both the systems in various test cases, we obtained unique fingerprint.

Figure 2 describes a user's individual device fingerprint attributes. The attributes consist of browser and its version, Java installed and its version, Plugins installed, engine and its version, CPU methods, OS and its version, IOS device, software versions, etc.

Thus, the existing systems which are used to obtain fingerprint are based on two methods:

- Client based method which uses a software to be installed in the remote computing device and hence obtain data.
- Server based method which uses JavaScript techniques to obtain fingerprint using different frameworks suitable.

Here, we are using server based method to implement our system. Instead of client based technique in which user privacy or integrity is not maintained. Here, the server based method has a drawback which is MAC address is not obtained in the unique fingerprint as MAC address is the main key to fraudster or even the violation of the user privacy [19, 20]. This can also be called as a drawback or data security key to user-friendly system [17, 18]. We consider pairing device fingerprinting with different application to enhance security, prevent fraudster, generate easy or automated system of you own. The information obtained from the device fingerprint can be used entirely



**Fig. 2** Device fingerprinting attributes

and can provide an automated help in selecting the best ways to present the data. Thus, device fingerprinting can also be made in a way to easily deploy the entire data in a much approachable way of presentation. For an easier understanding, consider your data to be great and unique information in terms of fingerprinting.

## 6 Conclusion

Device fingerprinting using client JS techniques may not be a perfect solution for obtaining unique fingerprint but on the other side device fingerprinting using angular JS techniques is providing the best solution by obtaining a unique fingerprint which includes Java and its version, software version, IOS device, CPU methods, engine and its version, etc. All these together are giving us an identification mark of the individual computing device and hence can be used in many applications. Some applications in which this unique device fingerprint can be used are banking services, identification purposes, user-friendly applications, automata system application. Through device fingerprinting, organizations or individuals can have access and can analyze the information of any individual computing device and can use it in the real-time applications efficiently. Hence, unique device fingerprinting is a research in itself.

## References

1. K. Abe, S. Goto, Fingerprinting attack on Tor anonymity using deep learning, in *Proceedings of the APAN—Research Work-shop 2016* (2016)
2. J. Hayes, G. Danezis, K-fingerprinting: a robust scalable website fingerprinting technique, in *USENIX Security, 2016*, pp. 1187–1203
3. G. Cherubin, J. Hayes, M. Juarez, Website finger-printing defenses at the application layer, PoPETs, vol. 2, pp. 186–203, (2017)
4. M. Ayenson, D. Wambach, A. Soltani, N. Mind blowing, C. Hoofnagle, Streak treats and security ii: Now with html5 and etag respawning, available at SSRN 1898390 (2011)
5. K. Boda, A.M. Foldes, G.G. Gulyás, S. Imre, Client watching and following on the web through cross-program fingerprinting, in *Proceedings of (game- plan of occasions) of the 16th Nordic Conference on Information Security Technology for Computer programs*, ser. NordSec’11, 2012, pp. 31–46
6. A. Hintz, Fingerprinting websites using traffic analysis, in *Privacy Enhancing Technologies*, pp. 171–178 (2002)
7. M. Perry, E. Clark, S. Murdoch, The plan and execution of the tor program [draft][online], joined states (2015)
8. G. Berk, Z. Andreas, E. Thomas, B. Sunar, PerfWeb. How to violate web privacy with hardware performance events, in *ESORICS*, pp. 80–97 (2017)
9. B. Krishnamurthy, K. Naryshkin, C. Wills, Confirmation spillage versus insurance measures: the making parcelled, in *Web 2.0 Security and Privacy Workshop* (2011)
10. S. Zhu, V. Saravanan, B. Muthu, Achieving data security and privacy across healthcare applications using cyber security mechanisms. *Electron. Libr.* **38**(5/6), 979–995 (2020). <https://doi.org/10.1108/el-07-2020-0219>
11. N.T. Le, J. Wang, D.H. Le, C. Wang, T.N. Nguyen, Fingerprint enhancement based on tensor of wavelet subbands for classification. *IEEE Access* **8**, 6602–6615 (2020). <https://doi.org/10.1109/ACCESS.2020.2964035>
12. P. Eckersley, How (like nothing else on the planet) is your web program?, in *Proceedings of (strategy of occasions) of the tenth International Conference on Privacy Improving Technologies*, ser. PETS’10 (2010)
13. F. Roesner, T. Kohno, D. Wetherall, Recognizing and protecting against outsider watching and following on the web, in *Proceedings of (strategy of occasions) of the ninth USENIX Conference on Networked Systems Design and Putting into utilization*, ser. NSDI’12, 2012, pp. 12–12
14. Etienne, J. Etienne, Customary Suzanne monkey from blender to kick your redirection off with threeex.suzanne
15. X. Dish, Y. Cao, Y. Chen, I don’t comprehend what you go by the past summer—shielding clients from outsider web watching and following with audit and following free program, in *NDSS* (2015)
16. D. Fifield, S. Egelman, Fingerprinting web clients through (game-plan of printed letters of a near style) numbers that measure things, in *(identified with directing cash) (the examination of making riddle codes) and Data Security* (Trapr, 2015), pp. 107–124
17. A. Lerner, A.K. Simpson, T. Kohno, F. Roesner, Web jones and the hooligans of the lost trackers: an (identified with considering individuals who carried on quite a while back) examination of web watching and following from 1996 to 2016, in *25th USENIX Security Symposium (USENIX Security 16)* (Austin, TX, 2016)
18. M. Mulazzani, P. Reschl, M. Huber, M. Leithner, S. Schrittweis, E. Weippl, F. Wien, Smart and solid program perceiving proof with javascript motor fingerprinting, in *W2SP* (2013)
19. A. Soltani, S. Canty, Q. Mayo, L. Thomas, C.J. Hoofnagle, Streak treats and security, in *AAAI Spring Symposium: Smart Information Privacy Management* (2010)
20. L. Lu, E.-C. Chang, M.C. Chan, Website finger-printing and identification using ordered feature sequences, in *ES-ORICS* (2010), pp. 199–214

# Adware and Spyware Detection Using Classification and Association



Kalyan Anumula and Joseph Raymond

**Abstract** In the android platform, the growth of application development reached 10 million per year. The application developers require to advertise their applications for describing its usage. Among all applications, the malware developers also introduced their applications with obfuscated names and allure to install them on our mobile devices. In this process, the applications run in the background and raise advertisements to bypass the users from the identification of malware. Here we proposed the system with adware and spyware detects by classification and association mechanism which defers specific data to be stolen. Analysis of the application data while running makes security scientists strategically tough tasks to accomplish.

**Keywords** Spyware · Adware · Classification and association

## 1 Introduction

In the current generation, mobile devices stuck as part of the human body. It functions as a pocket computer. The mobile trend has been increasing with new features and is compatible with computers. Every day the growth of application releases greatly increases. Because android is open-source, API libraries are pre-installed to develop, everyone can implement their own applications. To reach out to the users in which the third parties are developed, those are being distributed over android application stores for user installation. Applications emerge with features that are dependent on permissions come up with an android manifest XML file. This XML file is clenching with all permissions, intent-filters, broadcast receivers that are demanded by the application. The android operating system has been influencing mobiles for the past 10 years. As of the third quarter of 2020, android users were able to select between 2.87 million applications, structuring Google Play the application store with a high scale number of existing applications according to [1].

---

K. Anumula (✉) · J. Raymond

Department of Information Technology, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur 603203, India  
e-mail: [ak7954@srmist.edu.in](mailto:ak7954@srmist.edu.in)

When we look into development, not all applications are benevolent. Where some third parties are developed malicious threatened applications for robbing user information and their daily activities.

The divergent malicious applications are in the forms, such as Trojans, keyloggers, phishing, ransomware, spyware and adware. Malware is assembled by developing their own application or decompiling the existing application and add the malware content, compile it then give out over third-party application stores. Each malware type has its self-characteristics to handle. In this paper, we are proposing adware and spyware detection by specifying features and its association.

Spyware is an application that can hide its stealing behaviour and is exposed as a genuine application.

Some applications would not display an icon on the screen of mobile devices, and it runs as a native application. Spywares are used by organizations that are embedded in mobile devices because to trace out the activities of employees. Hackers also implemented their own applications to sneak and disclose user information. Spyware can also analyse the network traffic and monitoring system to know user objectives for phishing attacks. With the reference [injection], spyware developers have knowledge of victim proceedings. These applications are sophisticated the vulnerabilities existed in the operating system that would aid to access the root, then it would control the entire device. It could modify, delete and add the libraries to the operating systems and applications when it once accesses the root. It runs silently and frequently in the span of hours, days, years, etc.

Adware is an application that can advertise the ads and performing malicious actions such as device information transferring, collecting photos, call logs, location of the device. It brings down the performance and battery capacity of an android device. Often broadcasting the advertisements makes the user distressed. More often ad libraries are encapsulated in apk file for generating revenue and popularize the ad-related content. There are countable ad networks that provide ads to applications such as Admob, Ad colony, Facebook, Fyber and Unity ads. Adware applications will run on a user's device and will silently click ads in the background and without the user's knowledge to generate ad revenue. This type of application can probably download and install another apk inside a mobile device. Application marketing strategies have replaced in agile towards advertising. Ads are viewed while browsing online or while using an application [machine learning classification].

Adware and spyware are majorly interconnected in certain feature behaviours. They are focusing on device user information commonly such as location, camera, calls, network access and storage access. Adware can perform their motivations during advertisements, and spyware can perform malevolent motivations during runtime under background which is not corresponding with UI actions.

In the following section, we have split as Sect. 2 for the literature review, Sect. 3 for methodology, Sect. 4 for conclusion and finally references.

## 2 Literature Review

Dynamic analysis of android malware, particularly adware and spyware, detection can be handled based on feature selection. In Ref. [2], the authors introduced adware detection by using combinational features extracted from static and dynamic analysis. Extracted features make as datasets and trained with supervised machine learning.

Static can be done by decompiling apk packages and extract the strings, permissions from the manifest XML file. If the strings are obfuscated, identifying the words is being a severe task. Collecting the features by removing duplicates existed within them. It results in a 97% detection rate that even verified by third-party evaluators.

In Ref. [3], they implemented AdDetect framework with the mechanism called module decoupling. Analysed in-app automatic semantics of applications to identify the role. Module decoupling describes partitioned the non-ad-library and ad-library modules and then applied semantic analysis on its dependent functions. As refer this paper, some applications may raise ads with downloading packages after network connection. Applications would have irrelevant functions and are used to download ad packages followed by monitoring when application work. AdDetect carries out 95.34% accurate detection with minor false positives.

In [2, 3], support vector machine (SVM) practised with dataset that to detecting the application whether adware or not. In Refs. [4–6], the authors had proposed with multiple classifier machine learning algorithms and each compared with metrics such as accuracy, false positive, recall, precision and  $f$ -measure. In Refs. [7, 8], it is implemented based on network traffic with the nine feature measurements. It makes the comparison over random forest (RF), K-nearest neighbour (KNN), decision tree (DT), random tree (RT) and regression (R) over metrics and find the average as accuracy (91.41%), precision (91.24%) and false positive (0.085). It is applied for finding whether the application is malware or benign along with adware identification.

This entirely applied on network phenomenon. This paper used the adware families such as Airpush, Dowgin, Kemoge, Mobidash and shuanet for feature extraction.

- *Airpush*: it is working as spontaneously run the advertisement along with stealing the information.
- *Dowgin*: it is a library to generate ads and storing user information in background.
- *Kemoge*: it would like to control user's android device. This adware is a hybrid of botnet and disguises itself as popular applications via repackaging.
- *Mobidash*: Designed to monitor ads and to compromise user's private information.
- *Shuanet*: It makes try to bypass the antivirus tools and capture and transferring user device information.

In Ref. [9], the authors contributed the system with feature transformation-based android malware detection (FARM) which is irreversible. They used landmark-based transformation which expresses where the methods have been applying to perform

malevolent actions, feature clustering which finds specific type of malware, and generated correlation graph that represent similarities over applications that make closure to certain type. They permuted classification over good ware, rooting, banking Trojan and spyware. They enclosed basic features in the form of package followed by class followed by functions. These are held when variable names are behaviour related, if the name of variables, functions may obfuscate that make severe to the identification of background functioning.

In Refs. [10, 11], taken spyware applications and run-in android environment and analysed the activities done by them. They applied fuzz techniques to make results where they reach out to test cases made by authors. They even mentioned how the spyware gets injected and how to detect from mobile device.

From [1, 12, 13], contributed detection of malware with their respective mechanisms. In [1], used machine learning algorithms to spot malware or not. In [13, 14], enforced Boolean-valued vector to create inject into cluster algorithms for identification.

### 3 Methodology

In this proposed contribution, we have implemented adware and spyware detection. It implements classification on features and its association to implement datasets. Those datasets are become trained with clustering algorithm K-means. The cluster mechanism splits the features of adware and spyware to make detection analysis.

The step-by-step methodology can be explained in Fig. 1. As

- Running application
- Collecting logs
- Analysis
- Classification and association
- Cluster analysis.

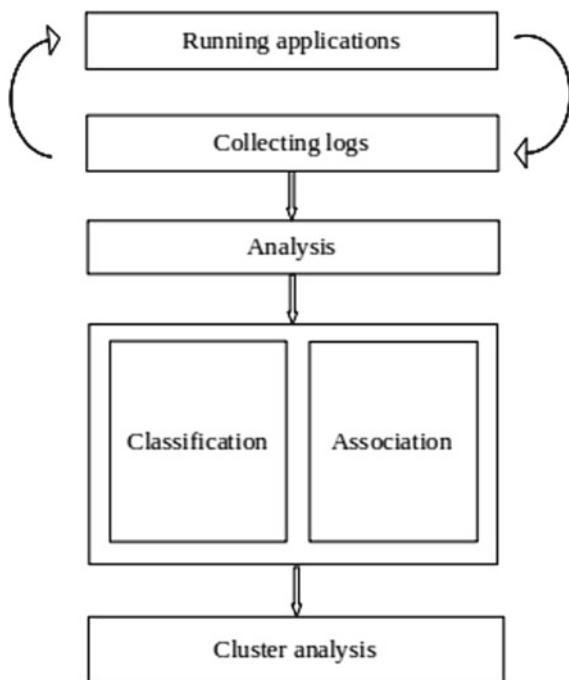
#### (i) **Running application:**

To run the application, user interacts with window interface which contains icons in android device that often senses for user interaction. Switching icons that open activity manager to run application. While running application, it creates logs that determine behaviour and functionalities. When starting activity manager, it generates process id that determines running application.

#### (ii) **Collecting logs:**

Logcat is a tool that helps to collect logs. In logs, we can filter based on process id, classes, methods and priority. These can differentiate the information specific to each application. Process id differentiates each application that is running. Classes and methods provide functionality information associated with it. There are five specific log severities such verbose, debug, information, warning and error. Verbose

**Fig. 1** Methodology of detection flow



determines every event of application, debug for identify errors, information for normal data, warning for exception access and finally error for unknown data given that rise errors.

#### (iii) Analysis:

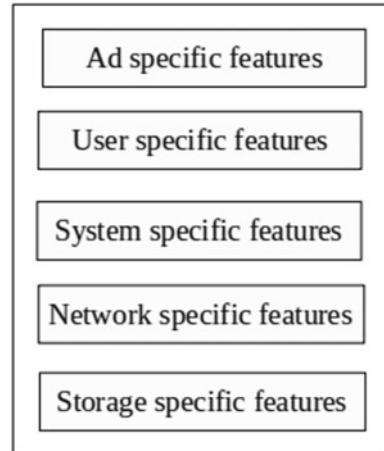
Analysis makes the process of filtering logs that are specific to each application. Each application is taken as package in android environment. It can use the components and actions determined in activity manager event log. Each package is handled by package manager that makes monitor intents between packages, intent-filters. Activity manager event monitors interconnection between activities that exist in the same application or other application. Activity represents user view that holds strings, layouts and animations.

#### (iv) Classification and Association:

In this proposed system, we have classified the features in terms of ad-specific, user-specific, system-specific, network-specific and storage-specific as shown in Fig. 2.

Ad-specific features represented as ads can be visualized by bypass the running application. It can use the web-view, Admob, Google Ads sources to display Ads with in in-built SDKs encapsulated in application.

User-specific features represented as call logs, contacts, images, live location. These are specific to know the user information. System-specific features related

**Fig. 2** Classification

to components such as camera, sensors and device information such as device id, IDMI, version of OS. Network-specific holds connectivity between user to remote. Finally, storage-specific such as database access, requesting information from another application.

Association refers test cases such as

1. Ads followed by network followed by user-specific, system-specific and storage-specific information
2. Ads followed by network-specific actions that download additional packages
3. Network followed by user-specific, system-specific and storage-specific
4. Network-specific actions that download additional packages.

(v) Cluster analysis:

In this clustering algorithm picking the dataset that provided by association to make training with K-means algorithm.

## 4 Conclusion

In this paper, we elaborate the way of dynamic analysis and relate the specific information that the malicious apps can do. It would identify category of stealing data. The log collection and filtering take more challenging to result the behaviour. It requires keep analysing on latest malware that expose how hackers implemented and what mechanisms used to bypass the antivirus applications.

## References

1. R. Agrawal, V. Shah, S. Chavan, G. Gourshete, N. Shaikh, Android malware detection using machine learning, in *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)* (2020). <https://doi.org/10.1109/ic-etite47903.2020.90491>
2. I. Ideses, A. Neuberger, Adware detection and privacy control in mobile devices, in *2014 IEEE 28th Convention of Electrical & Electronics Engineers in Israel (IEEEI)* (2014)
3. A. Narayanan, L. Chen, C.K. Chan, AdDetect: automated detection of android ad libraries using semantic analysis, in *2014 IEEE Ninth International Conference on Intelligent Sensors, Sensor Networks and Information Processing (ISSNIP)* (2014)
4. D. Vu, T. Nguyen, T.V. Nguyen, T.N. Nguyen, F. Massacci, P.H. Phung, A convolutional transformation network for malware classification, in *2019 6th NAFOSTED Conference on Information and Computer Science (NICS), Hanoi, Vietnam* (2019), pp. 234–239. <https://doi.org/10.1109/NICS48868.2019.9023876>
5. S. Zhu, V. Saravanan, B. Muthu, Achieving data security and privacy across healthcare applications using cyber security mechanisms. *Electron. Libr.* **38**(5/6), 979–995 (2020). <https://doi.org/10.1108/el-07-2020-0219>
6. J.Y. Ndagi, J.K. Alhassan, Machine learning classification algorithms for adware in android devices: a comparative evaluation and analysis, in *2019 15th International Conference on Electronics, Computer and Computation (ICECCO)* (2019). <https://doi.org/10.1109/icecco48375.2019.9043288>
7. A.H. Lashkari, A.F. Kadir, H. Gonzalez, K.F. Mbah, A. Ghorbani, Towards a network-based framework for android malware detection and characterization, in *2017 15th Annual Conference on Privacy, Security and Trust (PST)* (2017)
8. P. Kaur, S. Sharma, Spyware detection in android using hybridization of description analysis, permission mapping and interface analysis. *Procedia Comput. Sci.* **46**, 794–803 (2015)
9. Q. Han, V.S. Subrahmanian, Y. Xiong, Android malware detection via (somewhat) robust irreversible feature transformations. *IEEE Trans. Inf. Forensics Secur.* **15**, 3511–3525 (2020). <https://doi.org/10.1109/tifs.2020.2975932>
10. M.H. Saad, A. Serageldin, G.I. Salama, Android spyware disease and medication, in *2015 Second International Conference on Information Security and Cyber Forensics (InfoSec)* (2015)
11. H.M. Salih, M.S. Mohammed, Spyware injection in android using fake application, in *2020 International Conference on Computer Science and Software Engineering (CSASE)* (2020). <https://doi.org/10.1109/csase48920.2020.9142101>
12. I. Shhadat, B. Bataineh, A. Hayajneh, Z.A. Al-Sharif, The use of machine learning techniques to advance the detection and classification of unknown malware. *Procedia Comput. Sci.* **170**, 917–922 (2020). <https://doi.org/10.1016/j.procs.2020.03.110>
13. L. Cai, Y. Li, Z. Xiong, JOWMDroid: android malware detection based on feature weighting with joint optimization of weight-mapping and classifier parameters. *Comput. Secur.* (2020)
14. <https://www.statista.com/statistics/680705/global-android-malware-volume/>

# An Effective Approach to Explore Vulnerabilities in Android Application and Perform Social Engineering Attack



Joseph Raymond and P. Selvaraj

**Abstract** The Android applications and smart phones are used by many customers globally for personal and professional purpose. The migration from personal computers and laptops to smart phones has increased to a greater impact and more will be in future. In our paper, as part of social engineering attack in cyber security, we have done research on taking social app to modify the code using reverse engineering and explore its vulnerabilities. We have done privilege escalation of a victim by doing reverse engineering attack. We have technologies like Kali Linux, Metasploit and Genymotion for the implementation purpose. The goal of this paper is to throw limelight on how various kinds of social engineering attacks are done and make a study to countermeasure such kind of threat that can come to human society.

**Keywords** Cyber security · Android applications · Reverse engineering and activity monitoring

## 1 Introduction

The number of android malwares is increasing rapidly with the increase in the usage of the computing systems and gain unauthorized access to the computer systems and networks resulting in severe damage and consequences. The malware can be of types such as worms, virus and backdoors. There are techniques and methods to analyze and classify these types of malware. Static inspection of malware is done without execution of the code, whereas dynamic inspection is done by its execution in a sandboxed environment. In static analysis, the properties such as hash signatures and file size can be determined, whereas in dynamic analysis the details of the file system, registry keys, memory activities, etc., can be collected.

---

J. Raymond (✉) · P. Selvaraj  
SRM Institute of Science and Technology, Kattankulathur, India  
e-mail: [josephrv@srmist.edu.in](mailto:josephrv@srmist.edu.in)

P. Selvaraj  
e-mail: [selvarajp@srmist.edu.in](mailto:selvarajp@srmist.edu.in)

In [1], the authors have reasoned about install decision in mobile application and its privacy issues. They have collected low privacy-related fictional app and suggested improvements and thus safeguarding against external threat. The static analysis provides the basic information such as the file properties and hash signatures, whereas the dynamic analysis gives information about the runtime behavior of the malware. Dynamic analysis is more effective than static analysis as it can be used for the study of the android malware.

The further study of the malware tools and techniques has been discussed in [2] where the authors provide an in-depth discussion especially on the tools available for memory forensics, packet analysis, etc. The combination of the static and dynamic features can be used for the effective analysis of the malware as mentioned in [3]. This integrated system has been proven to give more accurate results than individually carrying out the static and dynamic analyses. The dynamic analysis has been performed using Kali Linux and Genymotion.

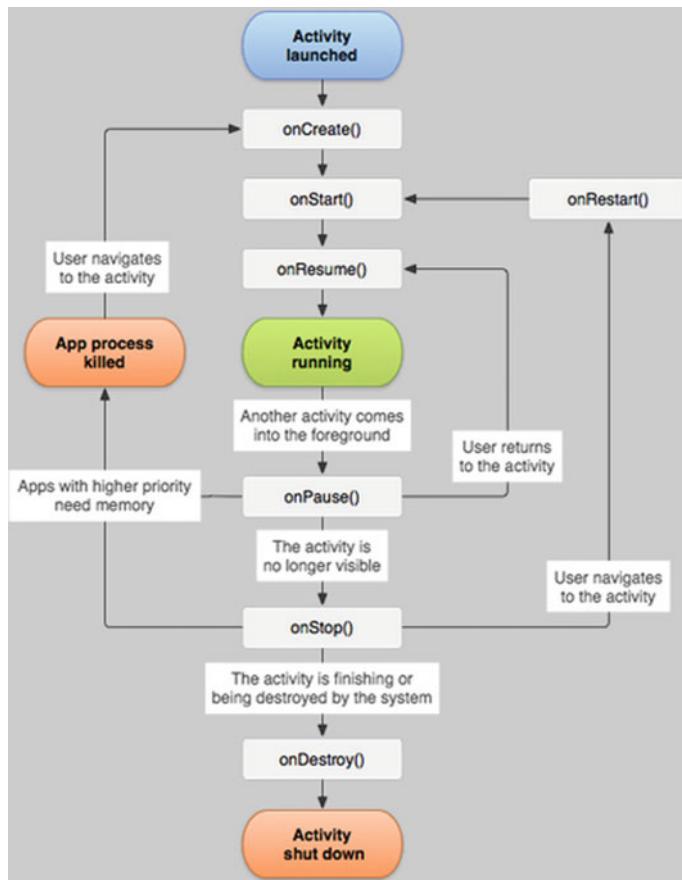
In [4–6], the authors have discussed Android signature-based and heuristic-based malware methods. The dynamic analysis involves feature extraction of file systems, registry networks, etc. The hybrid analysis involves combination of static and dynamic features such as import functions and call functions. The paper also discussed about the ways in which the malware obfuscation can take place, and the future works involve the use of memory-based artifacts for better malware analysis and classification. The obfuscation can take place in the form of classes, strings, variables, etc. The static, dynamic and hybrid analyses can be studied further, and the performance metric of each of the analysis types can be studied.

In [7], the authors discuss the knowledge mining of android application development downloading from App store. This reference helps us in gaining knowledge about vulnerable app and decoding sensitive information from the application.

In [8], the authors focus on gathering information about m-health App and find security and its countermeasures. The chronicle conditions are also focused, and effective mining process is done.

## 2 Terminologies

The life cycle of Android application involves three stages: creation, resume/restart and destroy. Activities bind action with implementation by two ways: Java or Kotin. Mostly, developers prefer Java for building Android apps. As shown in Fig. 1, an Android application will first launch the main activity and perform several state transition activities as discussed earlier. OnCreate() creates the main opening followed by OnStart() that represents the first phase in the life cycle. The two main operations are running status and process killed.



**Fig. 1** Life cycle of app activity

### 3 Implementation and Result

The purpose of this paper is to modify the mAadhaar application in android application and execute as rooted device. The methodologies are implemented using two important tools: APK tool and JADX. The APK file installed by Google Play Store, and if it needs to modify and perform reverse engineering, we can get from APK monk Web site. The goal of this paper is not to perform illegal operation but to throw limelight for gaining knowledge like these types of attacks can be done in online platform.

The key factor is how cost effective we are going to implement this framework and perform static analysis. To achieve this milestone, we need Android emulator like Genymotion and support of cyber forensics operating system kali Linux. We have

tried to implement in the recent platform installed and updated previous week from repository resources. The step-by-step implementation will be explained as follows.

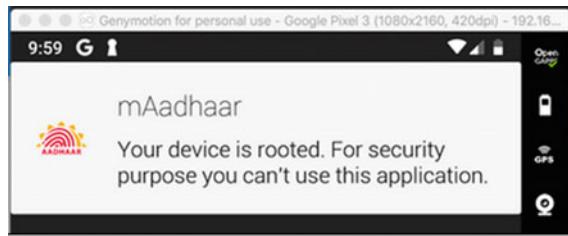
In Sect. 1, the first step is implementation of Android application, and the command is given below in figure, where adb means Android Debug Bridge used to map between the kali Linux and the emulator.

adb install in.gov.uidai.mAadhaarPlus 2018-09-26.apk

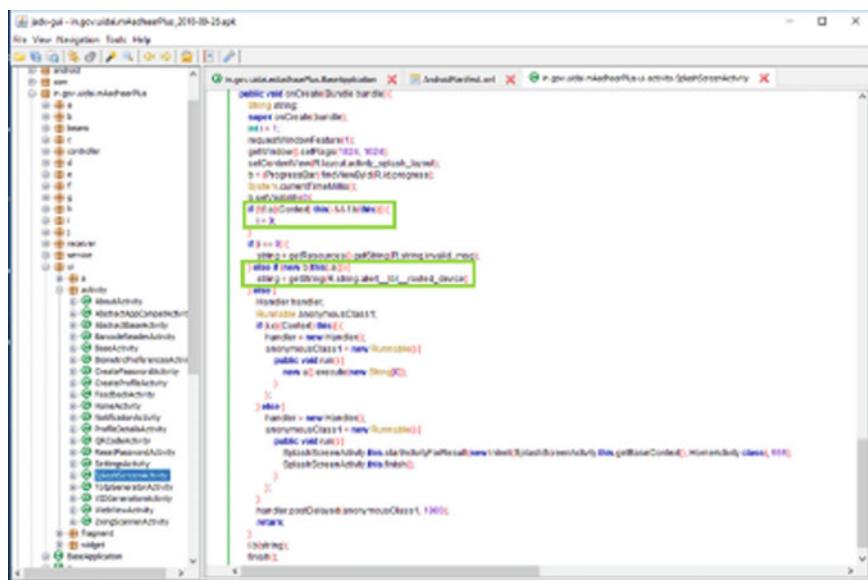
The second step is running the application in Genymotion environment. It is always recommended to install in lower version like Jelly Bean and explore more vulnerabilities. Figure 2 shows the snapshot.

The Jadex is both command line and GUI tools for producing Java source code from Android Dex and APK files. The tools look like as shown in Fig. 3 after opening manifest file and to be noticed launcher activity highlighted in green color.

**Fig. 2** M-Aadhar app



**Fig. 3** JADX command line tool



**Fig. 4** Splash activity

In Sect. 2, we mainly focus on the first major activity OnCreate () method. There are two process involved in splash screen activity: The first is integrity verification using methods f.a and f.b for modification. The second is method name b represents the rooted device. We have to disable both of these operations.

The integrity verification code to detect an unaltered app is shown in Fig. 4.

in.gov.uidai.mAadhaarPlus.ui.activity.b.f

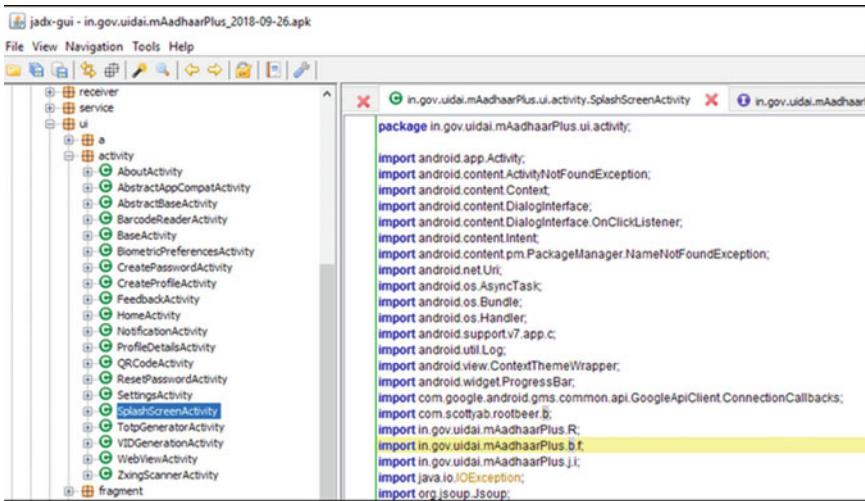
The last step to check whether it is an unaltered app using SHA-256 hash with a hard coded value as shown in Fig. 5.

In this section, we ensure the rooted device and the app is not modified, because it is to track the already affected application so gray hat hackers will always do this step before dropping the payload.

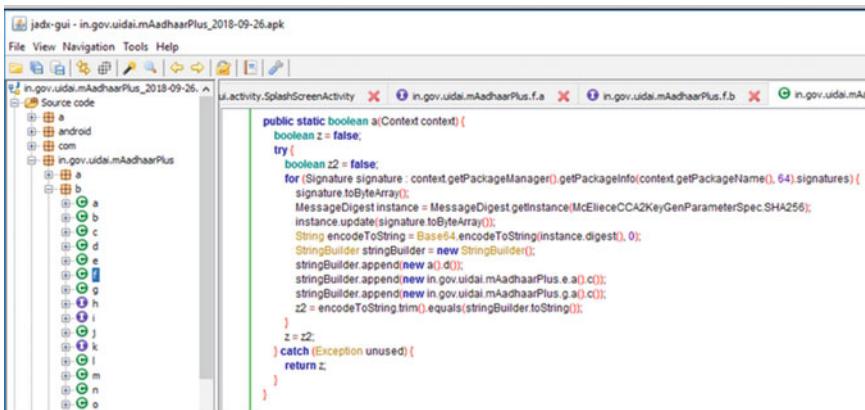
In Sect. 3, we will modify the application. The process is unpacking APK using APK tool as shown in Fig. 6.

```
apktool d -f -r in.gov.uidai.mAadhaarPlus 2018-09-26.apk
```

Now for disabling the integrity control, these steps are followed (Fig. 7)



**Fig. 5** Exploring small code



**Fig. 6** Checking SHA-256 hash

```

./in.gov.uidai.mAadhaarPlus_2018-09-26/small1/in.gov.uidai.mAadhaarPlus/ui/activity/SplashScreenActivity$1.smali:.class Lin/gov
./in.gov.uidai.mAadhaarPlus_2018-09-26/small1/in.gov.uidai.mAadhaarPlus/ui/activity/SplashScreenActivity$1.smali:.value = Li
./in.gov.uidai.mAadhaarPlus_2018-09-26/small1/in.gov.uidai.mAadhaarPlus/ui/activity/SplashScreenActivity$1.smali:.field final s
./in.gov.uidai.mAadhaarPlus_2018-09-26/small1/in.gov.uidai.mAadhaarPlus/ui/activity/SplashScreenActivity$1.smali:.method constr
./in.gov.uidai.mAadhaarPlus_2018-09-26/small1/in.gov.uidai.mAadhaarPlus/ui/activity/SplashScreenActivity$1.smali:.input-objec
./in.gov.uidai.mAadhaarPlus_2018-09-26/small1/in.gov.uidai.mAadhaarPlus/ui/activity/SplashScreenActivity$1.smali:.new-istan
./in.gov.uidai.mAadhaarPlus_2018-09-26/small1/in.gov.uidai.mAadhaarPlus/ui/activity/SplashScreenActivity$1.smali:.aget-objec
./in.gov.uidai.mAadhaarPlus_2018-09-26/small1/in.gov.uidai.mAadhaarPlus/ui/activity/SplashScreenActivity$1.smali:.invoke-dir
./in.gov.uidai.mAadhaarPlus_2018-09-26/small1/in.gov.uidai.mAadhaarPlus/ui/activity/SplashScreenActivity$1.smali:.invoke-vir
./in.gov.uidai.mAadhaarPlus_2018-09-26/small1/in.gov.uidai.mAadhaarPlus/ui/activity/SplashScreenActivity$1.smali:.class public Li
./in.gov.uidai.mAadhaarPlus_2018-09-26/small1/in.gov.uidai.mAadhaarPlus/ui/activity/SplashScreenActivity$1.smali:.field public st

```

**Fig. 7** Searching using grep keyword

```
grep SplashScreenActivity -r . | less -S
```

We can open the file with nano editor with admin rights or install gedit and then open the file identified through less command. We identify  $f > a$  and  $f > b$  methods as shown in Fig. 8.

We change the value of  $0 \times 0$  value to  $0 \times 1$  as shown in Fig. 9.

The next step is root detection and then disable it by identifying root bear function and commenting lines from there as shown in Fig. 10. Now the phone will run as a rooted device.

In Sect. 4, we will rebuild the app using the package name as shown in Fig. 11. We will use key tool to sign certificate for the app because it is mandatory to install the modified APK file.

```
keytool -genkey -v -keystore my-release-key.keystore -alias alias_name -keyalg RSA -keysize 2048 -validity 10000
```

```
invoke-static {p0}, Llin/gov/uidai/mAadhaarPlus/b/f;->a(Landroid/content/Context;)Z
move-result v0
if-eqz v0, :cond_0
invoke-static {p0}, Llin/gov/uidai/mAadhaarPlus/b/f;->b(Landroid/content/Context;)Z
move-result v0
if-eqz v0, :cond_0
goto :goto_0
:cond_0
const/4 p1, 0x0
```

**Fig. 8** Searching  $f1 \rightarrow a$  and  $f1 \rightarrow b$  methods

```
invoke-static {p0}, Llin/gov/uidai/mAadhaarPlus/b/f;->a(Landroid/content/Context;)Z
move-result v0
if-eqz v0, :cond_0
invoke-static {p0}, Llin/gov/uidai/mAadhaarPlus/b/f;->b(Landroid/content/Context;)Z
move-result v0
if-eqz v0, :cond_0
goto :goto_0
:cond_0
const/4 p1, 0x1
```

**Fig. 9** Modification code

```

invoke-static {p0}, Llin/gov/uidai/mAadhaarPlus/b/f;-->b(Landroid/content/Context;)Z
move-result v0

if-eqz v0, :cond_0
goto :goto_0

:cond_0
const/4 p1, 0x1

:goto_0
if-eqz p1, :cond_3

new-instance p1, Lcom/scottyab/rootbeer/b;

```

**Fig. 10** Checking conditions

```

root@kali:~/apk/maad# apktool b in.gov.uidai.mAadhaarPlus_2018-09-26
I: Using Apktool 2.3.3-dirty
I: Checking whether sources has changed...
I: Smaling smali folder into classes.dex...
I: Checking whether resources has changed...
I: Copying raw resources...
I: Copying libs... (/lib)
I: Building apk file...
I: Copying unknown files/dir...
I: Built apk...
root@kali:~/apk/maad# 

```

**Fig. 11** Compiling APK file

```

root@kali:~/apk/maad# adb install in.gov.uidai.mAadhaarPlus_2018-09-26/dist/
in.gov.uidai.mAadhaarPlus_2018-09-26.apk
Success
root@kali:~/apk/maad# 

```

**Fig. 12** Installing APK file

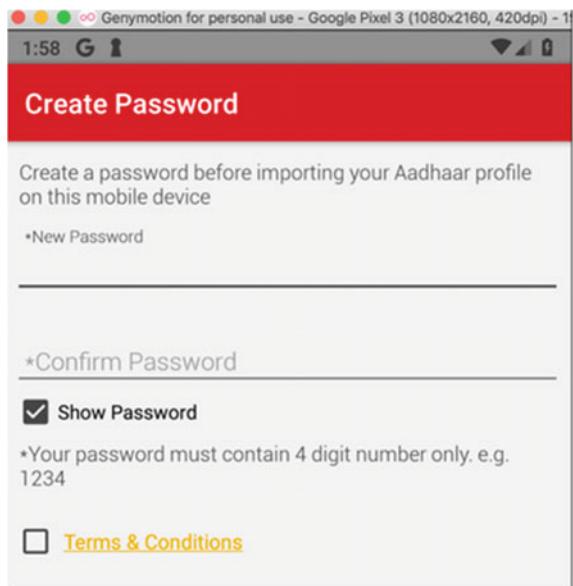
We have uninstalled the app installed before and then installed the modified application as shown in Fig. 12.

After launching the app, the user is prompted to make a phone call, and thus attacker takes privilege over the victim. Figure 13 shows how the integrity is broken down.

## 4 Conclusion and Future Work

In this paper, we confirm decision-making approach from a vulnerable app and discuss key finding with respect to research and major finding is user risk perception.

**Fig. 13** Modified output APK file



This idea complements privacy information communication tool and more privacy related to decisions.

## References

1. S.W. Tay, P.S. Teh, S.J. Payne, Reasoning about privacy in mobile application install decisions: risk perception and framing. *Int. J. Human-Comput. Stud.* **145**, 102517
2. L. Meftah, R. Rouvoy, I. Chrisment, Empowering mobile crowdsourcing apps with user privacy control. *J. Parallel Distrib. Comput.* **147**, 1–15 (2020)
3. C. Guo, D. Huang, N. Dong, J. Zhang, J. Xu, Callback2Vec: callback-aware hierarchical embedding for mobile application. *Inf. Sci.* **542**, 131–155
4. S. Ramesh, C. Yaashwanth, B.A. Muthukrishnan, Machine learning approach for secure communication in wireless video sensor networks against denial-of-service attacks. *Int. J. Commun. Syst.* **33**(12) (2019). <https://doi.org/10.1002/dac.4073>
5. T.N. Nguyen, B. Liu, N.P. Nguyen, J. Chou, Cyber security of smart grid: attacks and defenses, in *ICC 2020—2020 IEEE International Conference on Communications (ICC), Dublin, Ireland* (2020), pp. 1–6. <https://doi.org/10.1109/ICC40277.2020.9148850>
6. J. Yan, H. Zhou, X. Deng, P. Wang, R. Yan, J. Yan, J. Zhang, Efficient testing of GUI applications by event sequence reduction. *Sci. Comput. Program.* **201**, 102522
7. S. Gao, L. Liu, Y. Liu, H. Liu, Y. Wang, API recommendation for the development of android app features based on the knowledge mined from app stores. *Sci. Comput. Program.* **102556** (2020)
8. J. Lorca-Cabrera, R. Martí-Arques, N. Albacar-Riobóo, L. Raigal-Aran, J. Roldan-Merino, C. Ferré-Grau, Mobile applications for caregivers of individuals with chronic conditions and/or diseases: quantitative content analysis. *Int. J. Med. Inform.* **104310** (2020)

# History of Deception Detection Techniques



D. Viji, Nikita Gupta, and Kunal H. Parekh

**Abstract** Deception among humans has always existed. It has now become a part of our nature and has led to the research of various deception detection techniques. The longing for lie detection has been a never-ending age-old practice. The techniques have varied from measuring physical values of the person's body (heart rate, body temperature) to analysing the small expressions that the eyes made to micro-changes of the facial muscles, frequency of the utterance of the words, etc. In our survey, we have covered various techniques that have been employed in professional practice to techniques which are still developing and can show promising results in the near future. These techniques are widely useful in legal and lawsuits, police offices, crime and court trials. From the study of polygraph which dominated the twentieth century to facial and speech imaging techniques which are now widely used for various domains, the paper presents a comprehensive study of all the recognized techniques in lie detection and comparison among the different approaches. Finally, the paper summarizes the deception techniques and provides discussion on the existing problems and trends.

**Keywords** Lie detection · Polygraph · Multimodal deception detection · Psychology

---

D. Viji (✉) · N. Gupta · K. H. Parekh

Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur 603203, India

e-mail: [viji.d@ktr.srmuniv.ac.in](mailto:viji.d@ktr.srmuniv.ac.in)

N. Gupta

e-mail: [np4047@srmist.edu.in](mailto:np4047@srmist.edu.in)

K. H. Parekh

e-mail: [kh4061@srmist.edu.in](mailto:kh4061@srmist.edu.in)

## 1 Introduction

Deception has a simple meaning called misleading or hiding the truth. Since the beginning of human race, there have been instances where people have tended to lie in their everyday life. From small lies to big lies, lying has become a part of our nature. Though it is not considered a big problem, some lies can lead to serious consequences where it cannot be ignored. Deception is greatly concerned with the moral sense. The principle of detection technique is related to guilty conscience. The person who is lying must put more efforts to cover up his/her made-up story. This can stimulate the nervous responses leading to sudden changes in physiological values which forms the key for lie detection. Detecting deception has remained a great topic of interest for psychologists and legal professionals which spread among the researcher's community and ultimately led them to work on deception detection techniques. From the age-old era where non-scientific deception techniques were in practice to the new era of scientific knowledge where rational and calculated deception techniques have replaced the old ones, lie detection has continued to be a matter of concern and significance. The non-scientific practices like "Agni Pariksha", "Rice Test", "Bird's Egg", etc., [1] have been now overtaken by scientific techniques like polygraph, ADDs, EEG, PPG, thermal sensing, etc.

The deception detection has been actively used in law enforcements, court trials, police and other legal offices. From the invention of polygraph in twentieth century, this technique has ruled and overpowered others for a long time. Polygraph is a simple technique which records breathing, pulse, blood pressure, respiration and electrical resistance.

These values are recorded by attaching the appropriate sensors on the candidate's body. The candidate is put through high cognitive demanding questions, and his responses on the sensors are noted. This is done under the presence of a highly skilled examinee to evaluate the results and give weights on the attributes as concerned. It is acquired knowledge that when a person is put under stress or emotional stimuli, it leads to physiological changes in the body. A lot of techniques came into picture, some made use of sensors to detect changes, and the others were non-touch techniques using thermal, facial, speech imaging, blink rate calculation, etc. The researchers have shed light on the facial expressions and writing styles. A lot can be interpreted by minutely noting the facial expressions, utterances of the words, writing styles and micro-facial features. By focusing the pupils, shape of the eyebrow, perinasal region and lips a lot of facial features can be extracted. With the advancement in machine learning techniques, feature extraction and classification between liars and truth tellers have become easier and more accurate.

The ability to correctly distinguish between the innocent and a liar is a significant skill to be possessed by the investigators. There is always a great risk accompanied with the deception detection in crime cases because an innocent person can be given wrong judgement. Though a lot of progress has been shown and promising results have been obtained from the existing lie detection techniques, there are always some

error rates involved. Without the supervision of a highly skilled examinee, no final judgement has been passed yet in any lawsuit.

## 2 Related Work

The longing for lie detection has been a never-ending age-old practice. The techniques have varied from measuring physical values of the person's body (heart rate, body temperature) to analysing the small expressions that the eyes made to micro-changes of the facial muscles, frequency of the utterance of the words, etc. In our survey, we have covered various techniques that have been employed in professional practice to techniques which are still developing as well as can show signs of promising results in the near future.

Comprehensive studies in the field of psychology have shown that blink rate of a person's eye can reveal a considerable decrease when faced with high technical conscious situations [2, 3]. Someone who is going to lie will have to focus more on his made-up story or answer so that it is not suspected to be false and do not lead to any kind of ambiguity. In this process, there is an increase in the blink rate when the person is fabricating a lie followed by a sudden decrease in the blinking rate [4]. In [5], the authors have proposed a method of detecting whether a person is lying or not by calculating the differences in their eye blink patterns from the threshold blink rate to the target period blink rate. The high cognitive demanding questions are asked followed by low cognitive demanding questions, and the blink rate is noted. This period is called as the target period, and thus, blink rate during this period is called target period blink rate followed by post-target period blink rate. This gives rise to two cases—(1) if there is significant difference in the threshold and target period blink rate, then the person is lying. (2) If there is no significant difference in the two, then the person is telling the truth. Threshold blink rate is either set as "26 Blinks per min," founded by Bentivoglio et al. in [6] or by taking the average blink rate mentioned above. This technique [5] is not limited to only identifying lie detection but can also be used for testing fatigue, sleep driving, physical eye-related tests, Parkinson's disease, etc.

From the given knowledge of involuntary physiological changes happening due to high stress situations, it led to further research on physical values. Srivastava and Dubey [7] proposed a method using vocal features along with physical values to predict the lie detection using ANN and SVM.

The data set used here consisted of 50 subjects. The responses of polar questions having yes, or no answers were recorded. During the questionnaire, for physiological values it registered heart rate, blood pressure and breathing rate. From the responses (speech signals of subjects), features like basic frequency, zero crossing rate, function of frames and energy were extracted [8, 9] which further calculated Mel frequency cepstral coefficient. The features for each subject are extracted, concatenated, classified, error corrected and fed to two classification algorithms. ANN model with one hidden layer consisting of 20 neurons achieved 93% accuracy, whereas SVM

**Table 1** Polygraph CQT and CIT

Sensitivity and specificity of the polygraph		
Study type and measure	Sensitivity (%)	Specificity (%)
<i>CQT—polygraph</i>		
Laboratory studies	74–82	60–83
Field studies	84–89	59–75
<i>CIT—polygraph</i>		
Laboratory studies	76–88	83–97
Field studies	42–76	94–98

achieved 100% accuracy. Though the accuracy ranged high, the limitations laid inside the small data set.

From the twentieth century, polygraph has remained the most used scientific technique in law enforcement, court trials and offices. Polygraph was invented by Dr. John A Larson; it is nothing but the collective reading of various physiological stimuli one experiences under pressure or emotion. The principle of action is based on the characteristic changes of blood pressure, pulse rate, respiration, electrical resistance and sweating which is measured using GSR [10]. Different categories ranging from relevant to irrelevant and balanced to stress questions are asked. The response is inspected by a highly skilled examiner. It is under his control to provide different measure weights to different attribute changes. Generally, numerical scoring approach is used, positive scores for truth and negative scores for deceiving responses. In the end, total score is calculated.

The control question test (CQT) and concealed information test (CIT) remains one of the most reviewed and approved methods available for polygraph [11]. In Table 1, the sensitivity and specificity of polygraphs are noted for laboratory results and field results. The data of Table 1 is taken from Meijer and Verschueren (2015). When CQT and CIT are compared, CQT gives better results for liars and CIT gives better results for truth tellers. Though these give good accuracies, there are several countermeasures which can be used to deceive the polygraph results. Someone who is skilled can handle physical or emotional stimuli by pressing one's toe, doing mental calculations, etc., and deceive polygraph tests. Due to this reason, polygraph test alone cannot be used as the lie detector test yet.

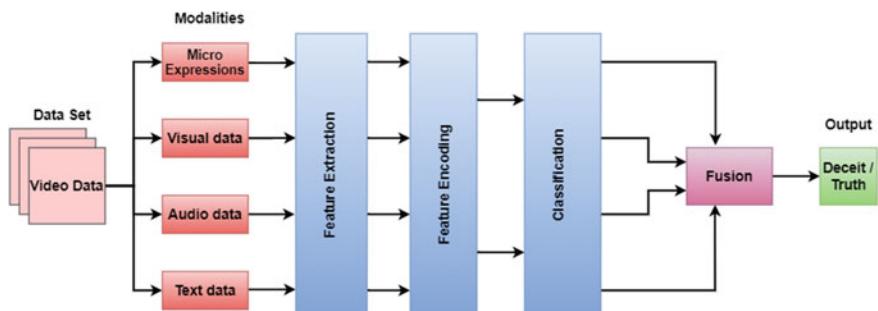
Thermal imaging is emerging as one of the most promising techniques for non-contact deception detection [12, 13]. Derakhshan et al. [14] proposed an algorithm which employed thermal infrared imaging technique. In [15], the authors have shown light on the thermal imaging approach. A comparative study and implementation are performed among thermal, GSR and PPG methods. The data set consisted of 41 subjects named THEPHY interviewed under two scenarios—mock crime and best friend scenario. The GSR and PPG response is noted down as usual. The thermal imaging is captured using a high-quality and high sensitivity thermal imager. Statistical features were implemented along with feature reduction techniques to recognize the best ROIs (“periorbital, forehead, cheek, perinasal and chin”).

Four classifiers, namely SVM, LDA, KNN and decision trees, were employed. The best results were achieved by ROC and decision trees with perinasal and chin attributes to be the most dominant for thermal changes. The highest accuracies obtained were 90.1% in mock crime and 74.7% in best friend plots for thermal technique. The thermal imaging results outperformed the GCR as well as PPG results. This research also gave way to the technique which did not rely on sensors being attached on the candidate.

In the next category, electrical signals on brain cells are utilized to analyse the lie detection capabilities of a person [16, 17]. In [18], the authors have worked on the principle of electroencephalogram (EEG) to identify lie, emotion and attention detection. The electrical activity of brain is captured using MUSE 2 headband device. The data set consisted of 30 healthy adults. The brain activity is classified into five categories of ranges as  $\delta$ (1–4 Hz),  $\tau$ (4–8 Hz),  $\alpha$ (8–12 Hz),  $\beta$ (12–30 Hz) and  $\gamma$ (>25 Hz). The data set is obtained differently for each category. For lie detection, binary answer questions are asked. For emotion detection, varieties of video fragments are shown. For attention detection, focusing-based tests are conducted. Features are extracted, and with the help of classification algorithms, lie and attention detection are binary classified and emotion detection is multiclass classified. The classification algorithms employed are KNN, logistic regression, CNN and random forest. The performance of the algorithm is average and has a major constraint on the subjective scenario as well as limited electrodes on the MUSE 2 device. There is lot of scope of improvement under this technique.

After the era of hardware sensors and thermal imaging came the approaches which relied on non-contact and more calculation-based techniques. Unimodal method came into use in which features like facial expressions, text, speech or micro-expressions were analysed to recognize deception. Furthermore, multimodal system which combines two or more unimodal features to detect deception came into the picture. In Fig. 1, we have shown a flowchart depicting the general algorithm for multimodal deception detection.

The authors [19] provide detail analysis of using facial affects along with interpretable features from modals like visual, text and speech, as describing factor for



**Fig. 1** Flowchart for multimodal deception detection

**Table 2** Performance of features and algorithms [22]

Modalities	Algorithm	CCR (%)
Audio	MFCC-SVM	64
	CC-SVM	66
	MFCC-LSTM	46
	CC-LSTM	58
	LE-LSTM	50
	MFCC-SRKDA	64
	<b>CC-SRKDA</b>	<b>76</b>
	LE-SRKDA	68
Text	LSTM	44
	SRKDA	68
	SVM	24
	Bag-of-words—SVM	66
	<b>Bag-of-N-grams—SVM</b>	<b>84</b>
Micro-expressions	<b>AdaBoost</b>	<b>88</b>
	Random forest	82
	SVM	75

deception detection. An additional point taken into consideration is the valance [20] (whether a state is pleasant or not) and arousal [21] (activeness or passiveness of the state) which is combined into a 2D space and used to detect deception. The features were extracted, and then relevant and significant features were selected.

For classification in both unimodal and multimodal system, SVM with linear kernel was used. The classification experiments were performed with fivefold stratified cross-validation. The results for unimodal facial affect were 80% accuracy while that of visual and vocal were 80% and 83% accuracy, respectively. The multimodal of facial affect, visual and vocal features with AdaBoost was the most effective with AUC of 91%. Their results suggest the facial affect makes an effective contribution in detecting deception [19].

With this achievement, a lot of research was done on multimodal systems. The authors in [22] have used four main functional units for their multimodal system, i.e., audio, text, micro-behaviour and the fusion subsystem. At the fusion subsystem, they aim to combine results of the other three subsystems by using majority voting. Two experiments were performed: one for comparison of existing and their proposed algorithm and the other for performance of their multimodal detector. According to this, for audio subsystem CC feature performed better when compared to MFCC and log-energy of MFCC, whereas SRKDA classifier performed better when compared to linear SVM and LSTM. From Table 2 [22], we note that CC-SRKDA performed the best with 76% accuracy. For text subsystem along with the linear SVM classifier, nag-of-N-grams (BoNG) functions were the best performer with 84% accuracy. For micro-expression subsystem, AdaBoost classifier was the best with 88% accuracy.

Deep learning techniques were also explored. Krishnamurthy et al. used neural networks approach for multimodal deception detection [23]. For feature extraction, the following has been used: for visual feature extraction 3D-CNN [24], for textual features extraction CNN [25–27] and for audio feature extraction OpenSMILE along with SoX (Sound eXchange) for background noise removal, and Z-standardization for voice normalization is used [28, 29]. The fusion of results is a simple concatenation of features from all modalities into a single feature vector or a Hadamard + concatenation in which the features are fused by Hadamard product. Though their model was able to provide an accuracy of 97.99%, their model relies on a small data set and may be prone to overfitting. Since data set contains limited scenarios, the model may not perform the same for various other scenarios.

The paper [30] states the following contributions: (1) subject-level deception detection, (2) exploration of the efficacy of a wide variety of derived features. (3) in order to obtain reliable scores, the experiment was performed three times, each with an arbitrary set of leave one out cross validation of every trial sample. (4) A semi-automatic system using a combination of automatic as well as manually annotated features provides an accuracy of 83.05% while fully automatic system provides an accuracy of 73%. The features extracted are unigram and LIWC for linguistic, facial display, hand gestures for manually annotated visual, FAUs for automatic visual, pitch and silence–speech histogram for acoustics. For subject-level feature integration, firstly the maximum values for each feature were taken and then were averaged corresponding to the video of each subject. The classifiers used were random forest, SVM and neural networks [31]. The best result was obtained by using neural networks, while fusion of visual features along with acoustic features gave an accuracy of 84.18%.

In [32], the method proposed is data driven and uses visual and verbal cues. Facial features were analysed using OpenFace with facial action unit recognition, while acoustic patterns were analysed using OpenSMILE. The classifier used was support vector machine. The action units (AU's) used in OpenFace are [1, 2, 4–8, 10, 12, 13, 18, 19, 22, 24, 24, 25, 27, 33]. The model “bag-of-words” was used for verbal feature extraction. OpenSMILE [34] was used in extraction of basic features of acoustics like MFCC, jitter and harmonics to noise ratio. After all the features are extracted, they combine resulting into a single feature vector.

With the popularity of automated systems, the authors of [35] present an automatic deception detection framework which consists of three phases: (1) feature extraction in multimodal system, (2) feature encoding and (3) classification. Fisher vector encoding is employed which aggregates a variable number of features into a vector with fixed length [36]. Perez-Rosas et al. [37] show the five most distinguishing micro-expressions for detecting deception: “Frown, Head Side Turn, Lips protruded, Lip corners up and Eyebrows raised”. Wu et al. [35] test four independent features first: IDT, score of high-level micro-expression, oral features and audio features as Mel frequency cepstral coefficients and then test their combinations in multimodal system. Classifiers used were linear SVM, Kernel SVM, random forests, Naive Bayes, AdaBoost, decision trees and logistic regression. In all tests, fivefold

and tenfold cross-validations have been used across various feature sets and classifiers. As a result, linear SVM works better on IDT features, RF works better on high-level micro-expression features, and on MFCC features, Kernel SVM performs better.

Though a lot of progress is made on the efficiency of deception detection techniques, it is still not enforced in various places due to its high cost and usefulness. The technique's usefulness can be measured by finding out the base rates or cost information through Bayesian analysis, information gain, or determining ROC points [33]. By doing this, it can find more applications and can be employed more at workplaces.

### 3 Inference from the Survey

From Table 2, we observed that old research was based on the signals and physical values received from the hardware sensors attached to the candidate's body [7, 14]. This method proved to be very reliable and dominated the law and crime enforcements for their application. But soon to be found that many countermeasures were discovered to escape the reliable results of polygraph and other methods which relied on physical values. Another drawback was the inconvenience of the requirement to attach the hardware instruments on the candidate.

This led to the new approaches which relied more on the non-touch methods. The other papers discussed in the comparison Table 2 are approaches which focused on facial and speech signals without the reliance of any hardware attachment. We have found some key points which were unique and improved the performance among the similar approaches. We observed that unimodal and multimodal systems on visual, text, vocal and micro-expression features are the current trend in deception detection. Visual features (91%) [19] and micro-expressions (AUC 0.9799) [20] have been identified as features having better deception discriminating power compared to vocal (76%) [21] and text (84%) [21]. However, multimodal systems are more accurate (the best result being 91%) [19] due to combination of features which provides better accuracy and reliability if trained properly. A key point to be observed here is the data set, which is common to all. Since it is small (121 videos), appropriate measures are considered for filtering and pre-processing it in order to obtain reliable output. Methods such as fivefold and tenfold cross-validation have been used along with random seeding to make the testing and training data more reliable. The commonly used classifiers for classifying features are support vector machine with Kernel, random forest, neural networks and AdaBoost. Since different features perform better with different classifiers, thus evaluation for the best accuracy for each model is necessary.

## 4 Summary and Future Prospects

From the survey, we have analysed that the old trend relied on hardware-based sensor values. The sudden changes in the values as a result of stress or emotional stimuli helped in the observation of guilt or ambiguity in the candidate. Under a skilled examinee, right judgement could be followed. The data sets availability is low. Since most of the data collected is confidential and thus kept private by the law enforcements, most of the researches are done by performing laboratory experiments and generating their own data set. When we go further, we observed the trend shifting towards video recorded data sets. Here the facial features and verbal signals extraction were employed. Since this approach overcame one of the drawbacks of polygraph and gave promising results, it paved the way for more research and development. The data set availability still lies low due to the sensitivity of the domain.

These researches have paved the way for future exploration of various possible combinations of features to obtain a more accurate and reliable output and finding new more reliable features to detect deception. Future work may involve precise selection of distinguishing features that are more reliable for detecting deception and elimination of features with less distinguishing capability (Table 3).

## 5 Conclusion

The review of current literature highlights the shifting trend from polygraph-based techniques to facial, thermal and speech imaging techniques. For a long time, deception detection has partly relied on the supervision of human interface. No technique has given results without any error costs. With the combination of feature reduction and classification techniques, accuracies can be improved to a great extent. It is yet too soon to be said that a lie detector can be built which can function without any human interference. Though a lot of techniques and methods claim for high accuracies, many of them use small/restricted data set and minimal environment domains, hence most of the models suffer with overfitting.

As deception detection using facial features is highly dependent on data set, people from varied regions and cultures have different facial features like skin tone, expressions, etc., and can contain scars/birthmarks too. These extra features may lead to inaccurate detection. Obtaining a near universal data set, which covers all possible facial features, is a complex and tedious task as the interviews of all accused are not recorded and from the ones which are recorded are not all available for public research. The data set most frequently used for research contains 121 videos with repeating faces. Using this, model may overfit the data set and may give inaccurate results. Thus, reinforced learning can be a possible option to improve accuracy with time. Also focusing on features which are common to most of the people (nearly 90%) can help overcome this problem. However, the accuracy of observing the change in these features may vary.

**Table 3** Comparison study of different deception detection techniques

Title	Dataset used	Methodologies used	Experiment results	Limitations
Deception detection using artificial neural network and support vector machine [7]	50 subjects polar questions having yes, or no answers recorded	Speech signal and physical values recorded Classified using SVM and ANN	ANN 93% accuracy SVM 100% accuracy	Small data set. Overfitting of the model
Identifying the optimal features in multimodal deception detection [14]	Data set named THEPHY 41 subjects interviewed under two scenarios—mock crime and best friend	Thermal, GSR and PPG imaging Feature reduction techniques to identify the best ROIs Four classifiers—SVM, LDA, KNN, decision trees	90.1% for thermal technique in mock crime and 74.7% for best friend scenarios Thermal imaging outperforms GCR and PPG	More work to improve the face and ROI trackers in thermal spectra Need more unobtrusive, realistic scenario increase the number of participants
Introducing representations of facial affect in automated multimodal deception detection [19]	121 real-world trial videos (60 truthful videos, 61 deceptive videos)	Facial, verbal, visual, vocal features considered for feature extraction in unimodal and multimodal system SVM and AdaBoost used for classification	Facial affect contributed for highest-performing multimodal approach, achieved an AUC of 91% through adaptive boosting training	Switch focus among faces in the courtroom Variations in video illumination, camera angles and face sizes
Robust algorithm for multimodal deception detection [22]	Real-life courtroom trial data. 121 videos, (61—deceptive and 60—truthful)	CC features with SRKDA for audio sub-system, bag-of-N-grams for text sub-system, AdaBoost for micro-behaviour sub-system and majority voting for fusion sub-system	Best performances-CC-SRKDA with 76% for audio Bag-of-N-grams with SVM with 84% for text AdaBoost with 88% for micro-expressions	Majority voting in fusion subsystem may not provide accurate result every time. Features should be weighted

(continued)

**Table 3** (continued)

Title	Dataset used	Methodologies used	Experiment results	Limitations
A deep learning approach for multimodal deception detection [23]	121 videos of courtroom trials (60 true and 61 deceit)	3D-CNN for visual features CNN for textual features OpenSMILE, SoX, for audio features Data fusion by concatenation	Model MLPH+C obtains an AUC of 0.9799 Visual and textual features play a major role	Prone to overfitting Limited scenarios in the data set thus the model may not show the same performance for out-of-domain scenarios
Multimodal deception detection using real-life trial data [28]	Trial hearing recordings from public sources	Speech features—unigram, LIWC Annotated visual features: facial displays, hand gestures Automatically extracted FAUs Classifiers used—random forest, SVM with radial basis function Kernel and NN classifiers	RF classifier performs best for linguistic with 64.41% and acoustic features with 71.19%, NN performs best with the visual features with 80.79%. NN classifier along with visual and acoustics features gave the best accuracy of 84.18%	Modal and data set processing is based only on ground truth establishment
The truth and nothing but the truth: multimodal analysis for deception detection [30]	61 deceptive and 60 truthful videos sourced from various YouTube channels	OpenFace for visual features, bag-of-words for lexical features, OpenSMILE for acoustic features	Feature level of fusion as automatic system best accuracy of 78.95% The accuracy in truth videos 81.10% while that in deceptive videos 76.80%	Soiled data set: videos accounted for more than one person speaking Lot of noise due to public courtroom proceedings thus reduced accuracy

(continued)

**Table 3** (continued)

Title	Dataset used	Methodologies used	Experiment results	Limitations
Deception detection in videos [32]	121 court room trial video clips	IDT is applied on motion features MFCC used audio features Micro-expression prediction model is used to check its effectiveness of micro-expressions Fisher vector used features encoding	Random forest works best on high-level micro-expression features, and Kernel SVM performs best on MFCC features Visual modality provides 0.8347 AUC After late fusion transcripts and MFCC, features provide 0.9221 AUC	The data set contains only 58 identities, which is less than the number of videos, and often the same identity is either uniformly deceptive or truthful

On an optimistic note, the topic is of great significance and grabs lot of attention of the research scholars. In the recent years, tremendous growth and inventions have been achieved. The benefits in the development of these techniques not only lie in crime, court and legal offices but also in detecting drowsiness, frame of mind, health check-up, etc. Future work on this topic can be a more optimized version of the existing models. With this paper, we aim to spread awareness and likeness among the masses so that more work and development can be done, and one day we can live in a world free of crime.

## References

1. H. Goswami, A. Kakker, N. Ansari, A. Lodha, A. Pandya, The deception clues in forensic contexts: the lie detection psychology. *J. Forensic Psychol.* (2016)
2. J. Bageley, L. Manelis, Effect of awareness on an indicator of cognitive load. *Percept. Mot. Skills* **49**, 591–594 (1979)
3. L.O. Bauer, R. Goldstein, J.A. Stern, Effects of information processing demands on physiological response patterns. *Hum. Factors* **29**, 213–234 (1987)
4. S. Leal, A. Vrij, Blinking during and after lying. *J. Non Verbal Behav.* **32**(4), 187–194 (2008)
5. B. Singh, P. Rajiv, M. Chandra, Lie detection using image processing, in *2015 International Conference on Advanced Computing and Communication Systems, Coimbatore* (2015), pp. 1–5. <https://doi.org/10.1109/ICACCS.2015.7324092>
6. A.R. Bentivoglio, S.B. Bressman, E. Cassetta, D. Carretta, P. Tonali, A. Albanese, Analysis of blink patterns in normal subjects. *Mov. Disord.* **12**(6), 1028–1034 (1997)
7. N. Srivastava, S. Dubey, Deception detection using artificial neural network and support vector machine, in *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore* (2018), pp. 1205–1208. <https://doi.org/10.1109/ICECA.2018.8474706>
8. I. Fujimasa, T. Chinzei, I. Saito, Converting far infrared image information to other physiological data. *IEEE Eng. Med. BiolMagaz.* **19**, 71–76 (2000)
9. O. Lartillot, P. Toiviainen, A matlab toolbox for musical feature extraction from audio, in *International Conference Digital Audio Effects* (2007), pp. 237–244
10. A. Vrij, B. Verschuere, *Lie Detection in a Forensic Context* (Oxford bibliographies, USA, 2015)
11. J. Masip, Deception detection: state of the art and future prospects. *Psicothema—J. Scholar Metrics* **29**(2), 149–159 (2017). <https://doi.org/10.7334/psicothema2017.34>
12. A. Merla, L. Di Donato, P.M. Rossini, G.L. Romani, Emotion detection through functional infrared imaging: preliminary results. *Biomed. Tech.* **48**, 284–286 (2004)
13. A. Merla, G. Romani, Thermal signatures of emotional arousal: a functional infrared imaging study, in *Proceedings of the 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, Lyon, France* (22–26 August 2007)
14. A. Derakhshan, M. Mikaeili, T. Gedeon, A.M. Nasrabadi, Identifying the optimal features in multimodal deception detection. *Multimodal Technol. Interact.* **4**, 25 (2020)
15. D. Cardone, P. Pinti, A. Merla, Thermal infrared imaging-based computational psychophysiology for psychometrics. *Comput. Math. Methods Med.* **2015**, 984353 (2015)
16. T. Alotaiby, F.E. Abd El-Samie, S.A. Alshebeili, I. Ahmad, A review of channel selection algorithms for EEG signal processing (2015)
17. M. Sung, A.S. Pentland, Stress and lie detection through non-invasive physiological sensing. *Int. J. Biomed. Soft Comp. Human Sci.* **14**, 11–118 (2009)

18. I. Lakshan, L. Wickramasinghe, S. Disala, S. Chandrasegar, P.S. Haddela, Real time deception detection for criminal investigation, in *2019 National Information Technology Conference (NITC), Colombo, Sri Lanka* (2019), pp. 90–96. <https://doi.org/10.1109/NITC48475.2019.9114422>
19. L. Mathur, M.J. Matarić, Introducing representations of facial affects in automated multimodal deception detection, in *ICMI '20: Proceedings of the 2020 International Conference on Multimodal Interaction* (October 2020)
20. J. Russell, A circumplex model of affect. *J. Pers. Soc. Psychol.* **39**, 1161–1178 (12 1980). <https://doi.org/10.1037/h0077714>
21. H. Gunes, M. Pantic, Automatic measurement of affect in dimensional and continuous spaces: why, what, and how?, in *Proceedings of the Measuring Behavior* (01 2010). <https://doi.org/10.1145/1931344.1931356>
22. S. Venkatesh, R. Ramachandra, P. Bours, Robust algorithm for multimodal deception detection, in *2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), San Jose, CA, USA* (2019), pp. 534–537. <https://doi.org/10.1109/MIPR.2019.00108>
23. G. Krishnamurthy, N. Majumder, S. Poria, E. Cambria, A deep learning approach for multimodal deception detection, in *Accepted at the 19th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)* (2018) arXiv:1803.00344 [cs.CL]
24. S. Ji, W. Xu, M. Yang, K. Yu, 3d convolutional neural networks for human action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 221–231 (2013)
25. N.T. Le, J.-W. Wang, C.-C. Wang, T.N. Nguyen, Novel framework based on HOSVD for Ski goggles defect detection and classification. *Sensors* **19**, 5538 (2019). <https://doi.org/10.3390/s19245538>
26. M. Balaanand, N. Karthikeyan, S. Karthik, R. Varatharajan, G. Manogaran, C.B. Sivaparthipan, An enhanced graph-based semi-supervised learning algorithm to detect fake users on Twitter. *J. Supercomput.* **75**(9), 6085–6105 (2019). <https://doi.org/10.1007/s11227-019-02948-w>
27. Y. Kim, Convolutional neural networks for sentence classification (2014). arXiv:1408.5882
28. F. Eyben, F. Weninger, F. Gross, B. Schuller, Recent developments in openSMILE, the munich open-source multimedia feature extractor, in *Proceedings of the 21st ACM International Conference on Multimedia. MM '13, New York, NY, USA, ACM* (2013), pp. 835–838
29. L. Norskog, Sound exchange (1991). <http://sox.sourceforge.net/>
30. U.M. Sen, V. Perez-Rosas, B. Yanikoglu, M. Abouelenien, M. Burzo, R. Mihalcea, Multimodal deception detection using real-life trial data. *IEEE Trans. Affect. Comput.* <https://doi.org/10.1109/TAFFC.2020.3015684>
31. A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, A. Lerer, Automatic differentiation in PyTorch, in *NIPS Autodiff Workshop* (2017)
32. M. Jaiswal, S. Tabibou, R. Bajpai, The truth and nothing but the truth: multimodal analysis for deception detection, in *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)*. <https://doi.org/10.1109/ICDMW.2016.0137>
33. D.P. Twitchell, C.M. Fuller, Advancing the assessment of automated deception detection systems: incorporating base rate and cost into system evaluation. *Inf. Syst. J.* **29**(3), 738–767 (2019). <https://doi.org/10.1111/isj.12231>
34. V. Perez-Rosas, M. Abouelenien, R. Mihalcea, Y. Xiao, C. Florian Eyben, F. Weninger, F. Gross, B. Schuller, Recent developments in openSMILE, the munich open-source multimedia feature extractor, in *Proceedings of the ACM Multimedia (MM), Barcelona, Spain, ACM* (2013), pp. 835–838. ISBN 978-1-4503-2404-5
35. Z. Wu, B. Singh, L.S. Davis, V.S. Subrahmanian, Deception detection in videos, in *Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*. arXiv:1712.04415 [cs.AI]
36. X. Han, B. Singh, V. Morariu, L.S. Davis, Vrfp: on-the-fly video retrieval using web images and fast fisher vector products. *IEEE Trans. Multimedia* (2017)

37. V. Perez-Rosas, M. Abouelenien, R. Mihalcea, M. Burzo, Deception detection using real-life trial data, in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction* (ACM, 2015), pp. 59–66

# Climate Change Misinformation Detection System



Sagar Saxena and K. Nimala

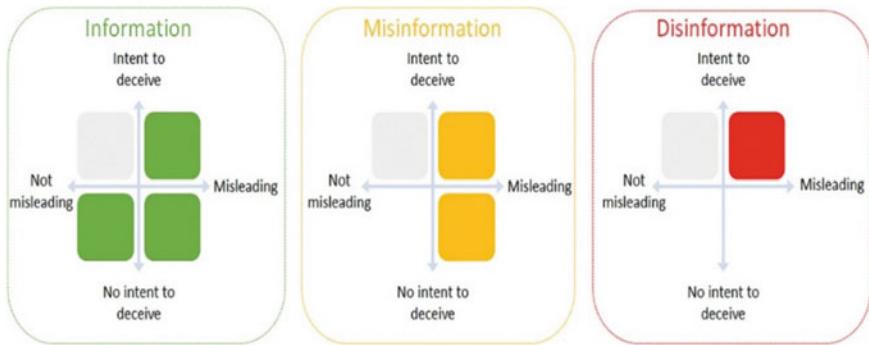
**Abstract** While there is overwhelming consistent simultaneousness on natural change, everyone has gotten enthralled over vital requests, for instance, human caused an overall temperature adjustment. Correspondence strategies to reduce polarization only here and there address the crucial explanation: insightfully decided misrepresentation scattered through sources, for instance, social and conventional press. To suitably counter online misrepresentation, we require fundamental structures that give broad understanding of the techniques used in environment trickiness, similarly as encourage verification-based approaches to manage slaughtering misleading substance. There has been an extraordinary climb in the quantity of text reports that need a more noticeable understanding of AI strategies to have the choice to viably portray the compositions in various applications. The achievement of applying diverse AI figuring depends upon how they can fathom complex models and non-straight associations inside the data. This endeavor inspects the potential strategies that can be used for text game plan of natural change trickiness. The test was to manufacture a system that recognizes whether a record contains ecological change duplicity. The goal was to achieve 0.65  $f1\_score$  for mark request. TF-IDF model was used to make features for combined getting ready and external data while for name portrayal pre-arranged models were used on the test dataset. The fact is to develop a system which stops the unpreventable of false information among online media stage customer.

**Keywords** Climate change · Misinformation · Disinformation

---

S. Saxena (✉) · K. Nimala  
SRM Institute of Science and Technology, Kattankulathur, India

K. Nimala  
e-mail: [nimalak@srmist.edu.in](mailto:nimalak@srmist.edu.in)



**Fig. 1** Difference between information, misinformation and disinformation

## 1 Introduction

In the world where information is available everywhere, it becomes very important to find whether the information provided is misclassified or not. This project is about building a machine learning approach that detects whether the climate change information is misclassified or not for the unlabeled data. Climate change misinformation task is to identify whether the climate change information is misclassified or not for the given dataset. The application of climate change misinformation is to identify the misinformation related that is regularly being distributed on mainstream and social media. In this project, we have been given the data with their text and positive label (meaning information misclassified) as a training data, and we need to predict the labels from the test data.

To predict the label of the text, we implemented different machine learning classification models. Classification models are used to identify category of the new observations based on the training set of data containing observations whose category is known but since the training data had only one label, so we did some web scraping to find external data that will be included with the training data to make it a binary supervised classification problem.

Classification models include linear models like SVM and nonlinear ones like K nearest neighbor, Naive Bayes. Difference between information, misinformation and disinformation is shown in Fig. 1.

## 2 Survey of Literature

Territory and attribution of biological change fuse surveying the reasons behind observed changes in the atmosphere framework through an exact evaluation of climate models and perceptions utilizing unmistakable quantifiable methodologies. Recognizing verification and attribution scrutinize are colossal for various reasons.

For instance, such assessments can help pick if a human effect on air factors (for instance, temperature) can be seen from standard instability. Recognizing evidence and attribution studies can help overview whether model ages are trustworthy with seen subjects or different changes in the environmental framework. Results from region and attribution studies can incite dynamic on climate methodology and assortment. There are a couple of general kinds of acknowledgment and attribution considered, including attribution of examples or long stretch changes in environmental factors; attribution of changes in cutoff points; attribution of atmosphere or climate events; attribution of air-related impacts; and the evaluation of air affectability using observational prerequisites. Paleoclimate mediators can moreover be useful for recognizable proof and attribution analyses, particularly to give a more broadened term perspective on environment capriciousness as a measure on which to take a gander at late air changes of the earlier century or something to that effect. Distinguishing proof and attribution studies should be conceivable at various scales, from worldwide to nearby.

Two of the main headways affecting US legislative issues are (a) the creating effect of private generosity and (b) the tremendous extension creation and scattering of lie. Notwithstanding their hugeness, the associations between these two examples have not been sensibly examined. This examination uses a refined assessment plan on a colossal grouping of new data, utilizing normal language taking care of and unpleasant string planning to review the association between the tremendous extension climate misrepresentation advancement and US noble cause.

The examination finds that over a twenty-year time period, associations of performers pronouncing intelligent misdirection about natural change were dynamically joined into the foundation of US charitableness. The degree of joining is foreseen by sponsoring associations with prominent corporate suppliers. These revelations uncover new data about large-scale tries to ravage public understanding of science and sow polarization. The assessment moreover contributes a fascinating computational approach to manage to be applied at this relentlessly huge, yet methodologically loaded, zone of investigation [1].

Policymakers, analysts, and specialists have all called attention to the issue of misrepresentation in the ecological change exchange. However, what is ecological change trickiness, who is incorporated, how might it spread, why does it have any kind of effect, and what should be conceivable about it? Ecological change double dealing is solidly associated with natural change uncertainty, repudiation, and contrarianism. An association of performers is locked in with financing, making, and upgrading double dealing. Once in the public space, characteristics of online casual networks, for instance, homophily, polarization, and resonance loads—ascribes in like manner found in ecological change chat—give productive ground to misdirection to spread. Principal conviction systems and typical practices, similarly as mental heuristics, for instance, certification tendency, are further factors which add to the spread of trickiness. An arrangement of ways to deal with fathom and address misdirection, from an assortment of controls, is discussed. These fuse educational, mechanical, managerial, and psychological-based approaches. No single system keeps an eye on all stresses over misrepresentation, and all have obstacles, requiring an interdisciplinary method

to manage tackle this multifaceted issue. Key investigation openings fuse understanding the scattering of ecological change misrepresentation through electronic media and reviewing whether misdirection contacts environment alarmism, similarly as climate refusal. This article researches the thoughts of trickery and disinformation and portrays disinformation to be a subset of lie. An assortment of disciplinary and interdisciplinary composing is reviewed to totally analyze the possibility of trickery—and inside this, disinformation—particularly as per ecological change [2].

While there is overpowering logical concurrence on environmental change, people in general have gotten energized over basic inquiries, for example, human-caused an Earth-wide temperature boost correspondence framework to diminish polarization on occasion addresses the fundamental explanation: thoughtfully decided lie scattered through sources, for instance, social and setup press. To effectively counter online trickiness, we require essential frameworks that give a total understanding of the methods used in climate deception, similarly as encourage confirmation-based approaches to manage slaughtering misinforming content. This part reviews assessments of environment lie, showing an extent of denialist conflicts and mysteries. Perceiving and deconstructing these different kinds of conflicts are essential to set up legitimate intercessions that effectively execute the misrepresentation [3].

Despite the fact that there is a vast comprehension among air scientists that overall natural change is certifiable, achieved by human activity, and a troublesome issue that sabotages our future, well-known evaluation is more mixed. The vast majority of everyone perceives that ecological change is authentic, yet a sizable minority is farfetched or denies natural change, and there is the lower public understanding for the presence of human-caused natural change. These revelations may be required somewhat since the natural change has become a significantly politicized issue with specific conservatives denying reality, cause, or truth of ecological change [4–6].

Duplicity can attack a well-working famous government. For example, public misinterpretations about natural change can incite cut down affirmation of the reality of ecological change and cut down assistance for help techniques. This assessment likely examined the impact of misrepresentation about ecological change and attempted a couple of preemptive interventions proposed to diminish the effect of misdirection. We found that false balance media consideration (giving foe sees identical voice with environment specialists) cut down clear understanding as a rule, regardless of the way that the effect was more noticeable among unregulated economy partners. Essentially, a misrepresentation that puzzles people about the level of coherent course of action concerning anthropogenic a perilous environmental deviation (AGW) had a polarizing sway, with unregulated economy partners reducing their affirmation of AGW and those with low unhindered economy maintain growing their affirmation of AGW. Regardless, we found that inoculating messages that (a) explain the blemished argumentation strategy used in the double dealing or that (b) include the sensible concurrence on natural change was effective in slaughtering those hostile effects of deception. We recommend that environment correspondence messages ought to consider habits by which intelligent substance can be distorted and join preemptive immunization messages [7].

### 3 Data Description

Three datasets were given for this project that is train.json, dev.json, and test-unlabelled.json data.

#### 3.1 *train.json*

The training data consists of articles with climate change misinformation, i.e., only (positive labels).

#### 3.2 *dev.json*

The dev.json file contains both labels (0 and 1) which will be used for measuring performance of the models and finding optimal hyperparameter.

#### 3.3 *test-unlabelled.json*

The test.json file contains text for which we need to predict labels using our model.

#### 3.4 *newtrain.json*

Since training data contains only positive label and for binary classification multi-labels are required, so I read the dev.json file with (0 label) and searched for articles that are related to (0 label) and made an external data.json file and combined it with the train.json to make a new json called newtrain.json which will be used for binary classification.

## 4 Background

In this section, we discuss the features and techniques that were adopted for the development of our system.

## **4.1 Feature Engineering**

We used text data to extract basic features. The features are as follows.

### **4.1.1 Stop Words Removal**

Stop words (they are the most common words that do not provide context in the language) they are processed and filtered out from the given text as they are of no use. Stop words act like connectors to the sentence, for example, conjunctions like “and, or,” articles like “a, an, and the” and prepositions. Stop words like these are less important as they contribute to valuable processing time. NLTK library was used to remove stop words, and I performed tokenization and lemmatization to remove that. The first thing when we do data pre-processing is to remove the stop words, but sometimes it helps us in gaining extra knowledge.

### **4.1.2 Removal of Punctuation**

Punctuation provides grammatical context to the sentence. Punctuations like comma might not add much value in understanding the sentence meaning. I removed all the punctuation from the text as it helps in reducing the size of training data.

### **4.1.3 Converting All to Lowercase Words**

I converted each word to lowercase because lowercase conversion of word helps in classifying the words that are present in uppercase as well as lowercase forms which can result in having words that are similar in the feature space. For example, in an article word like “Climate” might be given in uppercase or in lowercase form like “climate” which will be present in feature space. Both of them means the same, so conversion to lowercase was done so that we can classify these type of words as same.

### **4.1.4 Lemmatization**

It processes the text and groups together in the inflected forms of the word for analysis as a single item and returns dictionary form of word. Simply, it can be said that lemmatization is the conversion of words in the text to their base form that are present in the dictionary. Lemmatization was done so that I can see what the intended meaning of the word is rather than looking at literal meaning.

#### 4.1.5 Bag of Words

It processes each text as document and calculates the count of each word which creates word count vector, also called vector features of fixed length. It is simplifying the representation used in NLP. We achieved bag of words by using a Library YAKE (Yet another keyword extractor).

**YAKE:** It is an unsupervised keyword extractor which helps us in extracting keywords and provides their significance in the text. It helped me in generating slightly better results than the normal feature selection.

### 4.2 TF-IDF Vectorizer

It stands for term frequency and inverse document frequency for feature extraction. Term frequency and inverse document frequency are two components of TF-IDF. Term frequency is calculated using this methodology, and this frequency of the term tells us about the local importance of that term in the text. IDF tells us about the signature words which is important for the document and has high frequency but is not a common word in the text. This is used for filtering the stop words and for the creation of the features.

**Term Frequency:** Eq. 1 measures the frequency of the term occurring in the document.

$$\text{TF}(t) = \frac{\text{Number of times } t \text{ appear in a document}}{\text{Total Number of Terms in the document}} \quad (1)$$

**Inverse Document Frequency:** Eq. 2 measures how important the term is in the document.

$$\text{IDF}(t) = \log_e \frac{\text{Total numbers of terms in the document}}{\text{Numbers of document with } t \text{ in it}} \quad (2)$$

## 5 Methodology of Proposed System

### 5.1 Algorithm Used

This section discusses the challenges that were faced and the different approaches that were taken to perform the analysis.

### 5.1.1 Naive Bayes Classifier

Naïve Bayes is a probabilistic grouping set of rules. Each expression is considered as an unbiased word as it presently does not remember the area of a term in the sentence. Credulous Bayes is essentially founded on Bayes hypothesis to figure the opportunity of each term which compares to a label. $p(\text{label}|\text{features}) = p(\text{label}) * p(\text{features}|\text{label})/p(\text{features})$ , where  $p(\text{label})$  is the earlier likelihood of the mark in the dataset,  $p(\text{feature}|\text{label})$  is the earlier likelihood of a component identified with a name, and  $p(\text{feature})$  is the earlier likelihood of an element that has happened. The accuracy achieved through this model was not good because the Naive Bayes approach implicitly assumes that all attributes are mutually independent. It was not producing good results because the data became imbalanced due to categorical variable not observed in training dataset.

### 5.1.2 Support Vector Machines

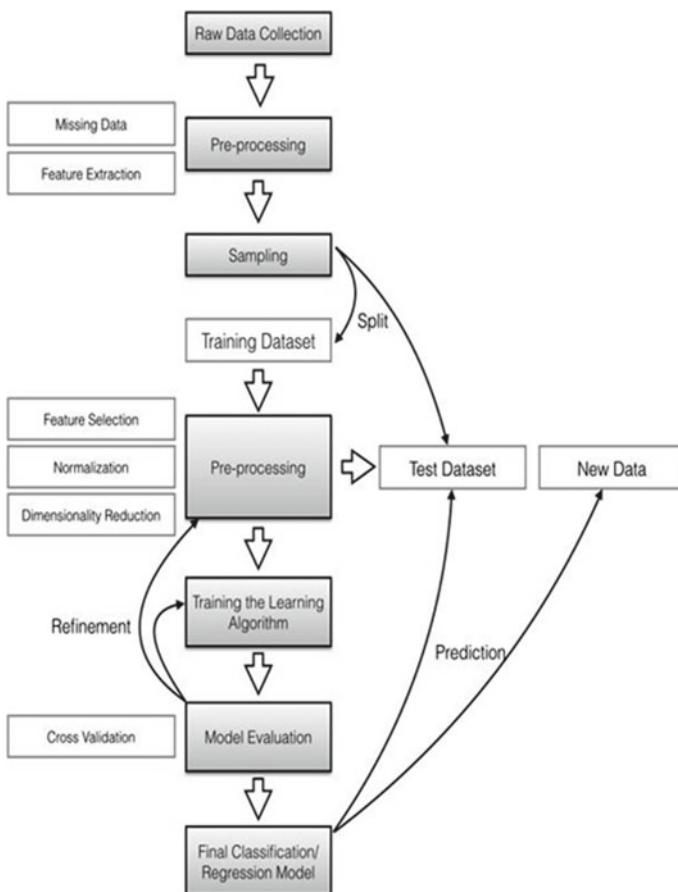
Backing vector machine introduced the essential chance to cure the issues of parallel classification. To portray the determination limits, the greater part of the information factors which are from uncommon preparing SVM has some expertise in deciding the fine hyperplanes that go about as a separator. To keep the greatest distance among control vectors of different preparing a hyperplane should be settled on. The support vector machine has the ability to control the straight and non-direct characterization errands. In SVM, chi-rectangular can be utilized as a component decision that is utilized for dimensionality decrease and commotion disposal. Since the highlights are many, I utilized SVM. It is not influenced by any exceptions when recognized. This is on the grounds that the calculation takes a shot at hyperplanes idea that are influenced by help vectors. The other motivation behind why I utilized this calculation was it functions admirably with information in content records and also scales up well when we have enormous measure of information. The motivation behind why SVM did not perform all around was a result of hyperplane choice which decreases effectiveness and exactness of the model.

### 5.1.3 Logistic Regression

Calculated backslide is a quantifiable model that in its fundamental structure uses a vital ability to exhibit an equal ward variable, but much more marvelous enlargements exist. In backslide examination, key backslide (or logit backslide) is surveying the limits of a determined model (a kind of twofold backslide). This calculation works dependent on the likelihood that it appoints perceptions dependent on discrete class. This calculation performed well when contrasted with others as the information was straightly divisible. This calculation overfits when we are having huge measurement information.

## 5.2 Features Used

On studying carefully about the project guidelines and description, I was able to judge that the accuracy for the model depends upon the various features that we will select to train the model. The system architecture is illustrated in Fig. 2. If we have better features, then better will be the accuracy of the model. At first, the aim was to get started with the problem. So, I started searching for more articles for false misclassification by web scraping, then combining the true and false misclassification data, and calculating the features using the most common approach which was normal feature selection, i.e., removing stop words, removing punctuation of each text, lemmatizing, and converting each word to lowercase. Using these features, I started applying different models on the same set of feature space with the statistical machine learning algorithm, i.e., Bernoulli Naive Bayes classifier, multinomial Naive



**Fig. 2** System architecture

Bayes classifier, logistic regression, and support vector machines. The accuracy of the model is as follows: Naïve Bayes 0.22  $f_1$ \_score, support vector machines 0.24  $f_1$ \_score, and logistic regression 0.28  $f_1$ \_score.

Since the accuracy was too low, I decided to change the method to calculate the features. I started searching about bag of words. This led me to search about something called Yet Another Keyword Extractor (YAKE). It is an unsupervised automatic keyword extraction method. It helps in calculating the keywords with more importance in the sentences with their importance value. The most important feature of YAKE was that it does not need data to be cleaned. It just takes the data and starts calculating features, and since it is domain and language independent, it helped in calculating way more better features. After getting the features, each text has its own sets of features and the number which tells us the importance of those words in the sentence. I applied the same three models on these set of features, and the results were as follows: Naïve Bayes 0.42  $f_1$ \_score, support vector machines 0.43  $f_1$ \_score, and logistic regression 0.44  $f_1$ \_score.

To improve the model accuracy, I was encouraged to search for new methods that will improve my features. Then I got to know about TF-IDF vectorizer. It is a model used for learning vector representations of words called “word embeddings.” Through this, I got word vectors which are fed into the model to do analysis. I performed all the above models as discussed. Then to improve the results, again I performed same models but after performing hyperparameter tuning of the models on the development data and then training it on train data. The accuracy achieved was as follows: Naïve Bayes 0.49  $f_1$ \_score, support vector machines 0.59  $f_1$ \_score, and logistic regression 0.65  $f_1$ \_score.

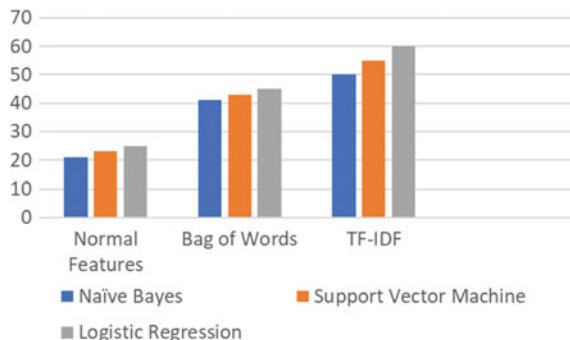
## 6 Results and Discussion

The accuracy of the model depends on how well it performs on training data with the features generated. The evaluation metric for the model was the accuracy of each model. A great machine learning model depends on how well it performs during training of the data and as well as on the test data. After the thorough hyperparameter tuning on the best performing model on test data, we know that model performance depends upon overfitting and underfitting. Underfitting as well as overfitting can lead to poor model performances.

Overfitting happens when the evaluation of train data and test data is different. To reduce overfitting of the model, either we can use a resampling technique for estimating the accuracy of the model or we can hold a validation dataset.

The most common sampling method is  $K$ -fold cross-validation. It helps us in training and testing the model  $k$  times. This is achieved by dividing the data in  $k$  different parts and taking  $k - 1$  for the test parts and rest for the training set.

**Fig. 3** Accuracy with different models for various features



## 7 Conclusion

I was able to arrive at the conclusion that logistic regression with TF-IDF features performed better. The reason behind this is given in error analysis. I had the option to come to the end result that logistic regression with TF-IDF highlights performed better. The outcome I accomplished before was 0.65 utilizing the TF-IDF vectorizer on hyperparameter tuned boundaries; however, after the last assessment, the outcome got changed and my last *f1\_score* diminished by 0.2; this can be because of model overfitting on the information. Furthermore, other explanation can be that we needed to extricate information from the web so the *f1\_score* would have diminished. The accuracy is given in Fig. 3.

It is moreover important that the inoculations in this examination did not indicate the specific trickiness that was presented after the immunization, yet rather advised about lie from a broader viewpoint by explaining the general technique being used to make question about an issue in the public's mind. The explanation behind such an intervention is to strengthen essential altogether thinking about the explanation of argumentative strategies, along these lines promising people to move past shallow heuristic-driven dealing with and participate in more significant, more key exploring of the presented information. An aftereffect of this approach is that generally plot vaccinations may execute different misdirecting disputes that use a comparable system or bogus thought.

## 8 Future Enhancement

I should endeavor LSTM count for future explanation behind future explanation. According to me, this count would have performed better and would have given better results on such data as “LSTM networks suit well for requesting, predicting, and taking care of the data reliant on time game plan”. In light of stops while playing out this on critical events between time plans for dark range, LSTM will be made to oversee vanishing point that is experienced while planning with regular RNNs.

Relative brutality toward opening length is a touch of elbowroom of LSTM over RNNs.

Social qualities, lone perception, social models, making headway, and a changing media scene all add to the multifaceted issue of distortion. Countering distortion requires a multidisciplinary procedure, including the mix of the divulgences of social, political, data, PC, and mental science in intertwined, exhaustive blueprints.

## References

1. J. Farrell, The growth of climate change misinformation in US philanthropy: evidence from natural language processing. *Environ. Res. Lett.* **14** (2019)
2. K.M.I. Treen, H.T.P. Williams, S.J. O'Neill, Online misinformation about climate change, in *WIREs Climate Change* (Wiley, 2020)
3. J. Cook, Understanding and countering misinformation about climate change, in *Handbook of Research on Deception, Fake News, and Misinformation Online* (2019)
4. M. Balaanand, N. Karthikeyan, S. Karthik, R. Varatharajan, G. Manogaran, C.B. Sivaparthipan, An enhanced graph-based semi-supervised learning algorithm to detect fake users on Twitter. *J. Supercomput.* **75**(9), 6085–6105 (2019). <https://doi.org/10.1007/s11227-019-02948-w>
5. N.T. Le, J.-W. Wang, C.-C. Wang, T.N. Nguyen, Novel framework based on HOSVD for Ski goggles defect detection and classification. *Sensors* **19**, 5538 (2019). <https://doi.org/10.3390/s19245538>
6. E.K. Lawrence, S. Estow, Responding to misinformation about climate change. *Appl. Environ. Educ. Commun.* (2017)
7. J. Cook, S. Lewandowsky, U.K.H. Ecker, Neutralizing misinformation through inoculation: exposing misleading argumentation techniques reduces their influence. *PLOSone* (2017)
8. R.E. Dunlap, P.J. Jacques, Climate change denial books and conservative think tanks: Exploring the connection (2013)

# Phishing Website Detection and Classification



D. Viji, Vaibhav Dixit, and Vishal Jha

**Abstract** Websites are used nowadays for a lot of online transactions like E-commerce which involve making online payments. These websites may or may not ask user sensitive data such as username, password, credit/debit card, and details. Websites which ask users for these sensitive data for malicious or illegal activities are termed as phishing websites. Detection of phishing website is a crucial task today to prevent phishing attacks. For detection and prediction of phishing/fraudulent websites, we propose a system that works on classification techniques and algorithm and classifies the datasets as phishing/legitimate. It is detected on various characteristics like uniform resource locator (URL), domain name, domain entity, etc. When the user makes the online transaction after the payment is done through the gateways of the website, data mining algorithm will be used to determine it as phishing/non-phishing. This application will find any real-life applications primarily in e-commerce, so that the whole transaction process is secure. The performance of data mining algorithm is better than the other classification algorithms which are used traditionally. This application enables the user to perform online transactions fearlessly. Phishing/fake website URL can be added by the admin in the system database, after which it will be scanned using the data mining algorithm, once the scan is completed, it will add new keywords in the database. Phishing has become one of the major methods of cyber-attack where credentials of users are obtained using illegal ways. URL of the website will used mainly as the input data, and apart from that the system will also scan for already present blacklisted keywords, link in the website and classify it as phishing/safe. In order to ensure maximum accuracy, exhaustive dataset will be used to train the model.

---

D. Viji (✉) · V. Dixit · V. Jha

Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai, India  
e-mail: [vijid@srmist.edu.in](mailto:vijid@srmist.edu.in)

V. Dixit  
e-mail: [vr1184@srmist.edu.in](mailto:vr1184@srmist.edu.in)

V. Jha  
e-mail: [vu6770@srmist.edu.in](mailto:vu6770@srmist.edu.in)

**Keywords** Phishing · Phishtank · URL · Domain entity · Fraudulent websites

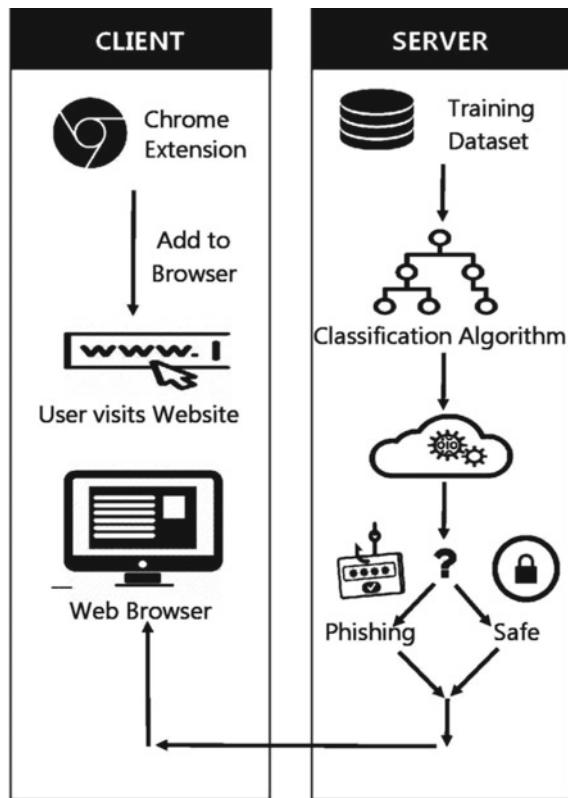
## 1 Introduction

Phishing is termed as an illegitimate/fraudulent attempt to gain access to sensitive and confidential data like password, user ids, and card details. Phishers/cyber-thieves disguise themselves and project themselves as a trustworthy organization. Phishing attacks are generally carried out by email spoofing, messaging [1]. In some cases, a replica of original website is created to scam the users to enter their personal and confidential information. These fake websites are the exact lookalikes of the original sites. Phishers also play with the psychology of the users by sending them attractive offers and discounts from trusted websites, social media platforms, banks, etc. They also pretend to be IT administrators of various organizations and ask for e-banking details of the customers via emails, phone calls, and text messages. A lot of attempts have been made by authorities to deal with phishing attempts which includes creating public awareness, adapting new and more secure technical measures, bring strict cyber-crime laws in legislatures. The word phishing is a homophone and is a more catchy/sensational spelling of “Fishing” which is also influenced by “Phreaking” [2]. Recent research on major road accidents showed that statistically most of the major road accidents are caused by heavily loaded trucks. To ensure trucks and other heavy loaded lorries do not involve in major accidents, a system was proposed called visible light communication (VLC) scheme [3]. In this, the weight of the load on the truck was connected with the traffic light system which automatically changed its lights to decelerate or slow down the heavy vehicles and other small vehicles at accident prone roads and junctions/intersections so as to ensure road safety and reduce the chances of accident to a maximum extent without applying of emergency brake by the driver in most of the cases which also leads to loss of control of the vehicle in many cases [4].

Moreover, the entire issue of road safety is not resolved by VLC communication system, with increase in the sale and purchase of vehicles traffic on roads is going to increase which can also lead to increase in road accidents if not monitored properly. Therefore, road traffic accident can be an important issue. Thus, an upgraded version of VLC termed as intelligent traffic system (ITS) was introduced which included many modern technologies like sensors, automatic control, communicators, and radar. It had many advantages over VLC such as high bandwidth, low transmission, long service life, and cost effectiveness.

Here, in our case too, many systems and techniques are in practice to ensure Web security, identify phishing, and fraudulent websites, but with ever increasing number of users on the Internet, efficiency of these algorithms is not sufficient to detect these websites, so a new more effective and time effective as well as cost-effective phishing classifier is the need of the hour in order to provide a secure and safe IT environment to the users.

**Fig. 1** Working of a phishing classifier



In the above-mentioned Fig. 1 [4], working of a phishing classifier is demonstrated, the communications and processes involved in the client and server side. When the website is visited by the client using any browser (here, it is chrome), an extension will be added to the browser which will fetch the URL and attributed of the website, page can be extracted [5]. This extracted data will serve the purpose of dataset for the classifier algorithm. This can be trained on a phishing website dataset. Then, the classifier will predict whether the URL entered is malicious or safe. It will alert the user if it is a phishing website otherwise if it is safe, it will give a green signal to user to carry on with the website [6].

## 2 Related Work

There have been different endeavors made to locate a hearty and dynamic answer for this difficult which can check if a site is phishing dependent on its present credits instead of recently characterized rules. Ankit Kumar Jain et al. introduced a far-reaching examination of all the phishing assaults known, alongside their subsequent

results. Additionally, it likewise gives a valuable knowledge over the different AI based methodologies for phishing discovery with the assistance of an imperative report. This opens up different perspectives regarding discovering more effective arrangements with assistance of AI in close future [7]. The last report consequently sets up that various phishing attacks can be set up by the detection component. In any case, there are still possibilities of receiving bogus alarms [8]. In powerfully heuristic enemy of deceitfulness framework [9], managing the presentation of phishing websites which are either old or known sees some improvement with this calculation [10]. Two calculations, namely backpropagation with SVM from Adaline Organization, were executed which are the rate of identification and utilization of the datasets of Alexa and Phishtank. Four boundaries were used in the assessment of the presentation of two calculations, namely mean square blunder, time for preparation, time for testing, etc. Accuracy provided by this algorithm was found to be around 99.14558% [11]. It utilized extremely less training time as compared to backpropagation network and SVM. A hybrid model was introduced which was capable of utilizing 30 highlights or parameters for tackling the phishing sites [12]. A solitary model cannot effectively distinguish the phishing sites, subsequently to upgrade the exactness, proficiency, and execution rate, at least two models are joined together to frame a stronger classifier. Right off the bat, the assessment of best classifier is performed with high precision and low rate of blunder is checked.

They have applied SVM metaheuristic calculation. In [13], Mohammad et al. presented system with versatile self—organizing neural. With the boosting calculation, it begins through appointing identical significance to each example of the preparation information. Next, the calculation orders for a classifier to be framed for this information, rechecking each example with respect to the classifier's yield. The significance of those accurately characterized diminishes, while the mistaken is expanded. With this, a bunch of simple cases of low significance and a bunch of "troublesome" are made [14]. Next, and in every single ensuing occasion, a classifier will be worked for the information that is rechecked (which consequently centers around arranging troublesome cases appropriately). Subsequently, the loads of the occurrences are enlarged or diminished concerning the yield the new arrangement. Subsequently, a portion of the troublesome occasions may change to simpler ones, and the other way around [15, 16]. These conceivable outcomes are conceivable. Following every occasion, loads can reflect precisely how frequently occurrences have been determined wrongly by classifiers created up till now. Keeping up some trouble, or "hardness" alongside, the occurrences can produce arrangement of correlative specialists that are viable with each other [17–19].

In Table 1, results of the study of combination of classifiers and their accuracy are shown, and their advantages and drawbacks are discussed.

In this paper, freely accessible phishing sites informational collections from the UCI AI repository† is utilized to survey the model exhibitions. We have utilized open source WEKA‡ AI instrument to assess the characterization correctness, zone under the ROC bend, and F-proportion of every method. Diverse AI strategies are utilized for arrangement utilizing a ten-overlap cross approval. Artificial intelligence methods yielded better accuracy and were instrumental in the identification

**Table 1** Study of combination of classifiers

Paper ID	Classifier combination used	Advantages and drawbacks	Accuracy (%)
[20]	KNN + random forest	Considered as best in terms of self-learning but has comparatively less accuracy	91.65
[21]	Linear regression + SVM	Very high in terms of detection speed	99.36
[22]	Decision tree with datasets stored in stacks	Showed impressive results in terms of speed but comparatively less accurate	97.25
[23]	SVM and KNN from Phishtank dataset	Best in terms of speed as well as accuracy	99.25
[14]	Random forest with backpropagation algorithm	Showed impressive results with text detection but was slow in detection graphics and multimedia objects	98.36

and detection of phishing websites and accomplished preferred execution over other classifiers. The outcomes generated from them are solid for all the other AI procedures, and it also established an outcome that the results yielded by RF model was more dependable and had very high accuracy [20].

ROC curves of other important and widely used algorithms were also implemented and analyzed such as K-nearest neighbor (KNN), support vector machine (SVM), and random forest. For base classification, decision tree algorithm was utilized [24]. A few AI approaches utilized different component extraction strategies. They were utilized to prepare phishing page classifier. AI techniques also found application in training the classifiers and datasets because it was not possible for a model of solitary nature to distinguish the phishing sites effectively to subsequently upgrade the exactness, proficiency, and execution rate, at least two models are joined together to frame a stronger and more effective classifier [23].

This proposal of joining of two or more different kinds of classifiers yielded some mind blowing and unbelievable outcomes which were never anticipated [14]. Earlier, the main issue while training most of the algorithms was the number and nature of classifiers on which they were supposed to work. Increasing the number of classifiers lead to reduction in the time used by these algorithms, whereas reduction in these parameters increased the accuracy but only for a brief period of time, It was not a viable solution in the longer run as most of the websites nearly 90 of them are following dynamic pattern of Web development, so it is very easy to make updates and changes to the content of the websites in order to overcome the detection algorithm as they do not remain static anymore. So with less number of classification parameters, it will no longer serve the purpose of the classification [24].

Hence, a conclusion was drawn that a single classification algorithm is not appropriate for this task as we needed more number of classification parameters for the detection and parameters which was adversely affecting the accuracy and speed of

the algorithm, So combination of two classification models was done on an experimental basis to check whether all the criteria for the detection and classification along with high speed and accuracy is achieved and with different kinds of combinations it yielded different results but was found to be a better option than using a single classifier model and overloading it with parameters [25].

### 3 Types of Phishing

#### 3.1 *Spear Phishing*

Spear phishing is termed as the phishing attempts which target specific companies or individuals. It differs from bulk phishing, as in spear phishing the attackers collect personal data and info to improve their success probability. A study on social phishing, a subcategory of spear phishing, revealed that its success rate was around 70%. Spear phishing targets employees and executives particularly in the financial departments to gain to financial data [3]. A threat group named Fancy bear targeted accounts linked to the campaign of Hillary Clinton in 2016 Presidential Elections using the tactics of spear phishing. More than, 1800 google accounts were targeted, and Google.com domain was used to threaten users.

#### 3.2 *Whaling*

Whaling is a term that refers to techniques that aim to target executives at senior levels and other employees with high authoritarian access. In whaling, they will craft a content specifically targeting senior-level managers [9]. The content can be a customer's complaint, an administrative or executive issue.

#### 3.3 *Catphishing*

In this, the attackers involve a deception method using which they personally get to know about the user with the intention of getting access to sensitive information or resources. They also usually take control over the target. Catfishing(f) is similar to Catphishing(ph), but is a slightly distinct concept, here it involves a person creating a presence on social network in order to involve the victim into a sentiments and involve with them romantically [26]. They start it online and make false promises of progressing into real-life romance. However, the main objective here is to gain access to victim's financial resources like bank accounts, personal property, etc.

### ***3.4 Voice Phishing***

It is not necessary to involve the use of a website in order to conduct a phishing attack. In voice phishing, numbers are sent to the victims in the form of messages claiming to be from bank authorities. They claim to be numbers of customer case and ask the users to dial those numbers to solve those problems. Once the user dials those numbers, it prompts the user to enter their account details like account number, PIN, etc. Voice phishing is also called vishing. They also use fake caller ids so as to look genuine and trustworthy [27].

### ***3.5 SMS Phishing***

SMS phishing is also called as smishing. It uses text messaging services to deliver messages to involve people to disclose their confidential information. These attacks usually prompt the user to click a link, dial a phone number, or an email address sent via SMS. The recipient is then asked for their private data such as credentials for other websites. Also, in many android browsers, full URL is not displayed, which makes it more difficult for the user to identify a fake page, as modern smartphones have very fast internet connectivity, and their UI (user interface) is designed such that it directly prompts to the page associated with the link [8].

### ***3.6 Clone Phishing***

Clone phishing is another phishing type in which contents of a previously delivered mail is copied and links in the previous mail is replaced with new malicious links to trap the user. In clone phishing, the email may claim to be a resend of the previous mail and ask for sensitive and confidential data from the recipient.

## **4 Phishing Techniques**

### ***4.1 Link Manipulation***

Major phishing types use technical deception in one form or other which is designed to make a link in an email. This link leads to the spoofed website or the malicious URL, or sometimes use common tricks like misspelt URLs, common subdomains, etc. [28].

## 4.2 *Filter Evasion*

Phishers sometimes use image in place of text to make it difficult for anti-phishing filters to detect the malicious content which is commonly used in phishing. To detect these images, more sophisticated algorithms are required to be implemented in anti-phishing classifiers, and Adobe flash is a tool used by phishers which is termed as flashing to avoid these anti-phishing classifiers that scans texts in images [29]. This technique is similar to a website, but here the phishing or malicious text is hidden inside a multimedia object.

## 4.3 *Website Forgery*

Website forgery is a major and one of the most commonly implemented phishing technique. JavaScript is used by some phishers which alters the address bar of the website they lead to. This is done in a number of ways. They either place an image or some other multimedia object over the address bar to hide the actual address and open a new link with the legitimate URL. Flaws present inside a trusted website's scripts is also used against the victim. These attacks are termed as cross site scripting which are the major problem as they trap the user into entering their login credentials for bank accounts which may appear correct to the user, but in reality, it is a forged website. This kind of forgery was used against PayPal in the year 2006 [30].

## 4.4 *Covert Redirect*

Covert redirect is another method which is used to perform phishing attacks where links may appear legitimate, but in reality, it redirects victim to a malicious website. This is usually masqueraded under a popup domain. It exploits the users on various well-known parameters such as OAuth 2.0 and OpenId as well. This also makes use of XSS and open redirect vulnerabilities in the websites based on third-party applications. Malicious browser extensions are also used to redirect users to these pages covertly. Normal link manipulation techniques are easy to detect as the malicious URL will be different from the genuine one. But in covert redirect, a real website can be used by the attacker by corrupting the original website with a malicious popup or login box or any other dialog box [19]. This differentiates covert redirect method from other phishing techniques. The whole process of covert redirect can be understood with the following case. If a user opens Facebook, a malicious popup window implanted by the phisher will show up and ask the user for authorization from the app, and it can expose sensitive information of the victim. The attacker can also gain access to more sensitive and confidential information like mail, chats, friends list,

etc. Even if the user denies authorization on the malicious popup, the presence of user on that page can be a major cause of threat.

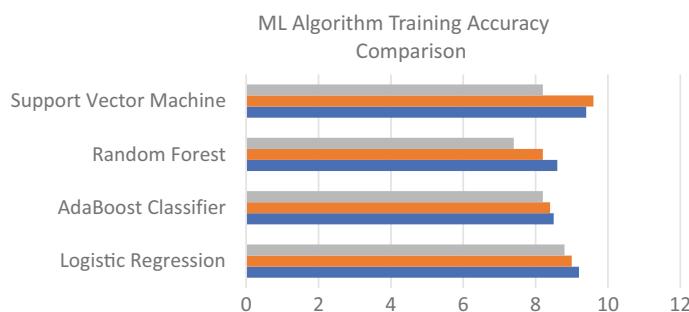
## 5 Classifier Comparison

Table 2 comprises of the study of various classifiers and the accuracy levels of different types of algorithms which followed different methods of detecting plagiarism from various research papers.

Figure 2 is a graphical representation of the average accuracy levels of most commonly used algorithm used as a classifier in the detection of phishing websites.

**Table 2** Accuracy level of different algorithms

Paper ID	Method	Classification accuracy (%)
[20]	System blacklisting	95.25
[21]	Hybrid detection	94.05
[22]	Web identity using web image	96.25
[31]	Website logo	92.49
[23]	Identity keywords extraction	96.14
[32]	Random forest classifier	97.36
[17]	CNN	96.35
[18]	Jail-Phish	98.61
[19]	Favicon leveraging	96.45



**Fig. 2** Accuracy level of most commonly used ML algorithms for phishing classification

## 6 Conclusion

Our aim with this proposed system is to implement the classification and detection of phishing websites with the help of data mining algorithms. Feature extraction of URL is done by when it is visited by the user. The extracted features will serve as the test data for the model, and this proposed model can be implemented by the random forest algorithm. The primary task and function of this system is the detection of phishing website and alert the user to prevent the leak of their sensitive and personal information. If a user wants to access the website even after the warning given by the system, the user can proceed at their own risk. Feature selection is proposed in the future models that will reduce the need for dependency on page content in the present model. Furthermore, more detailed study is needed for detection of phishing websites via mobile devices. Smartphones are a very popular and widely used offspring at present, and they also act as a common merger point for all the phishing attacks. The users of cell phones preferably read/check their mails on their phones immediately. Thus, finding a possible, potent, and efficient machine learning algorithm to detect phishing attacks on mobile phones is the need of the hour. A lot of study and research is needed to be done in the field of prevention and detection of cyberattacks and phishing on mobile phones.

## References

1. N. Abdelhamid, A. Ayesh, F. Thabtah, Phishing detection based associative classification data mining. *Expert Syst. Appl.* **41**(13), 5948–5959 (2014)
2. S. Abu-Nimeh, D. Nappa, X. Wang, S. Nair, A comparison of machine learning techniques for phishing detection, in *Anti-Phishing Working Groups Ecrime Researchers Summit* (2007), pp. 60–69
3. L.I. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, Classification and regression trees (cart). *Biometrics* **40**(3), 358 (1984)
4. APWG, Statistical highlights for 4th quarter 2016 (2016). [http://docs.apwg.org/reports/apwg\\_trends\\_report\\_q4\\_2016.pdf](http://docs.apwg.org/reports/apwg_trends_report_q4_2016.pdf)
5. V. Cherkassky, The nature of statistical learning theory. *IEEE Trans. Neural Networks* **8**(6), 1564 (1997)
6. G. Bottazzi, E. Casalicchio, D. Cingolani, F. Marturana, M. Piu, MP-shield: a framework for phishing detection in mobile devices, in *Proceedings—15th IEEE International Conference on Computer and Information Technology, CIT 2015, 14th IEEE International Conference on Ubiquitous Computing and Communications, IUCC 2015, 13th IEEE International Conference on Dependable, Autonomic and Se, 1977–1983* (2015)
7. I. Fette, N. Sadeh, A. Tomasic, Learning to detect phishing emails, in *Proceedings of the International World Wide Web Conference (WWW)* (2007)
8. iTrustPage (2013). <http://www.cs.toronto.edu/ronda/itrustpage/>
9. T.-C. Chen, S. Dick, J. Miller, Detecting visually similar web pages: application to phishing detection. *ACM Trans. Internet Technol.* **10**(2), 1–38 (2010)
10. L.-h. Lee, K.-c. Lee, Y.-c. Juan, H.-h. Chen, Y.-h. Tseng, Users' behavioral prediction for phishing detection, in *Proceedings of the 23rd International Conference on World Wide Web* (2014), No. 1, pp. 337–338

11. E. Medvet, E. Kirda, C. Kruegel, Visual-similarity-based phishing detection, in: *Proceedings of SecureComm 2008* (ACM, 2008)
12. Y. Pan, X. Ding, Anomaly based web phishing page detection, in: *Computer Security Applications Conference, ACSAC '06* (2006), pp. 381–392
13. N. Chou, R. Ledesma, Y. Teraguchi, D. Boneh, J.C. Mitchell, Client-side defense against web-based identity theft, in *Proceedings of the 11th Annual Network and Distributed System Security Symposium (NDSS)* (2004)
14. F. Feng, Q. Zhou, Z. Shen, X. Yang, L. Han, J. Wang, The application of a novel neural network in the detection of phishing websites. *J. Ambient Intell. Humaniz. Comput.* 1–15 (2018)
15. S. Smadi, N. Aslam, L. Zhang, Detection of online phishing email using dynamic evolving neural network based on reinforcement learning. *Decis. Support Syst.* **107**, 88–102 (2018)
16. T. Peng, I. Harris, Y. Sawa, Detecting phishing attacks using natural language processing and machine learning, in *presented at the 2018 IEEE 12th International Conference on Semantic Computing (ICSC)* (2018), pp. 300–301
17. S. Zhu, V. Saravanan, B. Muthu, Achieving data security and privacy across healthcare applications using cyber security mechanisms. *Electron. Libr.* **38**(5/6), 979–995 (2020). <https://doi.org/10.1108/el-07-2020-0219>
18. P.N. Hiremath, J. Armentrout, S. Vu, T.N. Nguyen, Q.T. Minh, P.H. Phung, MyWebGuard: toward a user-oriented tool for security and privacy protection on the web, in *Future Data and Security Engineering. FDSE 2019. Lecture Notes in Computer Science*, vol 11814, ed. by T. Dang, J. Küng, M. Takizawa, S. Bui (Springer, Cham, 2019). [https://doi.org/10.1007/978-3-030-35653-8\\_33](https://doi.org/10.1007/978-3-030-35653-8_33)
19. M. Babagoli, M.P. Aghababa, V. Solouk, Heuristic nonlinear regression strategy for detecting phishing websites. *Soft Comput.* **23**(12), 4315–4327 (2019)
20. B. Wardman, T. Stallings, G. Warner, A. Skjellum, High-performance content-based phishing attack detection, in *eCrime Researchers Summit* (IEEE, San Diego, CA, 2011), pp. 1–9
21. Wikipedia, Decision tree learning (2017). [https://en.wikipedia.org/wiki/Decision\\_tree\\_learning/](https://en.wikipedia.org/wiki/Decision_tree_learning/), [Online]
22. Wikipedia, Support vector machine (2017). [https://en.wikipedia.org/wiki/Support\\_vector\\_machine/](https://en.wikipedia.org/wiki/Support_vector_machine/), [Online]
23. W. Zhang, H. Lu, B. Xu, H. Yang, Web phishing detection based on page spatial layout similarity. *Informatica* **37**(3), 231–244 (2013)
24. P. Likarish, E. Jung, D. Dunbar, T.E. Hansen, J.P. Hourcade, B-apt: Bayesian anti-phishing toolbar, in *Proceedings of IEEE International Conference on Communications, ICC'08* (IEEE Press, 2008)
25. T. Ronda, S. Saroui, A. Wolman, itrustpage: a user-assisted anti-phishing tool. In: *Proceedings of Eurosys '08* (ACM, 2008)
26. C.Inc., Couldmark toolbar (2015). <http://www.cloudmark.com/desktop/ie-toolbar>
27. M. Dunlop, S. Groat, D. Shelly, Goldphish: using images for content-based phishing analysis, in *5th International Conference on Internet Monitoring and Protection (ICIMP)* (IEEE, Barcelona, 2010), pp. 123–128
28. J. Mao, W. Tian, P. Li, T. Wei, Z. Liang, Phishing website detection based on effective CSS features of web pages, in *The 12th International Conference on Wireless Algorithms, Systems, and Applications* (2017), pp. 804–815
29. M. Moghim, A.Y. Varjani, New rule-based phishing detection method. *Expert Syst. Appl.* **53**, 231–242 (2016)
30. A. Nourian, S. Ishtiaq, M. Maheswaran, Castle: a scocial framework for collaborative anti-phishing databases. *ACM Trans. Internet Technol.* (2009)
31. G. Xiang, J. Hong, C.P. Rose, L. Cranor, CANTINA+: a feature-rich machine learning framework for detecting phishing web sites. *ACM Trans. Inf. Syst. Secur. (TISSEC)* **14**(2), 21 (2011). <http://dl.acm.org/citation.cfm?doid=2019599.2019606>
32. Y. Zhang, J. Hong, L. Cranor, Cantina: a content-based approach to detecting phishing web sites, in *Proceedings of the International World Wide Web Conference (WWW)* (2007)

# Generations of Wireless Mobile Networks: An Overview



Burla Sai Teja and Vivia Mary John

**Abstract** Telecommunication and networking have been playing a crucial role when it comes to the evolution of mankind and technology itself. Unless it was not for such data transmission networks, we would still be in an age where technology is not as advanced as it is today. Cellular networking has developed phenomenally in multiple aspects. The journey of the same began with the then revolutionary 1G all the way to today's 4G and also the 5G and the 6G which is currently under research for the future generations. This paper gives an outline of mobile generation's development by the comparison of the features of each generation and an overview of changes made from the past to the current generation. This paper elucidates the 5G and 6G technology that has come to light recently.

**Keywords** Cellular network · Generations · Beamforming · Network slicing · Edge computing · Quantum computing · Integrated networks

## 1 Introduction

Mobile technology has progressed a lot since the first commercial mobile phone. Wireless technology growth increased people's ability to interact and work in both corporate and social areas. Every wireless technology has certain standards, new techniques, different capabilities, and new features compared with previous generations. Be it technology, services offered, technology, or speed, the changes have been recorded as generations. The first generation (1G) was used analogically only for calls involving voice. 2G promoted text messaging moving to a digital side. Mobile technology of 3G offers improved bandwidth, thereby giving a higher rate of data transfer and multimedia support. 4G integrates 3G using fixed Internet in assisting mobile Internet with improved data speed, thereby overcoming 3G's limitations. The

---

B. S. Teja (✉) · V. M. John  
CMR Institute of Technology, Bengaluru, Karnataka, India

bandwidth is also increased which leads to a reduction in costs. New wireless technologies are being introduced for the future in order to solve the limitations of 4G such as 5G and 6G.

## 2 Generations of Cellular Network

A cellular network extends via land areas known as “cells,” where each is served by a transceiver having at least a single fixed location but three cells or base stations. Cellular networks deliver various features such as more power than a single large transmitter, and portable devices utilize lower power compared to a single satellite as the towers are close and bigger area of coverage compared to a single transmitter.

New technology was developed in a sequence of generations. The word “generation” was not used until 3G was introduced. The further chapters look at the meaning and characteristics of all the generations up to now and the future generations and what their characteristics, advantages, and contribution might be to humanity.

### 2.1 Zero Generation (0G)

Wireless telephone systems preceded the current cellular technology. They were the ancestors of the first generation of mobile phones and are frequently mentioned as pre-cellular systems. Pre-cellular networks used technologies involving PTT or manual, MTS, and IMTS. Only one person can speak at a time using the PTT function. A temporary button was used in switching to transmission mode from voice reception mode [1].

### 2.2 First Generation (1G)

In 1979, Nippon Telephone and Telegraph (NTT) DoCoMo introduced 1G commercially in Tokyo, Japan. AMPS have since begun to be utilized in Australia and North America, Total Access Communications Systems (TACS) in the United Kingdom including several others [2]. 1G’s maximum speed was 2.4 Kbps. 1G technology was not digital even though it was the first cellular telecommunication system. Data transmission was achieved in the form of analog at speeds of 150 MHz and also above radio wave level—1G technology’s greatest downside. The transmission speed of data was also low and only practical for the purpose of phone calls [3].

## 2.3 Second Generation (2G)

2G or the second generation was first commercially introduced in Finland in 1991. This used digital voice transmission signals. This technology's fundamental emphasis was on digital signals and provided services to transmit content and provide low-speed (kbps) picture messages. This used the 30–200 kHz bandwidth. This has brought out devices smaller in size, a stable link, better quality for calling, and an elevated networking capacity. 2G has higher protection for both sender and recipient. All text messages are digitally encrypted, allowing data to be transmitted in such a way that only intended recipient can receive and read them. 2G network uses the TDMA and CDMA mobile communication technology [4, 5].

**2.5G or GPRS.** GPRS is a packet-oriented mobile communication system in GSM as 2G network. It is the best-effort technology, which means variable throughput and latency depending on the number of other users concurrently accessing the link. Whereas in circuit switching, a particular QoS is undertaken during communication. GPRS offered a data speed in the range of 56–114 kbps. This enables transfer of data at moderate speed, using unused TDMA channels in, for example, GSM.

**2.75G or EDGE.** EDGE is also known as enhanced data rates for global evolution, IMT-SC or EGPRS. It belongs to a cell phone technology that enables enhanced speeds of data transmission as an extension of GSM that is backward compatible. It is used for packet switching applications like the Internet connectivity. For 2.5G GSM/GPRS networks, EDGE is introduced as a bolt-on addition, making it easier for current GSM carriers to migrate to it. It will operate on any network that has GPRS installed on it, provided the carrier implements the required upgrade.

## 2.4 Third Generation (3G)

The next generation in modern mobile telecommunication technologies is 3G. It was an upgrade from the 2G and the 2.5G for quicker transmission of data. This was established on the guidelines set by International Telecommunications Union for cellular devices and mobile communications that use systems and networks adhering to the International Mobile Telecommunications-2000 (IMT-2000) policies. 3G includes cellular mobile Internet access, voice telephony, video calls, wireless Internet, and mobile television. 3G telecommunications networks support services that provide at least 144 kbps of information transfer rate [6].

**3.75G or HSPA+.** HSPA+ is the second step of HSPA which was implemented in version 7 of 3GPP and was further enhanced in later releases of 3GPP. HSPA+ is efficient enough to achieve data rates up to 42.2 Mbps. It introduces antenna array technologies such as MIMO communications and beam forming. Beam forming focuses an antenna's transmitted power in a beam, toward the direction of the recipient.

## 2.5 Fourth Generation (4G)

4G, the fourth generation in cellular network technology succeeded 3G. A 4G system should provide abilities defined by ITU. The potential or current applications involve modified mobile internet access, video conferencing, gaming services, IP telephony, mobile high-definition television, and 3D TV. WiMAX and LTE are the two main standards highlighted for 4G [7].

LTE, established on UMTS/HSPA and GSM/EDGE technologies, is a standard for wireless communication in cellular devices and data terminals. It increases capacity and speed along with improvements in core network using different radio interfaces. LTE is an improved path for carriers having both CDMA2000 and GSM/UMTS networks.

## 2.6 Fifth Generation (5G)

5G networks are those cellular networks which divide the serviceable area into small geographic units known as cells. All of the 5G wireless devices pertaining to a cell are interconnected using radio waves via the local antenna within a cell to the Internet and the telephone network. The biggest benefit of new networks is that these have larger bandwidth, offering bigger transmission rates of 10 Gbps eventually. The 3GPP is the standard setting for industry consortium for 5G. It describes any device using 5G NR software as 5G [8–10].

### 5G NR

*Waveform.* For any wireless access technology, the option of radio waveform is the central physical layer decision. 3GPP decided to implement OFDM with a CP for both downlink and uplink transmissions. For large bandwidth operations and MIMO technologies, CP-OFDM may allow for low implementation complexity and low cost [11].

*Frequency.* NR encourages spectrum activity from sub-1 GHz to millimeter-wave bands. In Rel-15, two FR are set.

- FR1—Less than 7 GHz frequently referred to as sub-6 GHz
- FR2—Greater than 24 GHz frequently referred to as millimeter-wave.

**Edge Computing.** Edge computing is used to enhance the response times and preserve bandwidth, which is distributed computing model bringing data storage and computing closer to the area where it is necessary. Edge computing seeks to remove computing away from the data centers to edge of network, to take advantage of smart items, cell phones, or network gateways to do tasks, and to facilitate cloud-based services. By moving facilities to the edge, service delivery, content caching, IoT management, and storage can be delivered that leads to faster response times and transfer speeds [12].

## 2.7 Sixth Generation (6G)

6G networks use higher frequencies than 5G networks, and this would allow for higher data speeds to be reached and a much higher overall capacity for the 6G network.

A much lower degree of latency is almost definitely going to be a must. 6G drivers aim to be a convergence of past developments (e.g., faster speeds, large antennas, and densification) with current trends involving new networks and the latest revolution in wireless devices—AI and computing [13–16].

**Artificial Intelligence.** It is possible to perform time-consuming tasks, such as handover and network selection, by using AI promptly. AI will also allow 6G to provide its users with MPS automatically and to submit and generate 3D radio environment maps.

**Quantum Computing and Communications.** Quantum communications can provide good protection by applying a quantum key dependent on uncertainty principle and the no-cloning theorem. It can remarkably enhance and accelerate AI algorithms which need massive training and large data.

**Integrated Networks.** Drones can be supported by providing access to hotspots beyond their position as users of 6G and to areas with scarce infrastructure to supplement terrestrial networks. Meanwhile, low orbit satellite and CubeSats satellite connectivity may be needed by both drones and terrestrial base stations for providing backhaul support with extra wide area coverage.

**6G with Satellite Network.** The coordinated 5G wireless mobile infrastructure in addition to the satellite network would be the 6G mobile infrastructure for global coverage. 6G's main goal is to provide high-data speed cell phone users with a range of networks as an identifier in a many location, multimedia related applications, and Internet associativity for mobile users without interrupting the network [11]. With excellent data speeds of up to 10–11 Gbps, it will have incredibly fast access to Internet services.

## 3 Conclusion

The mobile wireless networking world is changing rapidly. The wireless industry has seen incredible growth in the last few years. In this article, we saw an overview of all the different cellular networks till date from 0 to 4G. The technology of 5G and 6G will be a new revolution in the mobile industry. In the latest 5G and 6G wireless networks, there are several new techniques and innovations which could be used. An attempt was made to review few such technologies that might be implemented in these upcoming cellular networks. The world is trying to become fully wireless, with higher quality, high performance, improved bandwidth, and cost savings requiring unrestricted access to information at anytime and anywhere.

## References

1. A.S. Meril, M. Basthikodi, A.R. Faizabadi, Review: comprehensive study of 5G and 6G communication network. *J. Emerg. Technol. Innovative Res.* **6**(5), 715–719 (2019)
2. X. Lin et al., 5G new radio: unveiling the essentials of the next generation wireless access technology. *IEEE Comm. Stand. Mag.* **3**(3), 30–37 (2019)
3. Y. Wu et al., A survey of physical layer security techniques for 5G wireless networks and challenges ahead. *IEEE J. Select. Areas Commun.* **36**(4), 679–695 (2018)
4. V. Bioglio, C. Condo, I. Land, Design of polar codes in 5G new radio. *IEEE Commun. Surv. Tutorials* (2020)
5. I. Ahmed et al., A survey on hybrid beamforming techniques in 5G: architecture and system model perspectives. *IEEE Commun. Surv. Tutorials* **20**(4), 3060–3097 (2018)
6. A.A. Barakabitze et al., 5G network slicing using SDN and NFV: a survey of taxonomy, architectures and future challenges. *Comput. Networks* **167**, 106984 (2020)
7. H. Hui et al., 5G network-based Internet of Things for demand response in smart grid: a survey on application potential. *Appl. Energy* **257**, 113972 (2020)
8. M. Balaanand, N. Karthikeyan, S. Karthik, Envisioning social media information for big data using big vision schemes in wireless environment. *Wireless Pers. Commun.* **109**(2), 777–796 (2019). <https://doi.org/10.1007/s11277-019-06590-w>
9. B.-H. Liu, N.-T. Nguyen, V.-T. Pham, Y.-X. Lin, Novel methods for energy charging and data collection in wireless rechargeable sensor networks. *Int. J. Commun. Syst.* **30**, e3050 (2017). <https://doi.org/10.1002/dac.3050>
10. T.S. Rappaport et al., Wireless communications and applications above 100 GHz: opportunities and challenges for 6G and beyond. *IEEE Access* **7**, 8729–78757 (2019)
11. W. Saad, M. Bennis, M. Chen, A vision of 6G wireless systems: applications, trends, technologies, and open research problems. *IEEE Network* **34**(3), 134–142 (2019)
12. Z. Zhang et al., 6G wireless networks: vision, requirements, architecture, and key technologies. *IEEE Veh. Technol. Mag.* **14**(3), 28–41 (2019)
13. M.Z. Chowdhury et al., 6G wireless communication systems: applications, requirements, technologies, challenges, and research directions (2019). [arXiv:1909.11315](https://arxiv.org/abs/1909.11315)
14. F. Nawaz, J. Ibrahim, M. Awais, M. Junaid, S. Kousar, T. Parveen, A review of vision and challenges of 6G technology. *Int. J. Adv. Comput. Sci. Appl.* **11**(2) (2020)
15. S. Zhang, D. Zhu, Towards artificial intelligence enabled 6G: state of the art, challenges, and opportunities. *Comput. Networks* **107556** (2020)
16. G.K. Audhya, K. Sinha, P. Majumder, S.R. Das, B.P. Sinha, Placement of access points in an ultra-dense 5G network with optimum power and bandwidth, in *IEEE Wireless Communications and Networking Conference (WCNC)* (IEEE, 2018), pp. 1–6

# Quantum Networking—Design Challenges



S. Mohammed Rifas and Vivia Mary John

**Abstract** Quantum computing has revamped and paved a new stream of research diverted away from the classical means of computation and processing. One setback in this field was the usage of existing classical Internet coupled with a machine capable of quantum computing. Hence, extensive research has been done in this domain to provide and support the incorporation of quantum Internet with quantum computing to boost the overall performance of the quantum system. This paper aims to provide possible implementations of quantum Internet which can be coupled with the quantum computer being developed by various organizations. The major advancement would be done in the link layer as well as the routing processors, as these are the layers responsible for converting signals. The routing processor should be able to cope up with the quantum qubits to re-route and perform the necessary network layer computations to direct the packages to its respective destination addresses.

**Keywords** Quantum network · Quantum Internet · No-cloning theorem · Entanglement · EPR pair

## 1 Introduction

The vision of quantum Internet is establishing a parallel link to the existing Internet technology that we are all familiar with to connect between two points around the globe. The main drawback being the physical losses during transmission of the rapid data. The key challenge is the impossible nature of quantum information to be copied and does not allow signal amplification. The need to produce quantum entanglement between the qubits is a necessity for a reliable long-distance communication without amplification. Another challenge is implementation of classical protocols of cryptography; this can be ruled out with protocols such as quantum key distribution (QKD). The quantum network, just as the classical network, consists of the layers of components such as the physical layer to the link layer.

---

S. M. Rifas · V. M. John (✉)  
CMR Institute of Technology, Bengaluru, Karnataka, India



**Fig. 1** Given two inputs  $|-\psi\rangle$  and  $|-\gamma\rangle$  as the inputs along with a unary operation U. As per no-cloning theorem, we cannot create a clone of any of the given input states; thus, such a cloning machine CANNOT exist

The optimal medium for quantum communication is believed to be photons, due to their flying nature and their compatibility with the existing architecture of communication networks. The photon loss caused by such media can be limited to an extent by using optical fibers. An alternative is to deploy relayed based on satellites for the free-space channel of communication. The accepted and widely known quantum repeater has the swapping of entanglement, the purification of entanglement and the quantum memory as the basic components.

## 2 Background

This section reviews a few of the quantum mechanics principles as well as postulates hindering the development of quantum Internet.

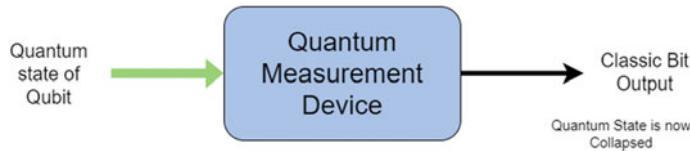
### 2.1 No-Cloning Theorem

This theorem states that an arbitrary qubit cannot be cloned in its entirety down to the last electron. This is in direct contrast with the principles and laws of quantum mechanics. Due to the no-cloning theorem, copying of information during long-distance transmission to boost the signals would not be allowed [1, 2] (Fig. 1).

### 2.2 Quantum Measurement

If a measurement can have more than one outcome, similar to the qubit possessing two states, once the measurement of the original quantum state breaks down, it can have irreversible effects on the qubit's original state. In a simpler form, the measurement corresponds to a physical system that is altered or manipulated to yield a numerical result that irreversibly alters the original state of the qubit.

The essential assumption from the above postulates is that they are applicable in closed and isolated systems only. In conclusion, we can state that systems



**Fig. 2** Quantum measurement leads to loss of quantum state after quantum measurement is applied on the qubit

that are closed and represented by Hamilton's unit time evolution can be determined by projective measurements. Systems are simply not closed and are therefore immeasurable employing projective measurements (Fig. 2).

### 2.3 *Quantum Communication*

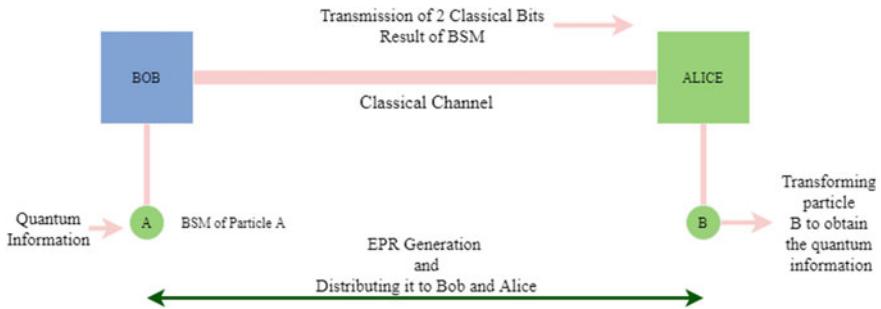
Quantum communication methodologies depend thoroughly on quantum mechanics to transmit data and exchange information. The quantum key distribution (QKD) and superdense coding are the only areas in which quantum communication has shown promising results. This, however, utilizes the quantum network to transmit classical bits, thus resulting in communication in the quantum realm, but essentially not quantum internet [3] (Table 1).

Quantum internetworking depends on the sharing of quantum information to the respective nodes. If a photon carrying such information is corrupted or altered during transmission, there exists no procedure to recover this information; hence, it would be said to be lost be destroyed. This is in direct relation with quantum mechanics principles and the no-cloning theorem. As a result, direct transmission of quantum information via photons would be unfeasible and other methodologies such as quantum teleportation could be employed (Fig. 3).

**Quantum Teleportation.** Quantum teleportation brings about an efficient stratagem for the transmission of the quantum information without the physical transmission of the photon storing the quantum information. This does so without any violation of quantum principles; it employs another operation utilizing bell state measurement (BSM) [4] and EPR Pairs which are shared between the parties involved. It is rather crucial to note that every transmission of quantum information through quantum

**Table 1** Classical versus quantum communication

	Classical communication	Quantum communication
Classical (BITS)	Traditional Internet	Superdense Coding
Quantum (QUBITS)	QKD	Quantum Internet



**Fig. 3** Quantum information from Bob is needed to be transmitted to Alice. This process would begin with Bob and Alice each receiving an EPR which is generated beforehand. The encoded quantum information in Particle A is extracted and converted to two classical bits using BSM and transmitted through a classical media. It is received by Alice and these two classical bits are then re-encoded to obtain the required quantum information.

teleportation would involve the generation of a new EPR pair for every qubit that is needed to be transmitted.

**The Quantum Networking Design.** This section would discuss the significant and key challenges to the incorporation of quantum Internet. It would discuss the various methodologies and strategies to be used to overcome various challenges in routing, protocol, interconnection, and others.

**Fidelity and Decoherence.** Fidelity [5] can be expressed as a measure of the distinctiveness of two states (quantum) between 0 and 1. In other words, if the imperfection between the measurements is high, the fidelity of that system is low. As stated before, teleportation consists of the quantum states incurring a set of operations being performed on them, any imperfections on such operations would thereby affect the fidelity of the qubit being transmitted.

The fragile nature of qubits causes it to lose information if there is any interaction between the qubit and its surrounding environment. Thus, decoherence is the loss of information from a qubit to its surrounding environment as time passes.

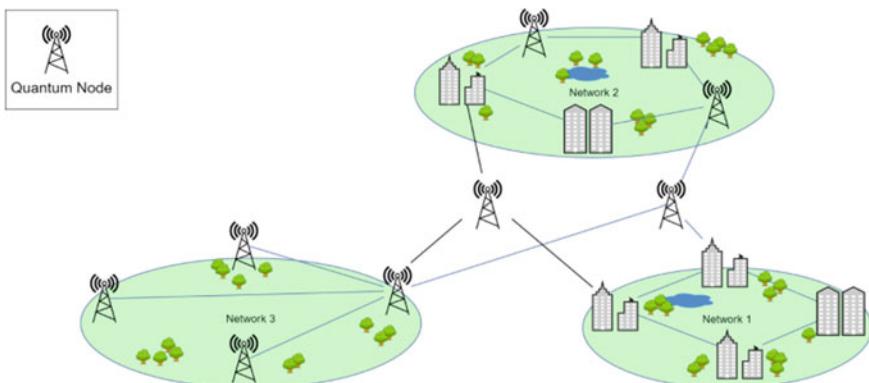
**Distribution of Entanglement.** From the above discussions, we can conclude that, when a quantum information, specifically the BSM outcome has to be transmitted, it is limited by the throughput of the classical bit. Unlike classical networks, quantum information can be transmitted only after the EPR pair has been communicated to both parties involved.

An area that has extensive research, even in the physics community, is the long-distance distribution of entanglement or simply entanglement distribution. Repeaters are used in classical networks to boost the signal, and there have been studies done to develop quantum repeaters [6].

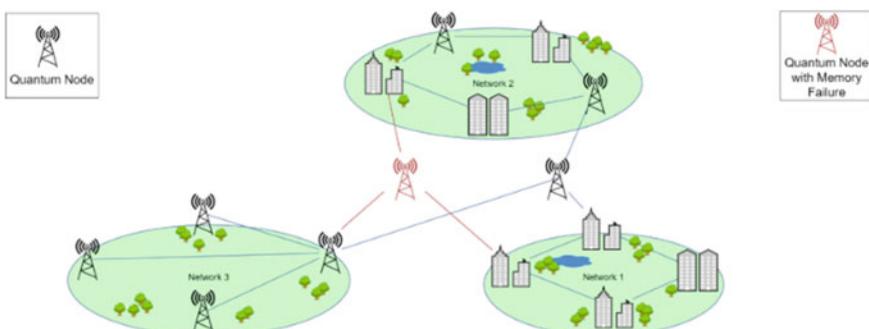
## Routing in Quantum Networks

*Adaptive Routing.* Within a quantum network, the quantum superposition states are stored in the central quantum archive consisting of its quantum routers. Any quantum impairment of memory makes this a new concern inside a quantum Internet environment to determine the fastest alternate route quickly and efficiently [7]. Backup routes exclude the points that have been impacted by quantum memory failure and provide smooth network propagation.

Throughout the solution, the fastest routes are calculated by a base network comprising all information upon this quantum network extension [8–10]. The calculation of the fastest node can be done by identifying the alternative path to the disconnected network subsystem by using various shortest path algorithms and also a set of mapped paths of the quantum network to facilitate the re-route as shown in Figs. 4 and 5.



**Fig. 4** Proposed network—system of quantum nodes—status and path to every node mapped



**Fig. 5** Memory failure at a single quantum interconnecting node causes the path to those nodes to be dropped, and an alternate path is mapped until the old path has been rectified in the master system

### 3 Conclusion

Few challenges faced during deployment as in this section. Quantum computers would be accessible in very few highly specific server farms capable of delivering the most demanding equipment required for quantum computers (ultra-high vacuum or ultra-low temperature). Industries and customers should be able to access quantum computing resources as a cloud service [11–13]. Configurations would have to take into consideration the rising importance of data packets (financially and in terms of quantum reliability) restricting both the volume of clusters and the use of networks for computing. Architectures in a hybrid nature are likely to be used for connectivity requiring either the use of such cryo fibers [14]. A substantial portion of computational research will be required to establish both innovative network architectures for quantum and classical protocols. Classical communication services are likely to be established by the convergence of conventional networks such as the global Internet with quantum Internet [15–17].

## References

1. N. Yu, C.-Y. Lai, L. Zhou, Protocols for packet quantum network intercommunication. (2019)
2. L. Gyongyosi, L. Bacsardi, S. Imre, A survey on quantum key distribution. *Inf Commun. J.* **XI**, 2 (2019)
3. Y. Hasegawa, R. Ikuta, N. Matsuda, Experimental time-reversed adaptive Bell measurement towards all-photonic quantum repeaters. *Nat. Commun.* (2019)
4. S. Brand, T. Coopmans, D. Elkouss, Efficient computation of the waiting time and fidelity in quantum repeater chains. (2019)
5. S. Kumar, N. Lauk, C. Simon, Towards long-distance quantum networks with superconducting processors and optical links. *Quantum Sci. Technol.* **4**, (2019)
6. Z.-D. Li, R. Zhang, X.-F. Yin, L.-Z. Liu, Y. Hu, Y.-Q. Fang, Y.-Y. Fei, X. Jiang, J. Zhang, L. Li, N.-L. Liu, F. Xu, Y.-A. Chen, J.-W. Pan, Experimental quantum repeater without quantum memory. *Nat. Photonics* (2019)
7. M. Caleffi, A.S. Cacciapuoti, Quantum Switch for the quantum internet: noiseless communications through noisy channels. *IEEE J. Sel. Areas Commun. (JSAC)* (2020, 2019)
8. B. Muthu, C.B. Sivaparthipan, G. Manogaran, R. Sundarasekar, S. Kadry, A. Shanthini, A. Dasel, IOT based wearable sensor for diseases prediction and symptom analysis in healthcare sector. *Peer-to-Peer Netw. Appl.* **13**(6), 2123–2134 (2020). <https://doi.org/10.1007/s12083-019-00823-2>
9. N. Nguyen, B. Liu, V. Pham, C. Huang, Network under limited mobile devices: a new technique for mobile charging scheduling with multiple sinks. *IEEE Syst. J.* **12**(3), 2186–2196 (2018). <https://doi.org/10.1109/JSYST.2016.2628043>
10. Z. Li, G. Xu, X. Chen, Z. Qu, X. Niu, Y. Yang, Efficient quantum state transmission via perfect quantum network coding. *Sci. China Inf. Sci.* **62**, 12501 (2019)
11. A.S. Cacciapuoti, M. Caleffi, F. Tafuri, F.S. Cataliotti, S. Gherardini, G. Bianchi, Quantum internet: networking challenges in distributed quantum computing. *IEEE Network* **34**(1), 2019 (2019)
12. M. Mastriani, Is instantaneous quantum Internet possible? (2019)
13. Y. Yu, F. Ma, X.-Y. Luo, B. Jing, P.-F. Sun, R.-Z. Fang, C.-W. Yang, H. Liu, M.-Y. Zheng, X.-P. Xie, W.-J. Zhang, L.-X. You, Z. Wang, T.-Y. Chen, Q. Zhang, X.-H. Bao, J.-W. Pan,

- Entanglement of two quantum memories via fibres over dozens of kilometres. *Nature* **578**, (2019)
- 14. S. Shi, C. Qian, Modeling and designing routing protocols in quantum networks. (2019)
  - 15. K. Pomorski, R.B. Staszewski, Towards quantum internet and non-local communication in position based qubits. (2019)
  - 16. S. Rothe, N. Koukourakis, H. Radner, A. Lonnstrom, E. Jorswieck, J. W. Czarske, Physical layer security in multimode fiber optical networks. (2019)
  - 17. A. Dahlberg, M. Skrzypczyk, T. Coopmans, L. Wubben, F. Rozpedek, M. Pompili, A. Stolk, P. Pawelczak, R. Knegjens, J. d. O. Filho, R. Hanson, S. Wehner, A link layer protocol for quantum networks. (2019)

# A Comparative Analysis of Classical Machine Learning and Deep Learning Approaches for Diabetic Peripheral Neuropathy Prediction



R. Usharani and A. Shanthini

**Abstract** The most prevalent chronic complication of diabetes mellitus, diabetic peripheral neuropathy (DPN), has become an increasingly prominent public health issue. Diabetic associated complications, which affects all major organs of the body, are common Diabetes Mellitus (Liu et al. in PLoS ONE 14:1–16, 2019 [1]; Jelinek et al. in J. Diabet. Complicat. Med. 01:1–7, 2016 [2]). In this analysis, the traditional machine learning algorithms (MLA) and deep learning method Multi-Layer Perceptron (MLP) has been compared in order to predict the DPN. The DPN data obtained from different hospitals. With the help of the Google-colab environment, the ML techniques SVM, Naïve Bayes, K-Nearest Neighbor (KNN) and DL technique MLP were implemented using the Python programming language. The model has a higher accuracy of 80.07%, while using the DL MLP method. It was also compared to the other ML strategies SVM obtained with 77.0%, Naive Bayes acquiring 72.06% and yielding 69.2% using KNN. The DL and MLA comparative study of DPN forecasting shows that DL-MLP provides faster and higher DPN diagnostic performance.

**Keywords** Diabetic peripheral neuropathy prediction · Deep learning · Multi-layer perceptron · Machine learning · SVM · Naïve bayes · KNN

## 1 Introduction

DPN, which poses a risk of chronic foot ulceration an eventual lower extremity amputation, may be caused by diabetic patients. While neuropathies may stay asymptomatic for prolonged periods of time, they have a substantial impact on the quality of life of the affected patients and are a massive strain on public health systems [3].

---

R. Usharani (✉) · A. Shanthini

Department of Data Science and Business Systems, School of Computing, SRM Institute of Science and Technology, Chennai, Tamil Nadu, India

A. Shanthini

e-mail: [shanthia@srmist.edu.in](mailto:shanthia@srmist.edu.in)

Artificial Intelligence (AI) approaches can be used to detect multiple diseases. (AI) intelligence methods such as Deep Neural Network, among other things, provides the highest results in classification and prediction problems [4]. DNN (Deep Neural Networks) has been used in recent years to identify various diseases. With the assistance of the above properties, by providing a dynamic decision surface, DNN is superior to other conventional neural networks in the classification problem. In that same research, DL architecture was applied and the recognition performances in the CNN and RNN architectures were examined. The model based on LSTM cells was applied for the RNN method and the negative log-likelihood cost equation has been used using Adam optimizers for the RNN method [5].

The authors concentrate on comparing three machine learning models, namely Random Forests, Multiple Linear Regression, Vector Regression and DL techniques such as Multi-layer Perceptron, with the forecast usefulness, using probe data gathered from Thessaloniki, Greece's road network. Although the model of Support Vector Regression works well with small variations in stable conditions, the experimental results indicate that the model of Multi-layer Perceptron adapts better to situations with greater variations [6]. Ophthalmology is well suited to the implementation of AI-, ML-, and DL-assisted automated scanning and diagnosis relative to other medical specialties due to the wide use of ophthalmic images that offer an array of data for a computer algorithm. In the coming years, unmanned digital AI, ML, and DL systems will be used to diagnose and treat ophthalmic diseases as a potential alternative to ophthalmologists, retina technicians, and trained human graders.

In the assessment of ophthalmic diseases, this study investigated the prospective promising therapeutic applications of AI, ML, and DL and introduced ophthalmic imaging modalities to the newest technologies. This research offered an important and detailed review of AI, ML, and DL applications in ophthalmology by both ophthalmologists and computer scientists and encouraged other potential promising clinical applications in the ophthalmology healthcare system [7]. A method for automatic diagnosis of serious diabetic neuropathy is introduced in the current research. In addition, the primary goal of this research is the efficiency analysis of various MLAs and DL techniques for DPN disease prediction. Different variants of MLAs and DL are discussed in the following sections, accompanied by introducing the methodology of this research. The findings and discussion of the analysis are discussed in the following sections.

## 2 Related Works

This study analyzes the DM factors that are influencing with DPN using DL and ML techniques. In addition, the estimation performance of these strategies is compared. Many studies have found the insight into various algorithms and related methods to detect possible differences in this field of research [2–7]. Firstly, it is most important to determine the parameters that will assist the analysis in evaluating the source of

the DPN in order to make the data right, valid and available. Of the 20 characteristics reported for each condition, 13 were discovered to influence the DM patient's progress toward DPN. The 13 characteristics described were age, type of diabetes, degree of education, BMI, history of blood pressure, systolic blood pressure, history of foot ulcer, prescription, weight, history of laser photocoagulation, length, mean blood glucose, and height [8, 9].

High blood glucose can damage nerves in the human body by fasting plasma glucose (FPG). A syndromic link between neuropathy associated with reduced glucose tolerance and sensory-predominant neuropathy typically found in early diabetes was identified from the findings of the tests. This supports studies on the relationship between neuropathy and glycemic regulation, HbA1C, age, BMI, triglycerides, and diabetes disease duration. The coexistence of neuropathy, cardiovascular and renal comorbid conditions is demonstrated by diabetic patients with high levels of urea, irregular lipid profile, elevated urea and decreased RBC levels. The diversity of neurological diseases involving the central nervous system and the peripheral nervous system is responsible for chronic renal dysfunction. The relation between the Glomerular Filtration Rate (GFR) and microvascular complications of type II diabetes mellitus such as diabetic neuropathy and nephropathy was explored in depth. The DPN-related factors have been identified in Table 1 in the relevant studies, and the American Diabetes Association recommends targeting the regular and abnormal spectrum [10–17].

In recent years, different diseases have been identified using deep learning. With the support of the above properties, DNN is superior to other traditional MLs in the classification problem by providing a dynamic decision surface [4–7]. Using the stacked autoencoders and softmax layer, the DNN classifier for diabetes dataset is constructed. In the input layer, the DM attributes are supplied as input. In the built-in DNN, there are two auto-encoder layers. The network, with 20 neurons each, has two hidden layers. The softmax layer is appended to the last hidden layer for the classification process. The output layer will give the probabilities for diabetic and non-diabetic class data records [4]. The DL architecture was applied and the recognition experiences in the CNN and RNN architectures were analyzed in the human activity monitoring system. A model based on LSTM cells has been incorporated for the RNN method. LSTM offers the chance to learn the data's long dependencies. By using several RNN cells, all the layers are stacked. The negative log-likelihood cost function using Adam optimizers is also used using accelerometer data for human activity recognition [5]. When training the model, the authors presented the refined ensemble deep learning approach known as the 'Liverpool Deep Learning Algorithm' (LDLA). On a patch basis, the qualified models were able to generate segmentations. The trained models were able to produce segmentations on a patch basis. The segmentation of a total Corneal Confocal Microscopy (CCM) image was achieved by using the majority vote on the overlap regions to combine the segmentations of all its patches. In order to retrieve the CCM images, which are clinically significant variables of corneal nerves, further research was carried out [9]. The Multi-Layer Perceptron is used to differentiate between infected and non-infected cases as a classifier. The results of adapting the technique of ANNs to the diagnosis of these

**Table 1** DPN risk factors ranges according to review of various researches specified in the related works and American diabetes association continuum

Diabetes mellitus factors	Ranges	
	Normal	Abnormal
Blood Glucose	<140 mg/dL	>140 mg/dL
Pre-breakfast Glucose tolerance test	70–130 mg/dL	>130 mg/dL
Post-breakfast Glucose tolerance test	130–180 mg/dL	>180 mg/dL
Age	30–90 (in years)	–
BMI (Body Mass Index)	18–25 kg/m <sup>2</sup>	>25 kg/m <sup>2</sup>
HbA1C	5.7% mg/dL	>6.5% mg/dL
Triglycerides	<150 mg/dL	150–199 mg/dL
Blood Pressure (BP)	130/80 mm hg	below or higher the normal range at risk
Urea-serum-creatinine	0.6–1.4 mg/dL	>1.4 mg/dL
Glomerular Filtration Rate (GFR)	Male: 97 to 137 mL/min Female: 88 to 128 mL/min	Male: > 137 mL/min Female: > 128 mL/min
hyperglycemia-Symptoms	No (0)	Yes (1)
hypoglycemic_symptoms	No (0)	Yes (1)
microalbuminuria	30–299 mg	>=300 mg
Diabetes mellitus (duration of years)	Tolerance with under control level	Elevated level

diseases based on chosen symptoms indicate the network's capacity to learn the patterns relating to the number of chronic conditions [18–24].

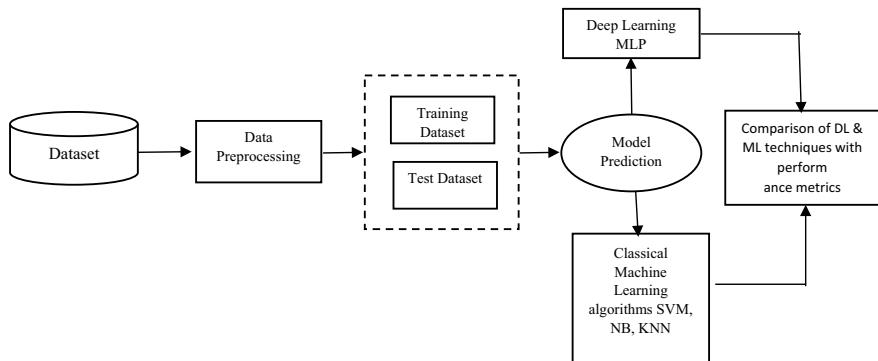
Therapeutic applications of DL and ML in this study for the diagnosis of DPN. An appropriate and thorough overview of DL and ML applications in the health care field will be included in the study findings by both physicians and data scientists.

### 3 Methodology

An overview of the DL and ML approaches used to evaluate and forecast DPN is provided in this study illustrated in the following Fig. 1. In addition, the accuracy and efficiency of the DL approach MLP have been compared with the other different ML strategies, SVM, Naïve Bayes (NB) & K-Nearest Neighbor (KNN).

#### A. Multi-layer Perceptron (MLP)

In this analysis, the Multi-Layer Perceptron (MLP) deep learning algorithm was used to evaluate the features of diabetic neuropathy and a Keras classifier to predict DPN [18–20]. The Scikit-learn library has also been used as a training

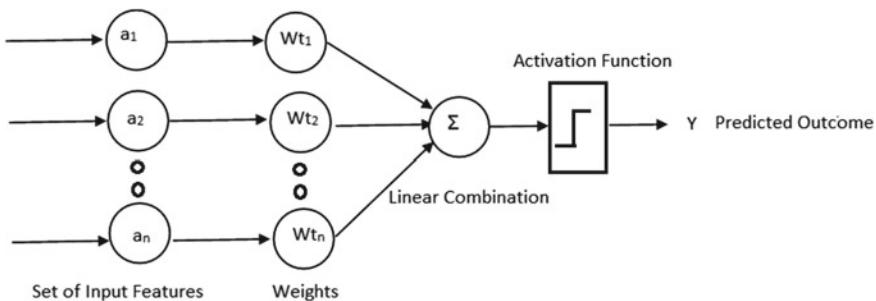


**Fig. 1** An overview of the DL and ML approaches used to evaluate and forecast DPN

set and testing set to break the dataset into two sets in this process. For the sequential function, the network uses better movement. Then, using the dense class, the three completely linked dense layers were introduced, two hidden and one output. Next, for the hidden layer, the Rectified Linear Unit (ReLU) activation function was used. For the output layer, no activation function is used because it is a regression problem that explicitly forecasts numerical values. The following Fig. 2 shows an illustration of MLP topology.

Input values ( $a_1 \dots a_n$ ) that are multiplied by weights are given ( $w_{1..n}$ ). In linear combination, these computations are summarized together and presented as feedback to the activation function. MLP topology, where weights are defined by connections for neurons of successive layers. In a linear combination of inputs and weights, the output of each layer is generated by adding an activation function. The output is obtained through the sequences of hidden layers with the observation made by the output layer.

For optimization, an appropriate ADAM optimization algorithm is used, integrated with a mean square error, binary cross entropy as a loss function. Build the neural network model as well as parameters such as the number of epochs and batch



**Fig. 2** Multi-layer perceptron (MLP) topology

size to pass forward to the model's fit () function later [21–24]. The final step is to verify and validate, using tenfold cross validation, this baseline model. To measure the model's efficiency, metric accuracy has been used. This model has been yielding a better accuracy of 80.07%.

### B. Support Vector Machine (SVM)

The kernel-based classifier, SVM, is an algorithm for supervised learning classifies data into two classes or more. Especially for binary classification, SVM is configured. SVM constructs the model in the training process, maps the determination boundary for each class, and defines the hyperplane that distinguishes the various classes [2]. With the aid of increased distance between the classes, the classification accuracy is estimated by measuring the margin of the hyperplane. On SVM, nonlinear classification can also be carried out perfectly [9, 19, 23]. Running a classification task using the Scikit-Learn SVM library's help vector classifier (SVC). A "sigmoid kernel" used for the implementation of Kernel-SVM. This model yields an accuracy of 77.0%.

### C. Naïve Bayes

Naïve Bayes is a fast and eager learning classifier. By using it, forecasts can be made in real time. The supervised learning and predictive approach are for classification and estimation. It helps to capture the complexity of the model in a principled manner by assessing the probabilities of the findings. By doing this, diagnosis and statistical problems can be overcome. The conditional likelihood function can be decomposed using Bayes' theorem as follows:

$$p(Yc|X) = \frac{p(Yc)p(Yc|X)}{p(X)} \quad (1)$$

This Bayes model yields an accuracy of 72.60%.

### D. K-Nearest Neighbor (KNN)

The KNN algorithm assumes that the new case/data is similar to the cases available and assigns the new case to a category far closer to the available categories. This algorithm stores all the available data and classifies a new data point, based on the similarities. This means that when new data appears, it can be quickly grouped into a well-suited category using the KNN process. KNN used for both regression and classification [2, 5]. This model yields an accuracy of 69.2%.

## 4 Results and Discussion

The proposed study on diabetes with DPN data collection gathered from different hospitals from 2017 to 2019 was carried out. This dataset contains age, blood glucose, triglycerides, urea-serum-creatinine, microalbuminuria, blood pressure, body mass

**Table 2** The DPN prediction performances using various MLAs

The DPN prediction performances using various machine learning algorithms			
DL-MLP (accuracy) (%)	SVM (accuracy) (%)	NB (accuracy) (%)	KNN (accuracy) (%)
80.07	77.0	72.6	69.2

index, diabetes duration in years and many others. The implementation has been carried out with Keras using the Python programming language. The dataset had further outliers and missed values excluded. The MLP concept is inspired by the human nervous system and is often referred to as “feed-forward neural networks.”

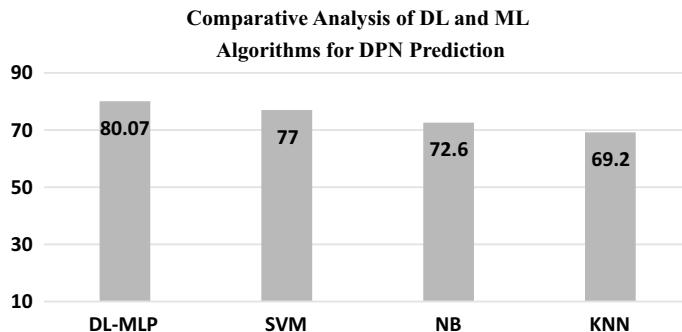
They consist of numerous coordinated neurons in layers. The number of layers is typically reduced to two or three, but technically, due to the real-time phenomenon, there is no limit. For the next layer, one layer’s outputs serve as the inputs. Each layer can be divided by an input layer, hidden layers, and an output layer. The advantages of MLP are strongly error resistant that is they continue to operate in the event of neuron failure and interconnections between them and it is nonlinear in design, so it is suitable for all types of real-world problems. In this test, used hidden layers, the Rectified Linear Unit (ReLU) activation function and the learning rate. Four classical ML classifiers, namely SVM, NB and KNN, were used after applying the MLP technique. The algorithms are tested on the basis of the accuracy of the performance metrics provided in Table 2. Accuracy is the basis for measuring any predictive model’s quality. This can be calculating the ratio of accurate estimates and the total number of measured data points. This paper consists of the best accuracies learned by DL-MLP and other different models of machine learning. For accuracy, Eq. (2) gives the equation.

Accuracy

$$= \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Negative} + \text{False Positive} + \text{True Negative}} \quad (2)$$

Table 2 shows that the highest accuracy achieved is 80.07% for MLP, 77.0% for SVM, 72.6% for NB, 69.2% yields for KNN. From this could be concluded, the performance of the DL-MLP performed better than other ML algorithms for predicting DPN.

The following Fig. 3 shows the DL-ML and other traditional ML algorithms performance analysis for DPN prediction.



**Fig. 3** Comparative analysis of DL and ML algorithms for DPN prediction

## 5 Conclusion

The research analysis was carried using the DPN dataset obtained from different hospitals. The record excluded all zero-value entries and outliers. In addition, features fed into the MLP DL technique and were implemented along with tenfold cross-validation in addition to the classical algorithms SVM, NB, and KNN. This made it possible to do thorough data analysis to evaluate the best outcome conclusively. Future studies will require a review for further examination of the dataset with other classifiers and the heatmap feature selection technique. The result obtained using the Multi-layer Perceptron classifier with a better accuracy of 80.07 percent was tested for this. It is also observed from the results that the Multi-layer Perceptron performed well compared to all the other classification algorithms in terms of all the other performance parameter accuracy. Future studies will require a review for further examination of the dataset with other classifiers and the heatmap feature selection technique.

## References

1. X. Liu, Y. Xu, M. An, Q. Zeng, The risk factors for diabetic peripheral neuropathy: a meta-analysis, *PLoS One* **14**(2), 1–16 (2019).0.1371/journal.pone.0212574
2. H.F. Jelinek, D.J. Cornforth, A.V. Kelarev, Machine learning methods for automated detection of severe diabetic neuropathy. *J. Diabet. Complicat. Med.* **01**(02), 1–7 (2016). <https://doi.org/10.4172/2475-3211.1000108>
3. I.I. Witzel, H.F. Jelinek, K. Khalaf, S. Lee, A.H. Khandoker, H. Alsafar, Identifying common genetic risk factors of diabetic neuropathies. *Front. Endocrinol. (Lausanne)* **6**(MAY), 1–18 (2015). <https://doi.org/10.3389/fendo.2015.00088>
4. K. Kannadasan, D.R. Edla, V. Kuppili, Type 2 diabetes data classification using stacked autoencoders in deep neural networks. *Clin. Epidemiol. Glob. Heal.* **7**(4), 530–535 (2019). <https://doi.org/10.1016/j.cegh.2018.12.004>

5. S.R. Shakya, C. Zhang, Z. Zhou, Comparative study of machine learning and deep learning architecture for human activity recognition using accelerometer data. *Int. J. Mach. Learn. Comput.* **8**(6), 577–582 (2018). <https://doi.org/10.18178/ijmlc.2018.8.6.748>
6. C. Bratsas, K. Koupidis, J.M. Salanova, K. Giannakopoulos, A. Kaloudis, G. Aifadopoulou, A comparison of machine learning methods for the prediction of traffic speed in Urban places. *Sustain.* **12**(1), 1–15 (2020). <https://doi.org/10.3390/SU12010142>
7. L. Balyen, T. Peto, Promising artificial intelligence–machine learning–deep learning algorithms in ophthalmology. *Asia-Pacific J. Ophthalmol.* **8**(3), 264–272 (2019). <https://doi.org/10.22608/APO.2018479>
8. F. Mao et al., Age as an independent risk factor for diabetic peripheral neuropathy in Chinese patients with type 2 diabetes. *Aging Dis.* **10**(3), 592 (2019). <https://doi.org/10.14336/ad.2018.0618>
9. M. Kazemi, A. Moghimbeigi, J. Kiani, H. Mahjub, J. Faradmal, Diabetic peripheral neuropathy class prediction by multicategory support vector machine model: a cross-sectional study. *Epidemiol. Health* **38**, e2016011 (2016). <https://doi.org/10.4178/epih/e2016011>
10. T.J. Oh et al., Association between deterioration in muscle strength and peripheral neuropathy in people with diabetes. *J. Diabetes Complications* **33**(8), 598–601 (2019). <https://doi.org/10.1016/j.jdiacomp.2019.04.007>
11. S. Ramesh, C. Yaashwanth, B.A. Muthukrishnan, Machine learning approach for secure communication in wireless video sensor networks against denial-of-service attacks. *Int. J. Commun. Syst.* **33**(12), (2019). <https://doi.org/10.1002/dac.4073>
12. N. Nguyen, B. Liu, V. Pham, T. Liou, An efficient minimum-latency collision-free scheduling algorithm for data aggregation in wireless sensor networks. *IEEE Syst. J.* **12**(3), 2214–2225 (2018). <https://doi.org/10.1109/JSYST.2017.2751645>
13. F. Booya, F. Bandarian, B. Larijani, M. Pajouhi, M. Nooraei, J. Lotfi, Potential risk factors for diabetic neuropathy: a case control study. *BMC Neurol.* **5**, 1–5 (2005). <https://doi.org/10.1186/1471-2377-5-24>
14. P.K. Bariha, K.M. Tudu, S.T. Kujur, Correlation of microalbuminuria with neuropathy in type-II diabetes mellitus patients. *Int. J. Adv. Med.* **5**(5), 1143 (2018). <https://doi.org/10.18203/2349-3933.ijam20183460>
15. P.C. Machado Aguiar, M.V. Della Coletta, J.J. Silva de Souza, The association of dyslipidemia and peripheral diabetic neuropathy: the influence of urea. *Diabetes Case Rep.* **01**(02), 2–4, (2017). <https://doi.org/10.4172/2572-5629.1000109>
16. N.C.G. Kwai, W. Nigole, A.M. Poynten, C. Brown, A.V. Krishnan, The relationship between dyslipidemia and acute axonal function in type 2 diabetes mellitus in vivo. *PLoS ONE* **11**(4), 1–12 (2016). <https://doi.org/10.1371/journal.pone.0153389>
17. M.U. Nisar, et al., Association of diabetic neuropathy with duration of type 2 diabetes and glycemic control. *Cureus*, no. September (2015). <https://doi.org/10.7759/cureus.302>.
18. M.O.G. Nayeem, M. N. Wan, and M. K. Hasan, Prediction of disease level using multilayer perceptron of artificial neural network for patient monitoring. *Int. J. Soft Comput. Eng.*, no. August (2015)
19. V. Janani, N. Maadhuryaa, D. Pavithra, S.R. Sree, Dengue prediction using (MLP) multilayer perceptron—a machine learning approach (2020)
20. P. Yildirim, Chronic kidney disease prediction on imbalanced data by multilayer perceptron: chronic kidney disease prediction, in *Proceedings of International Computer Software Application Conference*, vol. 2, pp. 193–198 (2017). <https://doi.org/10.1109/COMPSAC.2017.84>
21. M. Sultana, A. Haider, M.S. Uddin, Analysis of data mining techniques for heart disease prediction, in *2016 3rd International Conference Electrical Engineering and Information Communication Technology ICEEICT 2016*, (2017). <https://doi.org/10.1109/CEEICT.2016.7873142>
22. M.M. Kirmani, Heart disease prediction using multilayer perceptron algorithm. *8*(5), 1169–1172 (2017)

23. T.T. Hasan, M.H. Jasim, I.A. Hashim, Heart disease diagnosis system based on multi-layer perceptron neural network and support vector machine. *Int. J. Curr. Eng. Technol.* **77**(55), 2277–4106 (2017)
24. M. Durairaj, V. Revathi, Prediction of heart disease using back propagation MLP algorithm. *Int. J. Sci. Technol. Res.* **4**(8), 235–239 (2015)

# Zone-Based Multi-clustering in Non-rechargeable Wireless Sensor Network Using Optimization Technique



S. Srividhya, M. Rajalakshmi, L. Paul Jasmine Rani, and V. Rajaram

**Abstract** Real-time applications in wireless sensor network are power consuming. A lot of techniques related to various clustering and load balancing the network play a vital role. Direct diffusion is the traditional way of routing where all the non-rechargeable nodes communicate with the central server node (CSN). This will create more consumption of power. Non-rechargeable nodes' lifespan will be less, and duly it has to spend energy for both sensing and routing. The proposed optimized way of positioning the zone heads (ZHS) using satin bower bird optimization (SBO) improves the life span of the nodes in the network. The performance of the proposed work is compared with DEEC protocol in terms of the energy metric.

**Keywords** Wireless sensor networks · Zone · Clustering · Optimization

## 1 Introduction

A lot of sensors are self-organized to form a wireless sensor network. All the sensor nodes transmit the data packets to all the nodes within its radio communication range. The main features of these sensor nodes are smaller in size, less battery power, and non-rechargeable. If such nodes are deployed in unmanned regions for real-time monitoring, the status of the node is considered mainly [1]. The nodes are in live state or dead state is predicted by the battery power. Lots of research works are carried out to withstand the power level of the nodes by reducing the workloads. Prolonging

---

S. Srividhya (✉) · M. Rajalakshmi

School of Computing, SRM Institute of Science and Technology, Chennai, India  
e-mail: [srividhs1@srmist.edu.in](mailto:srividhs1@srmist.edu.in)

L. P. J. Rani

Department of Computer Science and Engineering, Rajalakshmi Institute of Science and Technology, Chennai, Tamil Nadu, India

V. Rajaram

Department of Information Technology, Sri Venkateswara College of Engineering, Sriperumbuthur, India

the lifespan of the network will be achieved by the proper optimized technique [2] for communication.

The rest of the work is narrated as follows. Part 2 relates different survey related to the clustering using optimization. Part 3 describes the assumptions taken for the simulation. Part 4 depicts the proposed architecture design. Part 5 describes the work flow of entire protocol. Part 6 speaks about results and the conclusion summarizes the work.

## 2 Literature Survey

Plenty of protocols are there related to optimizing the energy consumption of the nodes in the network. One of the traditional cluster-based protocol is called LEACH [3]. Here, the nodes are grouped together into numerous clusters. Each cluster is dominated by one head node. These head nodes are selected using the threshold value. Saeid et al. in [4] done improvement in the LEACH where sink nodes are movable in nature. Routing will be dynamic based on the location of the sink nodes at that time.

In [5], genetic algorithm is employed for the routing and clustering. Distance between the nodes with the sink and remaining energy of the node is considered to find the appropriate cluster heads for routing. Genetic algorithm-based approach enhances the performance of the network.

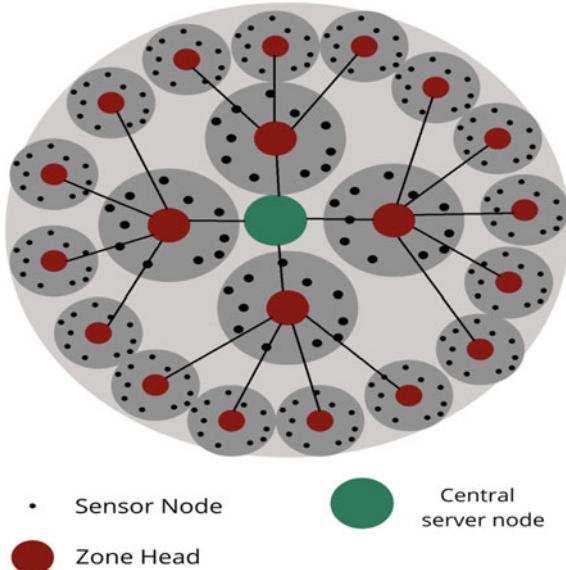
Particle swarm optimization is involved for defining multi-objective function based on node characteristics. In [6–8], Azharuddin et al. took failure scenario of the nodes to solve with particle encoding technique. Power consumption of the nodes and reliability of the communication have been balanced here.

Dervis et al. demonstrated the impact of ant colony optimization for suitable clustering using simulation in [9]. Clusters are framed using the ant pattern, and the performance of the work was compared with PSO optimization algorithm.

## 3 Architecture

The general structure of WSN communication is shown in Fig. 1. The network area is considered as circular region. Huge volumes of sensors are deployed in the network area randomly. The central server node is actually located at the center part of the region. The network region is again divided into several circular zones. Each and every zone has its own head for aggregating and forwarding the data packets. The optimized location of zone heads is identified by the satin bowerbird optimization algorithm.

**Fig. 1** WSN Zone-based architecture



## 4 Assumptions

The following are the assumptions taken for the simulation.

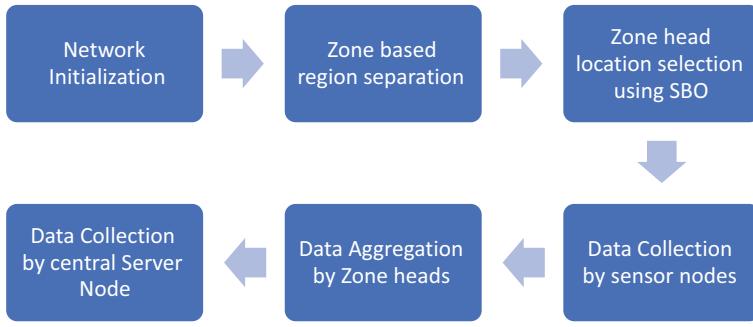
- All the sensor nodes are static in nature. It means that nodes are not movable.
- All the nodes homogeneous in nature.
- All the sensor nodes are initialized with power 1 J.
- The central server node is located at the center of the network area.
- All zone heads aggregate the data packets and forward to reach to the central server node.

## 5 Proposed Work

The flow of proposed ideas is depicted in the Fig. 2.

### 5.1 Energy Model

Power consumption by the nodes and zone heads are calculated by the energy free space formula. It is represented below as referred from [10]. The transmission energy is given by Eq. 1. Equation 2 is used for reception of data packets.



**Fig. 2** Proposed modules

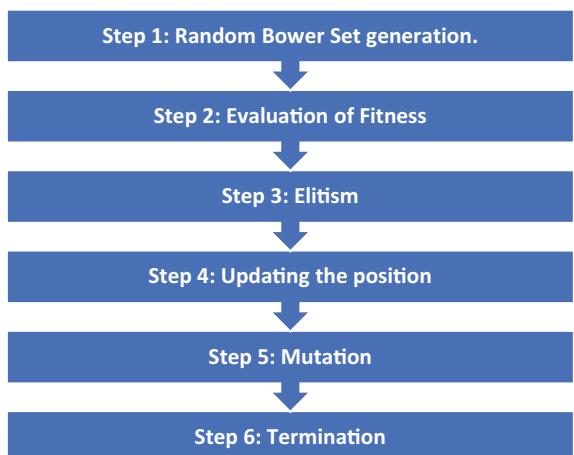
$$E_{tx} = \begin{cases} n * E_{elec} + n * \varepsilon_f * d^2 & \text{when } d \leq d_0 \\ n * E_{elec} + n * \varepsilon_{amp} * d^4 & \text{when } d \geq d_0 \end{cases} \quad (1)$$

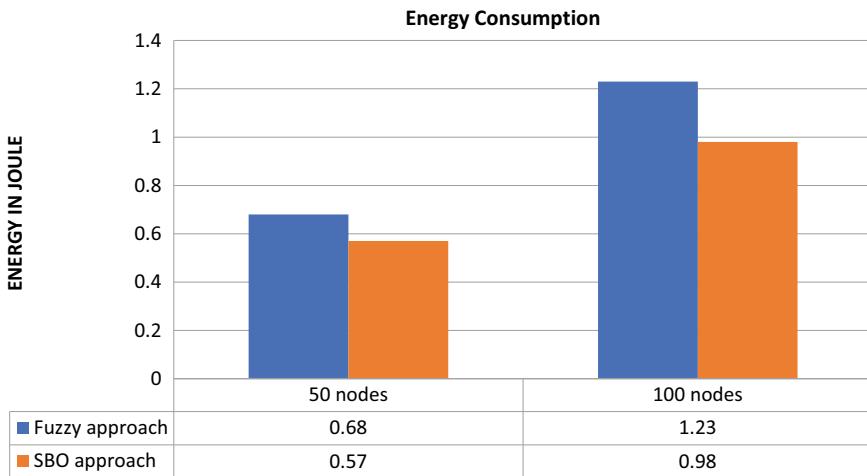
$$E_{rep} = n * E_{elec} \quad (2)$$

## 5.2 Position Update for Zone Head—SBO

Each and every zone head is selected based on its location using SBO [11] in Fig. 3. Updation in the position is based on the objective function. The objective function is given by minimum of  $F_1$  and  $F_2$ . The function  $F_1$  is given by the average distance between node to zone head and from zone head to CSN. The function  $F_2$  is based on the total residual energy of the all the zone heads.

**Fig. 3** SBO for Zone head location update





**Fig. 4** Energy consumption of proposed work

## 6 Experiments and Results

The proposed zone-based clustering using SBO is simulated with 50 nodes in the network area. The network area is  $100\text{ m} \times 100\text{ m}$ . It is compared with existing protocols using fuzzy-based head selection [12]. The proposed approach is simulated in MATLAB2020a with initializing static nodes.

**Energy consumption:** Energy is represented in joule. Initial energy of all the nodes is taken as 1 J. Total energy spend by the sensor nodes is the energy needed to transmit the sensed data. Total energy spent by zone heads are together for receiving, aggregating, and transmitting the packets. Average energy consumption of the network is visualized in the Fig. 4.

## 7 Conclusion

The effect of optimization in zone-based clustering reduces the energy consumption of the nodes. In turn, the life span of the network gets increased. The location of the zone head has to be in optimized location which will support to communicate with central server node and along with the member nodes. In the proposed work, satin bowerbird optimization is used to update the position of the zone heads.

## References

1. S. Chaudhary, N. Singh, A. Pathak, A.K. Vatsa, Energy efficient techniques for data aggregation and collection in WSN. *Int. J. Comput. Sci. Eng. Appl.* **2**(4), 37 (2012)
2. V. Rajaram, N. Kumaratharan, Multi-hop optimized routing algorithm and load balanced fuzzy clustering in wireless sensor networks. *J. Ambient. Intell. Humaniz. Comput.* **5**, 1–9 (2020)
3. W.R. Heinzelman, A. Chandrakasan, H. Balakrishnan, Energy-efficient communication protocol for wireless microsensor networks, in *Proceedings of the 33rd annual Hawaii international conference on system sciences*, p. 10. (IEEE, 2000)
4. S. Mottaghi, M.R. Zahabi. Optimizing LEACH clustering algorithm with mobile sink and rendezvous nodes. *AEU Int. J. Electron. Commun.* **69**(2), 507–514 (2015)
5. S.K. Gupta, P.K. Jana, Energy efficient clustering and routing algorithms for wireless sensor networks: ga based approach. *Wireless Pers. Commun.* **83**(3), 2403–2423 (2015)
6. Ramesh, S., Yaashwanth, C., & Muthukrishnan, B. A. (2019). Machine learning approach for secure communication in wireless video sensor networks against denial-of-service attacks. *International Journal of Communication Systems*, 33(12). doi:<https://doi.org/10.1002/dac.4073>
7. B.-H. Liu, N.-T. Nguyen, V.-T. Pham, Y.-X. Lin, Novel methods for energy charging and data collection in wireless rechargeable sensor networks. *Int. J. Commun. Syst.* **30**:e3050<https://doi.org/10.1002/dac.3050>
8. M. Azharuddin, P.K. Jana, PSO-Based approach for energy-efficient and energy-balanced routing and clustering in wireless sensor networks. *Soft Comput.* (2016)
9. M.H. Alsharif, S. Kim, N. Kuruoğlu, Energy harvesting techniques for wireless sensor networks/radio-frequency identification: a review. *Symmetry* **11**(7), 865 (2019)
10. B. Baranidharan, S. Srividhya, B. Santhi, Energy efficient hierarchical unequal clustering in wireless sensor networks. *Indian J. Sci. Technol.* **7**(3), 301 (2014)
11. M. Afsar, H. Mohammad, N. Tayarani, M. Aziz, An adaptive competition-based clustering approach for wireless sensor networks. *Telecommun. Syst.* **61**(1), 181–204 (2015)
12. V. Rajaram, S. Srividhya, N. Kumaratharan, V. Ganapathy, Fuzzy logic based unequal clustering in wireless sensor networks for effective energy utilization, in *2017 IEEE International Conference on Power, Control, Signals and Instrumentation Engineering (ICPCSI)* (pp. 2056–2061). (IEEE, 2017)

# A Review and Design of Depression and Suicide Detection Model Through Social Media Analytics



Michelle Catherina Prince and L. N. B. Srinivas

**Abstract** The COVID-19 pandemic has swept across the nations in the year 2020 and has affected people's daily lives by causing tremendous loss of lives and economy. This has gravely affected the mental health of people, plunging them into depression and suicide. The sooner depression and suicidal tendencies are detected, the easier it is to battle them. In the age of social media, an enormous number of people with mental health issues resort to social media to express their feelings without the fear of being judged. Due to this reason, it is possible to detect mental illnesses from social media data. This paper aims to design and develop a depression and suicide detection model, using tweets from Twitter to detect and avert depression and suicidal tendencies during the pandemic, using social network analytics. Patterns identified will aid psychologists, psychiatrists and hospitals with deeper cognizance to improve their care and diagnostics.

**Keywords** Depression · Suicide detection · Twitter · Machine learning · Social media analytics

## 1 Introduction

The COVID-19 pandemic reportedly originated on December 31, 2019, in Wuhan, China, which has a population of eleven million people and is the fulcrum of economics and business in Central China. Then, COVID-19 spread to the neighboring countries of Japan, Korea, Thailand and the rest of the world. The World Health Organization (WHO) officially declared the outbreak of COVID-19 as a global pandemic on March 11, 2020. As of October 30, 2020, WHO has reported 445,92,789 cases of COVID-19 with 11,75,553 deaths in over 219 countries. This has led to closure

---

M. C. Prince (✉) · L. N. B. Srinivas

Department of Information Technology, College of Engineering and Technology, SRM Institute of Science and Technology, SRM Nagar, Kattankulathur, Chennai, Tamil Nadu, India

L. N. B. Srinivas

e-mail: [srinival@srmist.edu.in](mailto:srinival@srmist.edu.in)

of educational facilities, nation-wide lockdown of businesses, travel restrictions and stay-at-home orders. The pandemic has not only taken away the lives of people but has also caused economic crises, business closures and job losses. This has caused emotional and mental turmoil among people.

Depression is a mental illness that has affected more than 264 million people across the globe. Depression rates have increased drastically due to the COVID-19 pandemic. Studies show that an estimate of 56 million people are suffering from depression in India alone. Infact, depression and mental illnesses are a major contributor to suicide. In India, 1,39,123 suicides were reported in the year 2019.

Many individuals hesitate to seek professional help due to the social stigma surrounding depression and as a result, they resort to social media for comfort and emotional peace. This explosion of social media usage has welcomed individuals to share their experiences, feelings, emotions and challenges of their mental state through online blogs, posts, micro-blogs, comments and tweets. The collection and analysis of these posts can aid in the detection of depression and suicidal tendencies and avert major catastrophes in an individual's life.

In this paper, we propose a novel model for detecting depression and suicidal tendencies using real-time data from Twitter. Data is collected from Twitter using the Tweepy API. These tweets are preprocessed with natural language processing techniques such as stemming, lemmatization and stop word removal. Proper feature selection and multiple combinations of features are essential for improving the accuracy in the detection of suicidal and depressive tendencies from social media tweets.

Hence, Linguistic Inquiry and Word Count (LIWC), sentiment analysis, N-gram model, Latent Dirichlet Algorithm (LDA) and depression ideation metrics are the major features that are going to be used for detecting depressive tendencies in this model. Linguistic Inquiry and Word Count (LIWC), sentiment analysis, Latent Dirichlet Algorithm (LDA), emotion analysis and suicide ideation metrics are the features that are going to be extracted for detecting suicidal tendencies.

In order to detect depression and suicidal tendencies from social media, convolutional neural networks, BERT, RoBERTa and autoregressive language model XLNet are the classifiers which will be used to detect the depressive and suicidal tendencies of tweets. We will compare these deep learning models and transformer-based models to determine which classifier works best for our model.

As depressive people often show suicidal tendencies, if a tweet is depressive, we further analyze it to see if it also shows suicidal tendencies by using the machine learning classifiers and transformer-based models.

## 2 Related Works

### 2.1 *Background*

Using social media as a tool for detecting mental illnesses gained popularity as people started expressing themselves on social media and other online platforms.

### 2.2 *Detecting Depression Stigma on Social Media*

Paper [1] demonstrates the potential of the social media platform Twitter, for predicting depression among individuals. A dataset was created from Twitter users containing a public profile of those who were diagnosed with clinical depression, through crowdsourcing method. The social media data of these users contain tweets for over a year preceding their depression was collected for analysis. The user engagement on social media, the time of the posting, ego network, usage of common depression terminology, emotions, linguistics, language and usage of anti-depression drug terminologies were used to measure behavioral attributes. These attributes were leveraged to build a supervised learning model and it was found that SVM classifier predicted the best with 70% accuracy. It was found that depressed individuals show lower social activity, display pessimistic emotions, are very self-centered, show increased anxiety in relational and medicinal concerns and find solace in religious activities. The challenge, however, was that due to the limited availability of data, it was difficult to prove the fact that depressed people are more connected to other depressed individuals.

Paper [2] displays the detection of depression through Facebook comments on posts using NCapture tool. The popular Linguistic Inquiry and Word Count (LIWC) dictionary has been used to encapsulate the psycholinguistic features of these comments. Depressive behaviors were measured using emotions, linguistics and temporal processes. Based on these features, a prediction model was built using four machine learning techniques: Support Vector Machine, K-nearest neighbor, ensemble and Decision Tree classifier to detect depression on social media. It was found that Decision Tree classifier outperforms other classifiers and Support Vector Machine was a close second for this FB dataset. It was observed that people showed depression at the AM time over the PM time and Facebook users were found to be depressed primarily because of their personal problems. There were two challenges faced by the authors. (i) Building the dataset was a manual process; it was tedious to find comments that were depressive in nature. (ii) This study uses a supervised learning mechanism and analyzes the Facebook comments for depressive tendencies. It does not identify a depressed individual.

Paper [3] proposes a new model, namely Multimodal Depression dictionary learning model (MDL) for detecting depression by harnessing social media data.

The authors argued that building a dataset using questionnaires and assessment is more expensive and time-consuming.

Hence, they built three data sets from Twitter using Twitter API which contained a testing dataset, depressed dataset and non-depressed dataset. The features extracted comprised of the features associated with the online social network, the features in the profile of users, emotions, visual features, specialized topic associated features, and depression-specific terminologies. Each extracted feature group was taken as a single modal. Through the amalgamation of the different models, a Multimodal Depressive dictionary Learning model (MDL) was created and through this, the representation of users was learned. Through experimental results, it was found that MDL outperformed the Naive Bayes, Wasserstein Dictionary Learning (WDL) model, and Multiple Social Networking Learning (MSNL) model. Patterns in depressed users were discovered and it was found that depressed users have more chances of posting tweets between 11 pm and 6 am. They portray more negative emotions and depression-related terminologies on social media and they are more self-aware.

In order to comprehensively understand an individual, Paper [4] illustrates a novel analytics framework created by collecting large amounts of data from Facebook and determining whether or not a particular Facebook user shows depressive tendencies, through machine learning techniques. It focusses on semantic, syntactic and pragmatic features to detect depression. Data was collected through my Personality dataset and the Facebook statuses of the users. It was an amalgamation of raw user data from social media as well as data from questionnaires. Features were extracted using LIWC tools, social network metrics, user personality, influence of friends and intention of the users. These features were extracted and tested on seven different machine learning algorithms namely Random Forest, Logistic Regression, Neural Network, K-nearest neighbor, Naive Bayes, and Support Vector Machine. Classification and Regression Trees (CART) was used in modeling the intentions of the users and its output was fed into the machine learning classifiers. The machine learning algorithm, Random Forest yielded the highest accuracy at 71%. The authors emphasize the importance of choosing the right features for analysis. Pragmatic feature extraction such as the intention of the user and the social influence of the user was introduced in this paper; however, incorporating them would be a challenging task.

Another popular social media platform is Reddit and data from it has been used to detect depressive tendencies. Paper [5] examines Reddit user's posts to detect depressive behaviors using natural language processing techniques and machine learning approaches. A pre-existing depression dataset was used for analysis. The dataset was preprocessed using NLP operations and features were extracted using LIWC dictionary, LDA and N-gram models. This paper shows the improvement of performance through proper feature selection as well as the combination of different selected features. The framework was developed with Logistic Regression, Support Vector Machine, Random Forest, Adaptive Boosting, and Multilayer Perceptron Classifier. It was concluded that multiple feature selection along with the Multilayer Perceptron Classifier demonstrated the highest accuracy at 91%. The best individual feature was a bigram with Support Vector Machine classifier as it was capable of detecting

depression with 80% accuracy. The challenge is the existence of absolute values in the metrics and hence, this is a tedious process.

### ***2.3 Detecting Depression on Multilingual Data from Social Media***

As social media engagement started increasing, social media platforms were being built to accommodate the non-English speaking population also. Weibo is a popular Chinese social media platform and in Paper [6], efficient detection of depression stigma through linguistic analysis was performed on Weibo. A content analysis was conducted to determine whether a post had depression stigma associated with it or not. After this, four machine learning algorithms—Support Vector Machines, simple Logistic Regression, Multilayer Perceptron Neural Network and Random Forest were used to build two groups of classification models based on linguistic features. It has been concluded that the use of linguistic features improves the performance and the detection of online stigmas and can help in programs that reduce the prevalence of such stigma. However, the only drawback was the limited and imbalanced availability of the dataset.

Popular social media platform, Twitter, incorporated other languages apart from English on its platform and this multilingual feature has increased the number of users. This in turn has led to the analysis of Twitter data in native languages. Paper [7] claims to be the first study that uses Arabic Twitter data from the Gulf region to diagnose online users with depression.

The study aims to capture those suffering from both clinical depression as well as from online depression. A corpus was built containing depressed tweets from users who have called themselves depressed and non-depressed tweets from users who have shown no signs of depression. The Weka tool was used to convert tweets into feature vectors and the features were extracted using unigrams and negation handling methods. Based on these features, a predictive model was built on four supervised learning algorithms and it was found that the Liblinear model predicted with the highest accuracy of 87.5% for Arabic data. The major challenge was in performing Arabic sentiment analysis as the underlying process had to be curtailed based on the differences in the language.

### ***2.4 Detecting Suicidal Tendencies and Other Mental Illnesses on Social Media***

Predicting national suicide rates has become possible through the analysis of social media data. Suicide is influenced by social factors, individuals and the environment. In paper [8], suicide-related and dysphoria-related weblogs from Naver were

collected over a period of 3 years from individuals who had committed suicide in South Korea. The purpose of the data collection was to find a relationship between suicides and social media activity. Suicide-related and dysphoria-related weblogs were the two primary social media variables used along with the classical, social, metrological and economic variables. Using univariate and multivariate models, a correlation between suicide and celebrity suicides was analyzed and it was observed that there was a sharp increase in suicide weblogs in close proximity to celebrity suicides. However, celebrity suicide did not affect the dysphoria counts.

Machine classifiers could detect and come to an agreement with human coders on suicide-concern tweets. A study was conducted in Paper [9] on Twitter data to see if machines could detect suicide-concern tweets as efficiently as human coders could. The dataset was derived from Twitter using an API containing suicide ideation phrases and terms. The machines used unigrams and term frequency-inverse document frequency (TF-IDF) to analyze the linguistics while Support Vector Machine and Logistic Regression were the machine learning classifiers that were used. It was determined by human coders and machine classifiers that humans express their suicidal concerns online. 14% of the tweets were found to be ‘strongly concerning’ suicidal tweets by the human coders and the computers were able to match this with an 80% accuracy. Although the machine showed 80% accuracy, the scope for improvements exists. The challenge expressed in this study is that it is impossible to retrieve all data since a lot of Twitter users are not public. Hence, suicide prevention through social media seems ambiguous. Also, certain suicide ideation phrases are vernacular utterances of a particular community or age group.

Paper [10] uses social media to predict people suffering from Major Depressive Disorders (MDD). Twitter data was extracted from crowdsourcing methods and compiled to a list of people who identify themselves as depressed or have post-traumatic stress disorder (PTSD). Each tweet is examined with its own document and is quantified using the bag of words approach. These quantified tweets were fed into four different classifiers, and was used to develop, compare and contrast different major depressive disorders classifiers from the signals derived. It was found that the unigram-based Naive Bayes approach showed 86% accuracy in detecting depression. Based on these findings, a multinomial approach to naive Bayes classifier can be further explored and developed.

Through the collection of public data on Twitter, Paper [11] analyzes for the presence of four major mental disorders, namely depression, bipolar disorder, post-traumatic stress disorder (PTSD) and seasonal affective disorder (SAD) among public Twitter users. On this data, LIWC was used to gather data quantitatively. Unigram model, 5-g language model, life analytics such as insomnia, positive feelings, negative sentiments and physical activity on tweets were recorded. These were classified with a Log Linear classifier to analyze the above-mentioned mental disorders. It was made clear from their findings that social media data can aid in understanding and analyzing the mental health of people. However, only a subpopulation suffering from the disorder is brought to light as not every individual freely posts

about their status of their mental health online. It was also observed that some individuals post fake tweets and distinguishing between fake and real tweets is a tedious process.

The objective of Paper [12] is to detect suicidal tendencies on the social media platform, Reddit through deep learning models and compare its performance with machine learning algorithms. A pre-existing suicide dataset of Reddit is used in this study. For the comparison between deep learning and machine learning models, two different frameworks were created, one for each model.

In the first framework, the Reddit dataset was preprocessed, and features were extracted using NLP techniques and traditional machine learning classifiers were used to process this data. The baseline models used were Random Forest, Support Vector Machine and XGBoost. In the second framework, the Reddit data was preprocessed, feature extraction was performed using word embedding and was classified using deep learning models. The deep learning model used was LSTM-CNN model which is a combination of two deep learning models. It was found that the deep learning model of LSTM-CNN outperformed the baseline models in suicide ideation detection as this hybrid model combines the strengths of both LSTM and CNN models. However, the limitation of this study was the deficiency in data and the biases present in the dataset. Also, building a manual dataset could lead to human errors.

Paper [13] discusses the different suicide ideation detection techniques available. There are three major factors for suicide; they are health, psychological and historical factors. Suicide detection can be performed in various methods, namely clinical methods, content analysis, feature engineering, affective characteristics and deep learning methods. In clinical methods, suicidal tendencies can be detected by taking physical tests as well as understanding the psychology of the patient. This process, however, will take many face-to-face visits with the doctor. Content analysis refers to the linguistic, statistical, lexicon-based and topic-related analysis on the user's content to detect suicidal tendencies. In feature engineering, NLP and machine learning classifiers are applied to understand whether the text has suicidal tendencies associated with it. Questionnaires and feature extraction from raw data are the examples of feature engineering. Affective characteristics determine the emotion in the suicide posts or notes. Advanced paradigms can be combined with deep learning models as they boost performance, understand and detect suicide better. Suicide detection can be applied on questionnaires, an individual's medical records, suicide videos, web blogs and notes and on online content such as social media. In order to detect suicide, a standard to classify the severity of the risk of suicide should be developed. The challenges of suicide detection are the availability of unsupervised training datasets, imbalance in data, biases in the data and lack of understanding the intent of the suicide attempter.

## 2.5 *Understanding the Mental Health of People During the COVID-19 Pandemic*

The mental health of people has been gravely affected due to the COVID-19 pandemic. Paper [14] examines the effect of depression during the COVID-19 pandemic by building a depression user's dataset from Twitter containing 2575 tweets. A linguistic analysis is performed using LIWC and demographic information. Also, social engagement of the users and the sentiments of the tweets are analyzed. This paper investigated the potential of transformer-based deep learning models such as bidirectional encoder representations from transformers (BERT), RoBERTa, XLNet and it was observed that they invariably outperform BiLSTM and convolutional neural network (CNN) regardless of the size of our training set. It was also observed that the performance of the machine learning classifier increases with the size of the training-value dataset. Based on these results, an application was built by XLNet to monitor the depression of the United States of America at the country level and the state level by building a tool merging deep learning models and psychological text analysis. It was found that in January, there was a drop in depression scores due to winter. Also, it was observed that the more the people discuss about COVID-19 the higher depression signals arise and eventually, the signals drop.

Paper [15] adopts a sentiment analysis method to understand the fluctuations of the mental health of people during the COVID-19 pandemic. A study on anxious depression which is an amalgamation of anxiety and depression was studied in quarantined individuals. Data of 1278 quarantined individuals were extracted with 214,874 tweets spanning a time of four weeks, which includes data before, during and after quarantine. Also, a dataset of 250,198 tweets was collected from people who were not quarantined. Lexicon-matching was used to analyze the tweets of both groups to understand depression among people. Text analysis was performed using features on MATLAB. Clear patterns were discovered for those who were quarantined. Depression and anxiety were common among people when quarantine began and it eventually diminished. However, it was found that anxiety and depression resurfaced at the end of the 14-day quarantine period. Also, quarantine has negatively affected the mental health of people. Due to these reasons, it is essential to introduce emotional and mental health management for the well-being of the mental health of individuals during quarantine and pandemic times. The major challenge in this study is the tedious collection of data, which is collected before, during and after the quarantine period. The overview of mental illnesses on social media is summarised in a Table 1.

## 3 Proposed System

Due to the increase of depression and suicidal rates during the pandemic, we aim to extract, analyze and examine the effect of depression and suicidal behavior on

**Table 1** Overview of mental health illnesses on social media

S. No.	Category	Inference
1	Detecting depression stigma on social media	Most depression detection frameworks have been created to aid health workers and clinicians to provide necessary treatments. Feature extraction is important as it directly influences the performance of the model. After essential features are extracted, different machine learning models and classifiers are used to determine depressive traits.
2	Detecting depression on multilingual data from social media	Non-English social media are quite unexplored areas in computational linguistics with their own conception of mental health and social stigma worthy of further study. As expressing feelings is different in different languages, sentiment analysis is commonly used for depression detection in non-English languages
3	Detecting suicidal tendencies and other mental illnesses on social media	It was found that machine classifiers could detect and come to an agreement with human coders on suicide-concern tweets. It was observed that suicide is influenced by society, depression and unemployment rates. Sentiment analysis and emotion detection methods are some of the common methods for detecting suicide and other mental illnesses.
4	Understanding mental health of people during the COVID-19 pandemic	The classification performance improves with the increase in size of our train-Val set. Transformer based models are extremely popular and it has been found that BERT, RoBERTa and XLNet invariably outperform BiLSTM and CNN regardless of the size of our training set. Also, there is a significant increase in depression signals as people talk more about COVID-19

tweets during the COVID-19 pandemic. A novel model is created by analyzing a tweet for depressive behavior, if the tweet shows depressive tendencies to analyze it further for suicidal tendencies. This model can be used to understand the mental health of people during pandemics and crises. The model also aims to understand the relationship between depression and suicide through social media analytics. The detection of depression and suicidal tendencies can be used by clinicians and analysts to understand the mental health of people during a pandemic and provide adequate treatment required.

### ***3.1 Dataset***

This paper proposes to use a testing dataset retrieved from Twitter using the hashtag #covid19 for a period of five months during the COVID-19 outbreak. A supervised training dataset for depression and suicide will be extracted from Twitter using the Twitter API.

### ***3.2 Dataset Extraction from Twitter***

Datasets are extracted from Twitter during the onset of the COVID-19 pandemic using Tweepy API for a period of five months.

### ***3.3 Data Preprocessing***

The tweets are cleaned and preprocessed using natural language processing (NLP) techniques such as stop word removal, URL removal, punctuation removal, lemmatization and stemming.

### ***3.4 Feature Extraction for Depression***

Proper extraction of features is critical for depression analytics. Hence, the feature extraction models used are:

1. Linguistic Inquiry and Word Count (LIWC) a psycholinguistic dictionary used for semantic and syntactic feature extraction
2. Sentiment analysis to understand the positive and negative sentiments of a tweet
3. LDA topic modeling for content analysis
4. N-gram modeling to understand the linguistic features
5. Depression ideation features, namely depression terminologies, antidepressants occurrences and Insomnia Index

Models with both single features and a combination of features will be used to deduce the most accurate model.

### ***3.5 Machine Learning Classifiers for Depression***

The machine learning classifiers and deep learning algorithms such as Support Vector Machine (SVM), XLNet, convolutional neural networks (CNN), BERT and

RoBERTa will be used to classify the tweets and determine if the tweet has depressive traits or not.

### ***3.6 Analysis of Results***

The tweet will be analyzed to determine if it has depressive traits associated with it. If the tweet does not have depressive traits, the tweet will be discarded else it will be taken for further processing.

### ***3.7 Feature Extraction for Suicide***

The processed depression positive tweet will be analyzed to determine if it shows traits of suicidal tendencies associated with it. Feature extraction models used to detect suicidality in tweets are:-

1. Sentiment analysis to determine positive and negative sentiment
2. Linguistic Inquiry and Word Count (LIWC) for semantic and syntactic feature analysis
3. LDA topic modeling
4. Emotion analysis to understand the feelings and emotions associated with the tweet
5. Suicide ideation metrics such as swear words and suicidal terminologies

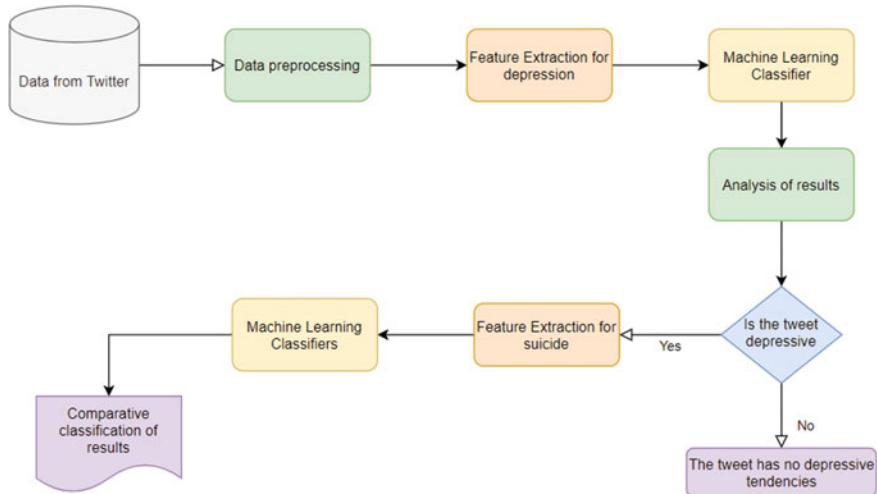
### ***3.8 Machine Learning Classifier for Suicide***

Once the important features are extracted, we apply Support Vector Machine (SVM), BERT, RoBERTa and XLNet data models to classify the tweet and determine whether the tweet has suicidal tendencies associated with it. The different data models will be used for analysis and comparison.

### ***3.9 Comparative Classification of Results***

The tweets will be analyzed to understand the relationship between the depressive and suicidal traits in it. The detection of depression and suicidal tendencies can be used to understand the mental health of people during a pandemic.

Figure 1 explains the architecture of the proposed system. It depicts a step-by-step process from the data collection of tweets to the comparative analysis of the results.



**Fig. 1** Architecture of proposed system

## 4 Conclusion

It has been observed that due to the explosion of social media revolution, people have expressed their feelings of joy, pain, depression and hardships freely on social media platforms. This has paved way to understand mental health disorders through the analysis of social media data. A review of depression models, suicide models and other mental health-related models have been analyzed in this paper.

The integration of depression and suicide detection model on social media offers a new direction for understanding the mental health of individuals during a pandemic. As research has shown a spike in feelings of depression during pandemics, a model is proposed to analyze the prevalence of depression and suicide tendencies on tweets during the COVID-19 pandemic. This will help in understanding the feelings of people during pandemics and crises.

We hope that researchers and mental practitioners will find this model useful to raise awareness of the mental health impacts of the pandemic. We also believe that our model will be a tool to help doctors, psychologists and clinicians as a tool for understanding the mental health of people through their online activities and provide adequate treatment.

## References

1. M. De Choudhury, M. Gamon, S. Counts, E. Horvitz, Predicting depression via social media, in *Proceedings of the International AAAI Conference on Web and Social Media 2013* Jun 28 (Vol. VII, No. I)
2. M.R. Islam, M.A. Kabir, A. Ahmed, A.R. Kamal, H. Wang, A. Ulhaq, Depression detection from social network data using machine learning techniques. *Health Inf. Sci. Syst.* **6**(1), 8 (2018)
3. G. Shen, J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T.S. Chua, W. Zhu, Depression detection via harvesting social media: a multimodal dictionary learning solution, in *IJCAI 2017* Aug 19 (pp. 3838–3844)
4. X. Yang, R. McEwen, L.R. Ong, M. Zihayat, A big data analytics framework for detecting user-level depression from social networks. *Int. J. Inf. Manage.* **1**(54), 102141 (2020)
5. M.M. Tadesse, H. Lin, B. Xu, L. Yang, Detection of depression-related posts in reddit social media forum. *IEEE Access.* **4**(7), 44883–44893 (2019)
6. A. Li, D. Jiao, T. Zhu, Detecting depression stigma on social media: a linguistic analysis. *J. Affect. Disord.* **1**(232), 358–362 (2018)
7. S. Almouzini, A. Alageel, Detecting Arabic depressed users from twitter data. *Procedia Comput. Sci.* **1**(163), 257–265 (2019)
8. H.H. Won, W. Myung, G.Y. Song, W.H. Lee, J.W. Kim, B.J. Carroll, D.K. Kim, Predicting national suicide numbers with social media data. *PLoS One* **8**(4), e61809 (2013)
9. B. O’dea, S. Wan, P.J. Batterham, A.L. Calear, C. Paris, H. Christensen, Detecting suicidality on twitter. *Internet Interventions* **2**(2), 183–8 (2015)
10. M. Nadeem, Identifying depression on twitter. arXiv preprint [arXiv:1607.07384](https://arxiv.org/abs/1607.07384) (2016)
11. G. Coppersmith, M. Dredze, C. Harman, Quantifying mental health signals in twitter, in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From linguistic Signal to Clinical Reality*, 2014, (pp. 51–60)
12. M.M. Tadesse, H. Lin, B. Xu, L. Yang, Detection of suicide ideation in social media forums using deep learning. *Algorithms* **13**(1), 7 (2020)
13. S. Ji, S. Pan, X. Li, E. Cambria, G. Long, Z. Huang, Suicidal ideation detection: a review of machine learning methods and applications. *IEEE Trans. Comput. Soc. Syst.* (2020)
14. Y. Zhang, H. Lyu, Y. Liu, X. Zhang, Y. Wang, J. Luo, Monitoring depression trend on twitter during the COVID-19 pandemic. arXiv preprint [arXiv:2007.00228](https://arxiv.org/abs/2007.00228) (2020)
15. W. Lu, L. Yuan, J. Xu, F. Xue, B. Zhao, C. Webster, The psychological effects of quarantine during COVID-19 outbreak: sentiment analysis of social media data. Available at SSRN 3627268 (2020)

# Artificial Intelligence Used in Accident Detection Using YOLO



S. V. Gautham and D. Hemavathi

**Abstract** Computer-based accident detection carried out using video surveillance which has become advantageous but a difficult job. Recent framework has been proposed to detect accidents of the vehicles. This framework utilizes YOLO for detection of vehicle accidents with better efficiency and would avoid false alarms. In this framework, detection is done based on speed and tangent of vehicle with respect to another vehicle. The purpose of this framework is for better accuracy and speed alert generation and avoiding false alarm due to various parameters that include daylight, dark visibility, weather conditions like rainstorm, hail, etc. When compared to other methodologies, YOLO has performed far better results in accident detection using the CCTV surveillance videos. This framework was relatively used for development in vehicle accident detection in present day-to-day life.

**Keywords** Computer vision · Deep learning · CCTV · YOLO

## 1 Introduction

Road traffic has become a prominent part in day life of humans which results in or disturbs the services and activities of a human on a daily basis. For smooth transit, especially at urban areas where public communicate customarily various techniques, management has been developed to control the road vehicular traffic. Annually, human fatalities such as deaths, injuries and disabilities and property damages are zooming up with respect to increase in vehicle production and number of vehicle accidents occurred. Despite various measures taken to control road traffic and accidents using CCTV [1] cameras which are placed at the interjections of roads and activities on National /State/District highways so that it can record incidents taking place, lives are lost due to delay in reporting of accidents due to which medication is delayed and lives are lost. Present road traffic managing methods heavily depend on human resources using the captured footage of the CCTV cameras [2]. Using the

---

S. V. Gautham (✉) · D. Hemavathi

School of Computing, SRM Institute of Science and Technology, Chennai, India

e-mail: [gs7083@srmist.edu.in](mailto:gs7083@srmist.edu.in)

human resources, the results are not spontaneous. According to the records, nearly 1.25–1.5 million people lives are lost and around 35–50 million people are either disabled or injured on an annual basis. Road traffic is termed as the ninth leading cause for loss of lives and 2.2% of the casualties worldwide. By 2030, road traffic will be predicted to be the fifth leading cause for human casualties.

At present, vehicle accident detection has become a prominent field using computer vision to overcome this challenging task of providing medication [3] to the injured on time without any human resources for monitoring the accident detection. So this was proposed to overcome the problems that were mentioned by suggesting a solution to detect spontaneously [4] which is very important for the traffic department to evaluate the situation in time. This proposal utilize deep learning framework which detects well-defined objects [5]. Later, we utilize the output of the neural network to detect road vehicular accidents by extracting feature points and our own created parameters that are used for detection of vehicular traffic accidents taking place. We have proposed this framework taking into consideration few of the drastic conditions and parameters that include broad daylight, dark vision, weather conditions, shadow and so on to make sure that our approach is suitable for present accident conditions [6]. Our approach will make sure that the results are spontaneous and accurate in detecting vehicle accidents so that medication can be allocated in time to the injured. Various variables and parameters are taken into consideration in this approach to detect road accidents with respect to the vehicular motion.

## 2 Problem Statement

The main problem that we are facing in the present situation is complication of object detection as not only do we want to classify image objects but also to determine the object position. The accuracy plays a major role in object detection; it is where most of the algorithms stumble with respect to the other datasets used.

Vehicular detection algorithms need not just exactly differentiate and localize important objects; they also need to be superfast at predicting time so as they meet the present demands of video processing. There are several different algorithms present today where most of them are unable to balance w.r.t the accuracy and speed.

There are many applications of object detection, items of interest which may appear in a wide range of sizes and aspect ratios. Practitioners hold several methods to ensure detection parameters or algorithms which are able to capture objects at multiple scales and views.

A very limited dataset is available for detection of vehicle accidents which is a hurdle for the approach. There are different datasets which are object detection dataset and image classification has above 1lakh classes. Further, image classification is easily done, and object detection is done accurately using boundary boxes, still the result are not satisfied.

### 3 Literature Survey

Over the course of time with advancement of technology and techniques in computer vision, various developments and surveys have evolved, many authors have proposed different approaches to detect vehicular traffic accidents. As such of them are.

CK Mohan and D. Singh have tried to detect vehicle accidents with CCTV camera by using spatiotemporal video using demonizing auto encoding (ANN). They have detected moving objects by tracking the objects and could be able to detect in normal traffic flow conditions. But this approach stumbled during heavy traffic [7].

Hui, Yaohua [8] recommended a format which uses Gaussian method to identify a vehicle. Using mean shift algorithm, the vehicles are tracked and handle occlusions during accidents, but they have failed to detect during changes in the traffic flow pattern and severe weather conditions.

Yki and D Lu et al. proposed a vision-based traffic accident detection algorithm; their approach was divided into two ways. First part, it takes the input and uses a grayscale image substance to detect and track the vehicle [9]. Second part applies feature extraction. But it snagged by overlapping their shadows.

Lu has used a hidden Markov model in Calogero Moser system for motion detection and feature extraction and analysis [10]. Hence, the presence of shadow and occlusion which made it lose the target which resulted in giving unsatisfactory results.

Boutheina et al. [11] have used traffic flow technique. Farneback Optical Flow was used for detecting motion of the vehicles and statistics method for detecting vehicle accidents. This was practically applied on few videos that are collected from the CCTV surveillance system. The results were proved to be efficient in which 240 frames were used for motion of the traffic.

Chen [12] extracted OF and scale invariant feature transform (SIFT) features for detection of accidents and then used bag of features (BOF) for encoding. Later, they finally used deep machine learning to detect accidents.

Wang Wei et al. [13] has designed an algorithm that generates alarm and provide the incident area.

Patel K. H. has come up with an Android application which uses sensors in smartphones, but the problem with this system is that any miss tilt of the smartphone which was placed in the vehicle would lead to false alarm which results in generating a false alert to the emergency service [14].

### 4 Proposed Approach

Road accident detection survey approach is done in three steps. They are

- (1) Object detection
- (2) Accident recognition
- (3) Alert generation

**Input:** To detect vehicle accidents, an Open Computer Vision [15] was used. OpenCV has more functions for computer vision and many of its functions are implemented on GPU. OpenCV plays a vital role in the proposed system by capturing the video feed and applied designed model.

Figure 1 describes the flow chart for the detection of accidents. In this, input has been given, after which the input tries for detection of object. If object is detected, then preprocessing is carried out and then model training is done. Later, it detects for any accident; if it is a yes, then alert is generated, and incase the result is no, then the process continues for further detection.

Figure 2 describes different steps that have been carried out in object detection in a one stage detector where input, backbone, neck and dense prediction are the different stages that have been taking place to detect an object using YOLO, only to make sure that the object detected is accurate.

**Object Detection:** There are various existing methodologies for object detection [16]; among them, YOLO has better scope and accuracy.

**YOLO:** YOLO [17] is meant for You Only Look Once. It is a real-object identifying system that can identify multiple objects in a single frame. YOLO can identify objects accurately and with greater speed when compared to other methodologies. It can identify up to 10,000 classes and even some unseen and unidentified classes [2, 18, 19]. In this image recognition system, it identifies multiple objects in a single stretch, and a boundary is made around the object in the form of a box. During this productive system, it gets deployed and easily trained.

YOLO is completely based on an isolated convolutional neural network (CNN). This CNN helps in dividing the image into different objects and predicts the boundary boxes and also predicts multiple boundary boxes for those classes. During training and testing time, complete image is observed by YOLO so that it completely encodes the information about class and appearance.

Figure 3 describes the SPP modified at different stages finally resulting in YOLOv4 where different stages like input image, convolution down sampling block, dense connection block, spatial pyramid pooling block and object detection block have been processed.

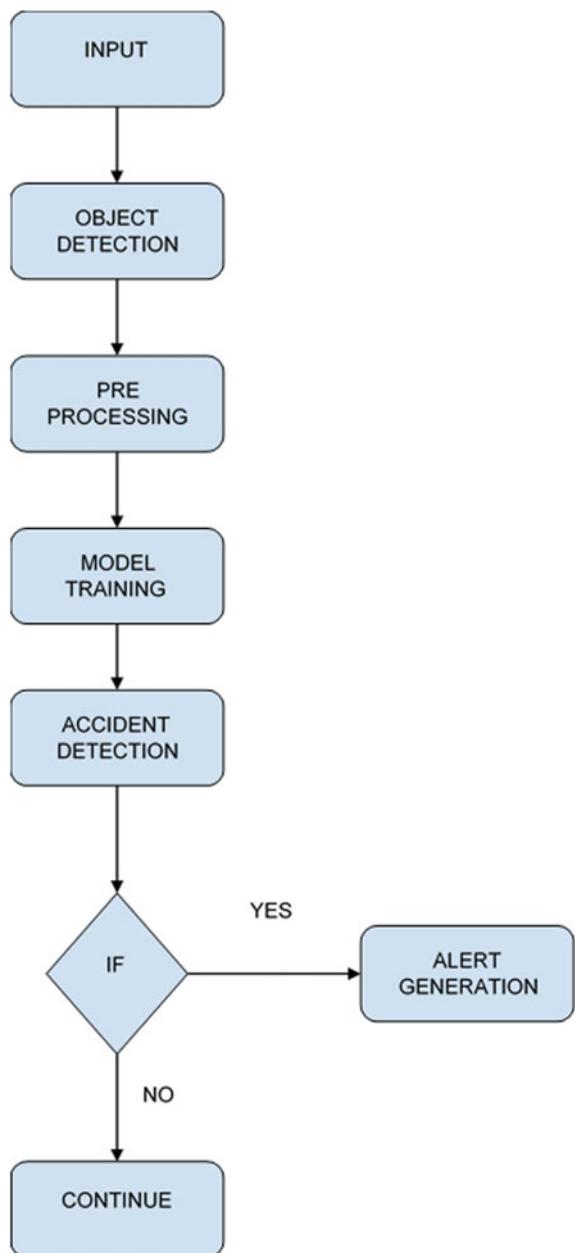
### Detection Framerate compared to other frameworks:

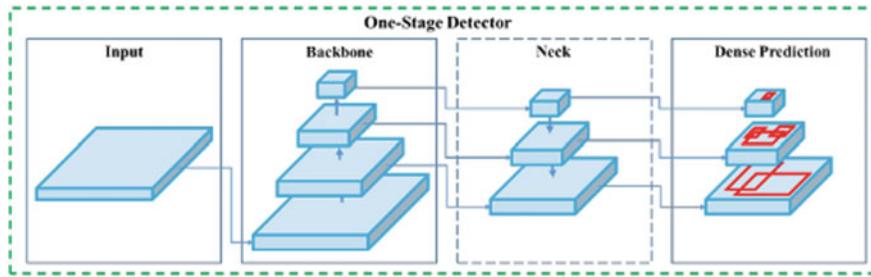
Table 1 describes the various methods which was implied for PASCAL VOC-2007 and VOC-2012. PASCAL VOC is a set of data for object detection which has a huge repository for various images existed, be it a small object or large object with different image resolutions [20].

PASCAL VOC has conducted a challenge in 2007 and 2012 in which various methodologies were tested where their respective score was graded with regard to their mean average precision.

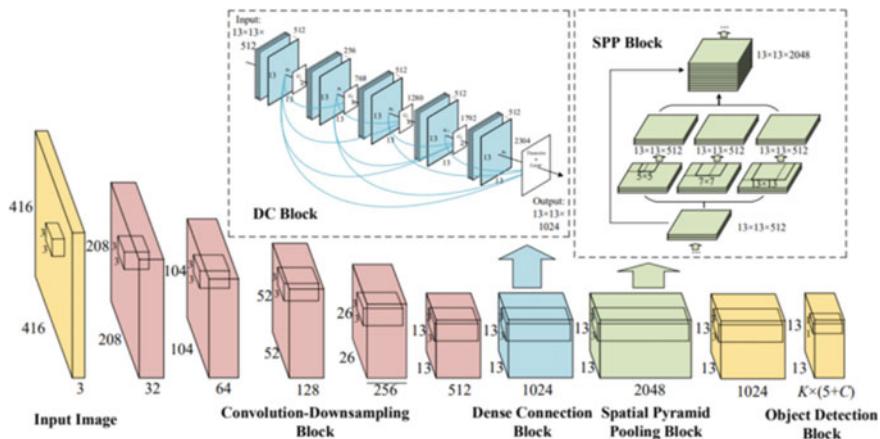
From Table 1, it is shown that various methodologies like fast R-CNN, fast R-CNN with VGG-16, faster R-CNN with ResNet, YOLO, SSD300(Single Shot Detector), SSD500, YOLOv2 with  $288 \times 288$  resolution, YOLOv2  $352 \times 352$ , YOLOv2  $416 \times 416$  and YOLOv2  $544 \times 544$  were listed above.

**Fig. 1** Workflow diagram describing the process of accident detection





**Fig. 2** YOLOv4 object detector



**Fig. 3** The diagram above demonstrates how modified SPP is integrated into YOLOv4

**Table 1** Shows the test result of PASCAL VOC-2007 and VOC-2012

Detection frameworks	Train	mAP	FPS
Fast R-CNN	2007 + 2012	70.0	0.5
Faster R-CNN VGG-16	2007 + 2012	73.2	7
Faster R-CNN ResNet	2007 + 2012	76.4	5
YOLO	2007 + 2012	63.4	45
SSD300	2007 + 2012	74.3	46
SSD500	2007 + 2012	76.8	19
YOLOV2 288 × 288	2007 + 2012	69.0	91
YOLOV2 352 × 352	2007 + 2012	73.7	81
YOLOV2 416 × 416	2007 + 2012	76.8	67
YOLOV2 480 × 480	2007 + 2012	77.8	59
YOLOV2 544 × 544	2007 + 2012	78.6	40

Training is performed on PASCAL VOC 2007 and 2012 dataset by the all methodologies which were listed above in which its respective mean average precision score was graded and average FPS is also mentioned.

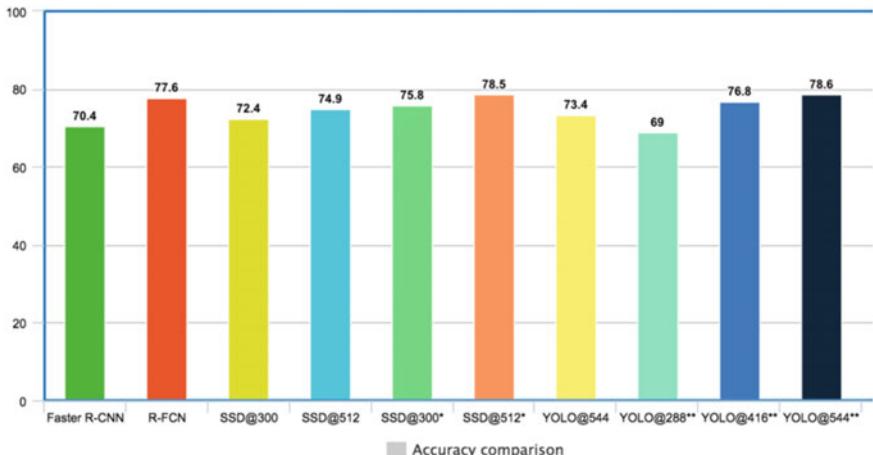
As Table 1 represents the accuracy score of YOLO is pretty high well compared with other methodologies and the average FPS is also more on YOLO algorithm.

Thus, it implies that YOLO algorithm works in detection of the object as its respective scores are pretty high.

Table 2 shows the detailed explanation of the various methodologies that are tested on different object, and their accuracy scores were listed above such that it would be easy to know that at how much accurate and detection method would predict an object and its accuracy will be listed out.

**Table 2** Shows the test result of PASCAL VOC-2007 and VOC-2012 on various objects

Methods	Fast R-CNN	Faster R-CNN	YOLO	SSD300	SSD512	ResNet	YOLOv2 544
Data	2007 + 2012	2007 + 2012	2007 + 2012	2007 + 2012	2007 + 2012	2007 + 2012	2007 + 2012
mAP	68.4	70.4	57.9	72.4	74.9	73.8	73.4
Aero Plane	82.3	84.9	77.0	85.6	87.4	86.5	86.3
Bike	78.4	79.8	67.2	80.1	82.3	81.6	82.0
Bird	70.8	74.3	57.7	70.5	75.8	77.2	74.8
Boat	52.3	53.9	38.3	57.6	59.0	58.0	59.2
Bottle	38.7	49.8	22.7	46.2	52.6	51.0	51.8
Bus	77.8	77.5	68.3	79.4	81.7	78.6	79.8
Car	71.6	75.9	55.9	76.1	81.5	76.6	76.5
Cat	89.3	88.5	81.4	89.2	90.0	93.2	90.6
Chair	44.2	45.6	36.2	53.0	55.4	48.6	52.1
Cow	73.0	77.1	60.8	77.0	79.0	80.4	78.2
Table	55.0	55.3	48.5	60.8	59.8	59.0	58.5
Dog	87.5	86.9	77.2	87.0	88.4	92.1	89.3
Horse	80.5	81.7	72.3	83.1	84.3	85.3	82.5
mBike	80.8	80.9	71.3	82.3	84.7	84.8	83.4
Person	72.0	79.6	63.5	79.4	83.3	80.7	81.3
Plant	35.1	40.1	28.9	45.9	50.2	48.1	49.1
Sheep	68.3	72.6	52.2	75.9	78.0	77.3	77.2
Sofa	65.7	60.9	54.8	69.5	66.3	66.5	62.4
Train	80.4	81.2	73.9	81.9	86.3	84.7	83.8
TV	64.2	61.5	50.8	67.5	72.0	65.6	68.7



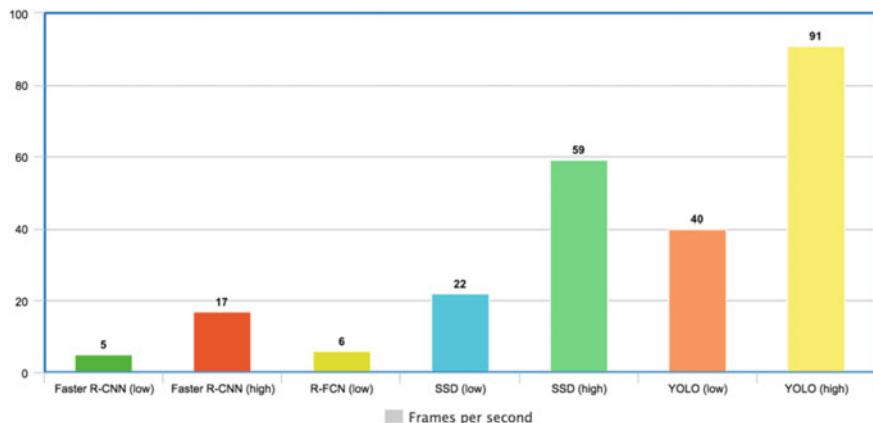
**Fig. 4** The accuracy comparison among various methodologies

Figure show the graph (Fig. 4) of different methodologies that have been used for detection of accidents and their accuracy of object detection, among which YOLO has better accuracy than any other method.

\* denotes small object data augmentation is applied.

\*\* indicates as VOC2007 [12] test which has better results when compared to VOC 2012 [13], so VOC 2007 is taken as reference for YOLO

Inputs are speed and resolution of the object that is detected. Below attached graph (Fig. 5) indicates higher and lower frames per second (FPS) results of different test



**Fig. 5** Shows accuracy score (mAP) by various methodologies on FPS

papers. At different mAP [21], they are measured which are highly biased with respect to the below result.

When coming to this Fig. 5, it shows the clear result of FPS accuracy of different methodologies used for object detection. Even this has YOLO to be the most accurate among all other methodologies.

**Accident Detection:** There is no specific dataset available for the road accidents and have prepared a dataset with images containing the accident.

The model score is generated by applying the formula on a real-time video feed to check for whether an accident occurred or not; if the score reaches a given threshold, accident is detected and an alert will be generated or continued with the video.

$$\text{mAP} = \frac{1}{|\text{classes}|} \sum_{c \in \text{classes}} \frac{\#\text{TP}(c)}{\#\text{TP}(c) + \#\text{FP}(c)}$$

where

mAP Mean Average Precision

TP True Positive

FP False Positive

Well, getting into the details, the mean average precision has metrics in order to judge the performance of a classification model that is precision and recall.

In order to know the operations of average precision for object detection, it is important to know about the intersection over union which means the intersection area and union of area of the expected boundary boxes around the image and ground of the boundary boxes are calculated.

IOU means Intersection over Union using that mathematical calculation an accident detection model is built.

**Alert Generation:** Generation of the alert [22] will be done when threshold score is above the limit; if not, alert will not be generated.

## 5 Conclusion

In this study for accident detection of the vehicle was proposed, using the regression-based algorithm known as YOLO (You Only Look Once) which is tested using the livestream video data from the CCTV. The proposed system is accurate and faster than other existing object detection methods and predicts the object better than other object detection algorithms. From the provided input, system tries to detect the accident and generate an alert to the nearby respective emergency services.

## References

1. A. Franklin, The future of CCTV in road monitoring, in *Proceeding of IEE Seminar on CCTV and Road Surveillance*, pp. 10/1–10/4 (1999). <https://ieeexplore.ieee.org/document/793939>
2. Object detection: speed and accuracy comparison (faster r-cnn, r-fcn, ssd, fpn, retinanet and yolov3)
3. Y. Ki, J. Choi, H. Joun, G. Ahn, K. Cho, Real-time estimation of travel speed using urban traffic information system and CCTV, in *Proceeding of International Conference on Systems, Signals and Image Processing (IWSSIP)*, pp. 1–5 (2017). <https://ieeexplore.ieee.org/document/7965582>
4. An accident detection system on highway through CCTV with calogero-moser system. <https://ieeexplore.ieee.org/document/6388248>
5. .R.J. Blissett, C. Stennett, R.M. Day, Digital CCTV processing in traffic management, in *Proceeding of IEE Colloquium on Electronics in Managing the Demand for Road Capacity*, pp. 12/1–12/5 (1993). [https://ieeexplore.ieee.org/document/280164?denied=\\_](https://ieeexplore.ieee.org/document/280164?denied=_)
6. R. Gavrilescu, C. Zet, C. Foșalău, M. Skoczyłas, D. Cotovanu, Faster R-CNN: an approach to real-time object detection
7. D. Singh, C.K. Mohan, Deep spatio-temporal representation for detection of road accidents using stacked autoencoder. *IEEE Trans. Intell. Trans. Syst.* **20**(3), 879–887 (2019). <https://ieeexplore.ieee.org/abstract/document/8367975>
8. Z. Hui, X. Yaohua, M. Lu, F. Jiansheng, Vision-based real-time traffic accident detection, in *Proceeding of World Congress on Intelligent Control and Automation*, pp. 1035–1038 (2014)
9. Y. Ki, D. Lee, A traffic accident recording and reporting model at intersections. *IEEE Trans. Intell. Transp. Syst.* **8**(2), 188–194 (2007)
10. W. Hu, X. Xiao, D. Xie, T. Tan, S. Maybank, Traffic accident prediction using 3-d model-based vehicle tracking. *IEEE Trans. Vehicular Technol.* **53**(6), 677–694 (2004). <https://ieeexplore.ieee.org/document/1300862>
11. Adaptive video-based algorithm for accident detection on highways. <https://ieeexplore.ieee.org/document/7993382>
12. PascalVoC 2007 visual object class classification. [http://host.robots.ox.ac.uk/pascal/VOC/voc\\_2007/](http://host.robots.ox.ac.uk/pascal/VOC/voc_2007/)
13. PascalVoC 2012 visual object class classification. [http://host.robots.ox.ac.uk/pascal/VOC/voc\\_2012/](http://host.robots.ox.ac.uk/pascal/VOC/voc_2012/)
14. T. Lin, M. Maire, S.J. Belongie, L.D. Bourdev, R.B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollar, C.L. Zitnick, Microsoft COCO: common objects in context, in *CoRR*, vol. abs/1405.0312, 2014. [Online]. Available <http://arxiv.org/abs/1405.0312>
15. OpenCV. <https://docs.opencv.org/master/>
16. Object detection for dummies part 3: R-CNN family. <https://lilianweng.github.io/lillog/assets/images/rcnn-family-summary.png>
17. A. Bochkovskiy, C.-Y. Wang, H.-Y.M. Liao, YOLOv4: optimal speed and accuracy of object detection. <https://arxiv.org/abs/2004.10934>
18. A. Daniel, K. Subburathinam, B.A. Muthu, N. Rajkumar, S. Kadry, R.K. Mahendran, S. Pandian, Procuring cooperative intelligence in autonomous vehicles for object detection through data fusion approach. *IET Intel. Trans. Syst.* **14**(11), 1410–1417 (2020). <https://doi.org/10.1049/iet-its.2019.0784>
19. N.-T. Nguyen, M.C. Leu, S. Zeadally, B.-H. Liu, S.I. Chu, Optimal solution for data collision avoidance in radio frequency identification networks. *Internet Technol. Lett.* **1**, e49 (2018). <https://doi.org/10.1002/itl2.49>
20. K.H. Patel, Utilizing the emergence of android smartphones for public welfare by providing advance accident detection and remedy by 108 ambulances, *Int. J. Eng. Res. Technol.* (2013). ESRSA Publications
21. S.M. Beitzel, E.C. Jensen, Ophir Frieder, mAP. [https://doi.org/10.1007/978-0-387-39940-9\\_492](https://doi.org/10.1007/978-0-387-39940-9_492)

22. S. Taylor, J. Janelle, Automatic alert code and test generation system. <https://ieeexplore.ieee.org/document/111270>
23. R. Girshick, Fast RCNN. <https://arxiv.org/abs/1504.08083>
24. J. Ren, Y. Chen, L. Xin, J. Shi, B. Li, Y. Liu, Detecting and positioning of traffic incidents via video-based analysis of traffic states in a road segment. IET Intel. Transport Syst. **10**(6), 428–437 (2016)
25. W. Wei, F. Hanbo, Traffic accident automatic detection and remote alarm device, in 2011 *International Conference on Electric Information and Control Engineering (ICEICE)*. (IEEE, 2011)

# An Ensemble Learning Method on Mammogram Data for Breast Cancer Prediction—Comparative Study



T. Sreehari and S. Sindhu

**Abstract** Cancer remains a serious problem for people nowadays because of its largest death rate in the world. The biggest rate is full of the absence of adequate early detection of cancer. Breast cancer carcinoma is one of the main killer illness among ladies in the world, and the commonly diagnosed non-skin cancer in women. Early detection of this deadly disease can considerably increase the possibility of successful treatment. The traditional way of cancer diagnosis primarily depends on the doctor's or technician's experience to identify the abnormalities in our human body. Bosom malignancy happens once the cell tissues of the breast become irregular and wildly divided. By investigating, the mammogram information can help doctors in recognizing disease cases at a whole lot sooner stage and accordingly can altogether improve the capability of patient treatment. It is as yet a test to anticipate the forecast of disease patients, due to its high heterogeneity and multifaceted nature. This study focuses on the forecast of breast cancer from mammogram pictures and assesses the precision of various CNN models.

**Keywords** Cancer prediction · Mammogram data · CNN

## 1 Introduction

Malignancy has been portrayed as a gathering of related infections including anomalous cell development with the possibility to partition ceaselessly and spread into enveloping tissues. According to the GLOBOCAN project, in the year 2012 alone, roughly 14.1 million new instances of malignant growth happened universally (excluding skin disease other than melanoma), which caused around 14.6% of the demise. Since disease could be a significant purpose behind dismalness and mortality,

---

T. Sreehari (✉) · S. Sindhu

Department of Information Technology, SRM University, Chennai, India

e-mail: [st7170@srmist.edu.in](mailto:st7170@srmist.edu.in)

S. Sindhu

e-mail: [sindhus2@srmist.edu.in](mailto:sindhus2@srmist.edu.in)

diagnosing and identifying malignant growth in its early stage is critical for its fix. Over the previous many years, the interminable development of malignancy investigation has been performed. Among the different strategies and methods created for malignant growth forecast, the use of mammogram information is one of the research hotspots in this field. Data analysis of mammogram pictures has encouraged malignancy conclusion and treatment by and large. Right forecast of disease is one of the most significant and dire assignments for the specialist.

With the quick advancement of computer-aided techniques recently, the use of machine learning—ML and deep learning—DL techniques is assuming an inexorably significant function in malignancy analysis, and different prediction algorithms are being investigated constantly by researchers.

Breast disease is the second most general sickness among females, and it is the number one driving reason for malignancy deaths around the planet. The seriousness of bosom malignant growth can be split into five phases (0-IV), which depicts the degree of bosom disease inside the patient's body. The intricacy of mammary gland disease and its considerably different clinical impacts currently make it trying to foresee and treat. Along these lines, having the option to anticipate malignancy infection at its beginning phase with all the more precisely not just helps breast cancer patients to think about their future but additionally helps doctors to settle on educated choices inside a limited ability to focus time and further guide clinical treatment.

On the off chance that you have been determined to have a tumor, the essential advance your primary care physician can take is to search out whether it is dangerous or normal, as this can influence your treatment course of action. So, the significance of the term harmful is malignant, and the importance of benign is noncancerous. So, the terms malignant(harmful) and benign (noncancerous) are significant to familiarize the analysis of disease and hence influence your well-being. Previously, doctors do this using manually seeing mammogram photo but now with help of previous patient data, we can create a machine learning model or deep learning model to learn from it and make a future prediction whether it is malignant or benign, the patient is having cancer or not.

To decrease metastatic risks, most breast cancer patients receive systemic adjuvant therapy after surgery. However, this therapy can also have serious facet outcomes on health, and the majority of the sufferers will be disease-free even barring it. It is challenging to decide whether the affected person desires the therapy or no longer only depending on the clinical performance.

The computer-aided detect and diagnosis (CAD) frameworks are as of now being utilized in the clinical field to offer pivotal help with the dynamic cycle of radiologists. Such frameworks may likewise broadly lessen the measure of exertion needed for the appraisal of a sore in clinical practice while limiting the scope of false positives it leads to pointless and discomforting biopsies. The computer-aided detect and diagnosis also known as CAD frameworks with respect to mammography may address two distinct assignments: to find the location of dubious injuries in a mammogram image (CADe) and conclusion of recognized sores (CADx), i.e., grouping as normal and harmful.

Deep learning (DL) is considered as a major advancement innovation of proceeding years since it has shown execution a long way past the cutting edge in various machine learning (ML) tasks, for example, object discovery and grouping of objects. In spite of standard machine learning (ML) methodologies, which require a hand-created features extraction stage, which is a troublesome undertaking since it relies upon domain information, deep learning (DL) strategies adaptively gain proficiency with the reasonable component extraction strategy from the input information to the objective yield. This takes out the dull strategies for designing and exploring the separation capacity of the choices while encouraging the dependability of the techniques.

## 2 Related Works

### 2.1 Literature Study

The paper [1] audits the procedure of convolutional neural network also known as CNN applied in a particular field of mammographic breast cancer diagnosis or MBCD. It plans to give data about how to utilize a CNN for related undertakings. MBCD is a long-standing issue, and enormous computer-aided diagnosing and detection (CAD) models have been proposed. With help of CNN, a mammographic breast cancer diagnosis can be comprehensively sorted into three gatherings. The first one is to configuration shallow or to alter existing models to diminish the time cost just as the number of occasions for preparing; second one is to make use of a pretrained convolutional neural network (CNN) model by transfer learning strategy to fine-tune the model; the third and last one is to exploit convolutional neural network (CNN) models for the feature extraction, and the distinction of benign ones and harmful injuries is satisfied by utilizing the ML classifiers. This investigation enlists peer-inspected paper distributions and presents specialized subtleties and the upsides and downsides of each model. Besides, the paper incorporates the discoveries, difficulties, and impediments are summed up, and a few signs on the future work are likewise given.

The paper [2] contains the present day and age computerized picture preparation which is vital in various regions of innovative work. Advanced image processing is utilized to handle computerized pictures and create valuable attributes from the information, which would then be able to be utilized to settle on basic choices with high exactness. These strategies are likewise utilized in the clinical field, for instance, in the recognition of mammary gland cancer. In the present-day situation, breast malignancy is one of the primary drivers of death among ladies on the planet, and it is hard to forestall bosom disease as the fundamental reasons basic bosom disease stay obscure. All things considered, the explicit trademark signature of bosom disease, which encompasses microcalcifications and masses that can be found in mammograms, can be utilized for early discovery, and henceforth are massively truly supportive

for ladies who may also be an opportunity of developing dangerous tumors. The chief check utilized for screening and early finding is mammography, and the best possible investigation of the clinical report is imperative for bosom malignant growth expectation; however, the choice might be inclined to error.

The paper [3] is with respect to CAD which is generally utilized in mammography. The standard CAD utilization prompts to point out the expected malignant growths on the mammograms with an improvement in analytic precision. Due to the advances in the ML, particularly with the utilization of deep (multilayered) convolutional neural networks (DCNNs), AI has gone through a change that has a superior nature to the forecasts of the models. These days, such DL algorithms have been applied to mammography and digital breast tomosynthesis (DBT). Here in this paper, they clarified how a DL model functions inside the setting of mammography and digital breast tomosynthesis and characterized some significant difficulties in it. A short time later, they talk about this standing and future perspectives on AI-based medical applications for mammography, DBT, and radionics. As of now, accessible deep learning algorithms are progressed and approach the exhibition of radiologists—particularly for malignancy recognition and danger expectation at images of mammography. Notwithstanding, clinical approval is basically missing, and it is not satisfactory how the ability of dl should be utilized to advance practice. Further improvement of dl models is important for digital breast tomosynthesis and to play out this requires the assortment of bigger databases. It is normal that DL can in the end have a crucial part in digital breast tomosynthesis, just as in the age of manufactured pictures.

The mammary gland malignancy is the most widely recognized disease among ladies around the world [4]. Notwithstanding colossal clinical advancement, bosom malignant growth has remained the subsequent driving reason for death around the world; consequently, its initial conclusion significantly affects diminishing mortality. In any case, it is hard to analyze bosom irregularities. There are various devices, for example, mammography, ultrasound, and thermography, which have been created to screen mammary gland malignant growth. Thusly, the computer assists radiologist technicians to distinguish chest irregularities all the more proficiently by utilizing artificial intelligence (AI) and picture processing instruments. Here in this review paper, they analyzed different techniques for artificial intelligence (AI) utilizing picture handling to recognize bosom malignant growth. They directed an examination through a library and Internet look. Via looking through various databases and research papers alongside looking for bosom malignant growth catchphrases, AI and clinical picture handling methods were extracted. The acquired results from the investigation were arranged to show various methods and their outcomes over ongoing years. From the examination, the outcomes indicated that SVM had the most elevated exactness rate for various kinds of pictures, for example, ultrasound, mammography, and thermography.

Mehdy et al. [5] Here in this paper, the medical imaging procedures have broadly been being used in the area of finding and discovery of bosom malignancy. The disadvantage of applying these methods is enormous time utilization within the nonautomatic detection of each picture design by an expert radiologist. Computerized classifiers may significantly overhaul the discovery technique, as far as both

precision and time prerequisite by recognizing benign and malignant examples consequently. Neural network (NN) assumes an indispensable part in this, particularly in the field of bosom malignancy identification. Regardless of the enormous number of examination paper distributions that depict the utilization of neural networks (NNs) in different clinical procedures, just a few review papers are accessible that direct the advancement of those algorithms to support the detection strategies concerning particularity and affectability. This survey paper intends to dissect the substance of as of late distributed writing with special attention to techniques and state of art of neural networks (NNs) in clinical imaging. They talk about the utilization of neural networks (NNs) in four totally extraordinary clinical imaging applications to bring up that NN is not confined to a couple of territories of medication. Different types of NN utilized, alongside the different kinds of feeding information, have been audited. They likewise address mixture neural networks (NNs) transformation in mamma gland malignancy identification.

This paper [6] meant to study both customary machine learning (ML) and deep learning (DL) writing with specific applications for bosom disease determination. The survey additionally gave a concise knowledge into some notable deep learning (DL) networks present a review of machine learning (ML) and deep learning (DL) methods with specific applications for bosom malignancy. In particular, they search in the PubMed, Google Scholar, MEDLINE, ScienceDirect, Springer, and Web of Science information bases and recover the examinations in deep learning (DL) for as far back as five years that have utilized multi-view mammogram datasets.

Breast malignancy is among the world's second most happening diseases in a wide range of malignancy. Charan et al. [7] They explain that in these days, the most widely recognized disease among ladies overall is bosom malignant growth. There is consistently a requirement for progression with regards to clinical imaging. In the event that we identify malignancy in its early phase and give proper treatment, the danger to the life of the patient can be decreased substantially. Machine learning (ML) can assist clinical experts in recognizing the infection with more precision and less time interval, whereas deep learning (DL) or neural network (NN) is one of the advanced procedures which can be utilized for the arrangement of benign and malignant bosom location. Convolutional neural networks (CNNs) can be utilized for bosom malignant growth location. Here, they utilized the mammograms-MIAS dataset for this project. This dataset contains three hundred and twenty-two (322) mammogram pictures, just about one hundred and eighty-nine (189) pictures are benign and one hundred and thirty-three (133) are malignant bosoms. Empowering exploratory discoveries have been acquired, which speaks to the viability of depp learning (DL) models for bosom malignancy discovery utilizing mammogram pictures and further encourages the utilization of deep learning (DL)-based most recent component extraction and grouping techniques in different clinical imaging applications, particularly in the zone of bosom disease detection utilizing mammogram images. It is as yet progressing research and further advancements are being made by enhancing the CNN design and furthermore utilizing different pretrained network models which will ideally prompt higher exactness measures. Appropriate segmentation methods are required for effective feature extraction and classification.

In [8], it is imperative to distinguish bosom malignant growth as ahead of schedule as could be expected under the circumstances. Here, another system for characterizing mammary gland disease utilizing deep learning (DL) and some division strategies are presented. Another computer-aided detection (CAD) framework is presented for characterizing benign and harmful mass tumors in bosom mammography pictures [9]. Here, computer-aided detection (CAD) framework and two division approaches are utilized. The first one includes deciding the region of interest (ROI) physically, and the subsequent one uses the method of limit and area based. Here, a deep convolutional neural network also known as DCNN is utilized for the element extraction measure. A notable deep convolutional neural network (DCNN) design named AlexNet is utilized and is adjusted to classify two classes rather than 1,000 classes. In that model, the last fully connected layer is connected to the support vector machine (SVM) classifier to acquire precise exactness. For this paper, they utilized two freely accessible datasets; the first one is DDSM, and the remaining one is CBIS-DDSM. Training on numerous data gives a high precision rate [10]. However, the clinical datasets contain a nearly modest number of tests because of fewer patient's information. Appropriately, data enlargement is a method for expanding the dimension of the input data by methods for creating new data from the first info inputted data. There are numerous structures for data expansion; the one utilized here is rotation. The exactness of the new-prepared deep convolutional neural network (DCNN) engineering is higher than past works utilizing a similar condition.

Mammary gland disease is one of the huge explanations behind death among women [11]. Numerous kinds of studies have been done on the finding and discovery of bosom disease utilizing different picture processing and grouping strategies. In any case, the infection stays probably the deadliest sickness, having conceived one out of six ladies in the course of her life. Since the reason for breast malignancy remains obscure, avoidance gets inconceivable. Thus, the early discovery of tumors in the mammary gland is the best way to fix the bosom disease. Utilizing computer-aided diagnosis (CAD) on the mammography picture is the most proficient and least demanding approach to analyze bosom malignancy. Precise revelation can adequately lessen the passing pace of breast malignancy. Masses and microcalcifications clusters are significant early side effects of conceivable bosom malignant growths. They can help detect bosom disease in its infant state. The digital database for screening mammography (DDSM) is taken for this project, which contains around 3000 cases and is being utilized worldwide for malignancy research. The paper collectively speaks about the testing framework used for highlighting the affected area for the recognition of breast cancer growth. These surface highlights are separated from the ROI of the mammogram picture to portray the microcalcifications as innocuous, normal, or undermining. From that point onward, by utilizing principle component analysis (PCA), we can diminish the highlights for the better capturing of Masses. After that, features are additionally compared and proceed through the back-propagation algorithm (neural network) for a superior comprehension of the malignancy design in the mammography picture.

The restrictions of convolutional computer-aided detection and diagnosis (CAD) frameworks for mammography explained in [12], the most extreme significance of

the early detection of mammary gland malignancy, and the colossal effect of the bogus discovery on an affected individual drives scientist to explore deep learning (DL) techniques for mammograms (MG). The recent breakthroughs in deep learning (DL) techniques, and advancements in convolutional neural networks(CNNs) has helped AI to make striking advances in the clinical fields. In detail, convolutional neural networks (CNNs) are utilized in mammography for sore confinement and identification, hazard assessment, recovery of pictures, and characterization undertakings. Convolutional neural networks (CNNs) additionally help radiologists and cancer specialists by providing a more exact conclusion by conveying the exact quantitative examination of dubious sores. In this review paper, they run an exhaustive report on the qualities, impediments, and execution of the latest convolutional neural networks (CNNs) applications in the field of dissecting mammogram pictures. From the entire study, they reached the conclusion that around 83 examination reads for applying convolutional neural networks (CNNs) on different assignments in mammography. It centers around finding the prescribed procedures utilized in these examination studies to improve the recognition of malignancy exactness. This review likewise gives a profound knowledge into the engineering of convolutional neural networks (CNNs) utilized for different errands. Likewise, it clarifies the most well-known openly accessible mammogram image databases and features their principle highlights and qualities.

The rapid advancement of deep learning (DL) and a group of machine learning (ML) procedures have prodded a ton of curiosity in its application to clinical imaging issues [13]. Here, they build up a deep learning (DL) model which will precisely analyze bosom malignancy on mammogram screening by utilizing an “end-to-end” training approach that productively uses training datasets with either entire clinical comment or just the disease status (label) of the entire mammogram picture. While performing this study project, injury comments are required only at the beginning stage of the training purpose; after that, the leftover stages require just picture level marks, deleting the dependence on seldom accessible sore explanations. The entire convolutional neural network (CNN) technique for ordering screening mammograms accomplished magnificent execution in contrast with past strategies; by utilizing an autonomous test dataset of digitized film mammogram pictures, they utilized the pictures from the database called digital database for screening mammography (CBIS-DDSM).

In [14], a method is given to classify the breast cancer masses in keeping with the new geometric features. After getting the computerized breast mammogram pictures from the digital database for screening mammography (DDSM), picture preprocessing was performed. Then, by using image processing strategies, an algorithm was developed for automatic extracting of masses from other normal parts of the breast mammogram image. In this study paper, nineteen final completely different features of every image were extracted to generate the feature vector for classifier input. The proposed system not only determined the boundary of masses but also classified the type of masses into benign and malignant ones. The neural network (NN) classification strategies, for example, k-nearest neighbors (KNN), probabilistic

neural network (PNN), and the radial basis function (RBF), and so on are utilized for an official conclusion of the mass class.

In this paper, [15] breast malignancy is a significant underlying driver of death in most malignant growth influenced ladies, and consistently, a large number of ladies are biting the dust far and wide in light of this destructive infection. In these days, different image processing strategies are utilized to extract the highlights from mammogram pictures. Mammogram images are similar to X-ray; mammograms are grayscale pictures of the mammary gland which are utilized to envision the early indication of bosom cancer. Cancer Specialists are utilizing different picture handling methods for compelling choices making, by looking at the mammographic pictures. Microcalcifications are little calcium stores that resemble white bits on mammograms. These are generally not a consequence of malignant growth but rather in the event that they show up in a specific example and bunch together, can prompt the beginning phase of the disease. Different image processing techniques can be utilized for recognizing these microcalcifications. This paper aims to distinguish the dangerous cells by the utilization of various existing image processors with their examinations of yield execution, and the upsides of this technique are future expectations of breast cancer. On the off chance that the malignancy cells are identified at the beginning phase, numerous valuable lives can be spared by giving early treatment.

Here [16], they explain nowadays breast cancer disease is probably the biggest reason for ladies' passing on the planet. The serious designing strategy of characteristic image classification and artificial intelligence (AI) techniques has widely utilized for the mammary gland mammogram picture characterization task. Utilizing the computerized picture grouping will support the doctors, and the radiologist a subsequent sentiment, and it will spare the technicians' and doctors' time. Notwithstanding a few study papers on mammary gland picture characterization, a couple of survey papers are accessible which gives definite data on bosom malignancy picture arrangement procedures, including feature extraction, and feature selection process, and picture grouping discoveries. Here, they put an extraordinary accentuation on the convolutional neural network (CNN) technique for bosom picture order. Alongside the convolutional neural network (CNN) technique, we have likewise depicted the contribution of the customary neural network (NN), and logic-based classifiers, for example, support vector machines (SVM), random forest (RF) algorithms, and a couple of the supervised and semi-supervised strategies which have been utilized for mamma image classification.

In this paper [17] breast malignancy is the most well-known illness in ladies and represents an incredible danger to ladies' life and well-being. Mammography is one of the viable strategies for distinguishing bosom malignancy; however, the outcomes are generally restricted by the clinical experience of radiology staffs or cancer specialists. Subsequently, the primary motivation behind this exploration paper is to play out a two-stage characterization of normal/abnormal and benign/malignant of two-view mammograms through a CNN [18]. Here in this investigation paper, they developed a multi-view extraction of feature network model for the characterization of mammogram pictures from two perspectives, and they suggested a multi-scale consideration DenseNet as the spine network for extraction of features [19]. In this paper, the

proposed model contains two autonomous branches; these are utilized to extract the features of two mammogram images from various perspectives. Their works are chiefly focused on the development of a multi-scale convolution module and consideration module. The outcomes show that the model has acquired great execution in both classification tasks. Here, they utilized the digital database for screening mammography (DDSM) database to assess the proposed technique.

## ***2.2 Dataset Description***

In this project, we used datasets from the digital database for screening mammography (DDSM). It contains digitally stored mammogram images. These images are used by researchers and scientists for their research purpose. The Sandia National Laboratories, Massachusetts General Hospital, and the University of South Florida Computer Science and Engineering Department are the main contributors of this database. It contains roughly 2500 examinations. Each examination comprises of two photographs of each bosom, alongside some connected affected individual's information data such as age at the time of the investigation, breast density rating, etc., and mammogram picture records, for example, details of the scanner, information about spatial resolution, and so on.

## ***2.3 Problem Statement***

Mammary gland disease happens once the cell tissues of the breast become irregular and uncontrollably partitioned. By examining the mammogram image, data can help doctors in identifying disease cases at a whole lot sooner stage and in this way can essentially improve the capability of patient treatment. It is still a challenge to predict the prognosis of cancer patients, because of its high heterogeneity and complexity.

Accurate cancer prognosis in patients is crucial for timely treatment. Prediction models have been built over the years that contributes a valued proposition to assess the risk and forecasting it by identifying individuals at high risk, aiding the design and planning of clinical trials, and enabling better living.

By using mammogram images, we can detect cancer at its beginning stage. In the past decades, doctors do this using by manually seeing mammogram photos but now with help of previous patient data, we can create a machine learning (ML) model to learn from it and make a future prediction whether its malignant or benign patient is having cancer or not. Early detection of cancer can be cured easily.

This project aims at the prediction of breast cancer from mammogram images and compares the accuracy of various CNN models.

## 2.4 Relative Study Table

Author	Application	Dataset	Methods	Accuracy
Yan [20]	To recognize body parts	Dataset contains CT image slices of 12 body organs	Two stage convolutional neural network	~ 92.23%
Tulder [21]	To classify lung texture and airway detection	ILD (interstitial lung diseases) CT scans	Convolutional restricted boltzman machine	~ 89%
Anthimopoulos [10]	Lung pattern classification	ILD (interstitial lung diseases) CT scans	Convolutional neural network	85.5%
Sirinukunwattana [21]	Detection and classification of nuceli	histology images of colorectal adenocarcinomas	Two architectures of CNN	~ 80.2%
Payan [22]	Prediction of alzheimer disease	MRI Images	CNN with 2D convolutions and 3D convolutions	~ 89.4% with 3D convolutions 85.53% with 2D convolutions
Hosseini [23]	Alzheimer disease diagnosis	MRI Images	DSA- 3D CNN	94.8%
Farooq [24]	Classification of alzheimer disease	MRI Images	GoogleNet ResNet-18 ResNet-152	98.88% 98.01% 98.14%
Ma [25]	Thyroid nodule diagnosis	Ultrasound Images	Pretrained convolutional neural network	~ 83%
Sun [26]	Breast cancer diagnosis	Mammographic Images with ROIs	CNN using semi supervised learning	~ 82.43%
Pratt [27]	For diabetic retinopathy	Kaggle dataset	Convolutional neural network	75%
Farahnaz Sadoughi[4]	Diagnosis of breast cancer	Various medical images	Artificial intelligence	Ultrasound (95.85%), Mammography (93.069%) Thermography (100%)
Charan et al.[7]	Detection of breast cancer	Mammograms-MIAS dataset	Convolutional neural network	65%
Ragab [8]	Breast cancer detection	Digital database for screening mammography (DDSM)	CNN with support vector machines (SVM)	DCNN = 71.01% SVM = 87.2%

(continued)

(continued)

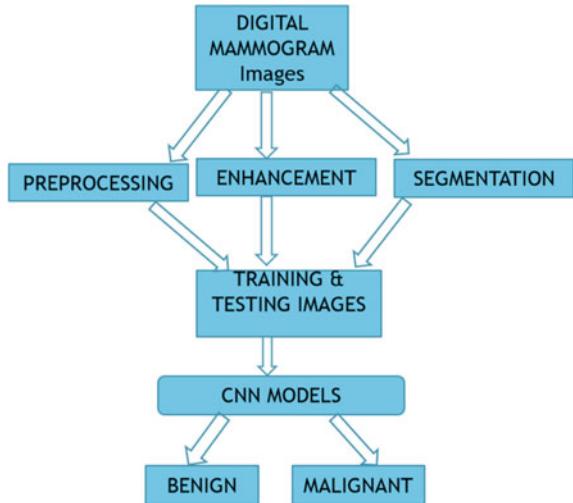
Author	Application	Dataset	Methods	Accuracy
Shen [13]	Breast cancer detection	Digital database for screening mammography (CBIS-DDSM)	Deep learning using an “end-to-end” training approach	~ 90%
Safdarian [14]	To detect and classification of breast cancer	Digital database for screening mammography (DDSM)	Radial basis function (RBF), Probabilistic neural network (PNN), and multi-layer perceptron (MLP), TakagiSugeno-Kang (TSK) fuzzy classification, the binary statistic classifier, and the k-nearest neighbors (KNN) clustering algorithm	~ 97%
Zhang [17]	A model for screening and diagnosis of mammograms	Mammogram images from DDSM	a multi-view feature fusion neural network model	94%
Ayelet Akselrod-Ballin	Predicting breast cancer	52,936 images were collected in 13,234 women from the range of 2013 and 2017,	Deep learning	~ 91%

### 3 Proposed System

Concerning the writings, this composition presents a computer-aided diagnosis and detection (CAD) framework to arrange normal and dangerous mass sores from mammogram picture tests with the assistance of different deep learning (DL) design models and assess their exhibition. Here, we will utilize residual networks otherwise called ResNet, VGG, GoogLeNet/Inception, and AlexNet (Fig. 1).

From the above figure, we are taking the digitally stored mammogram images from the publicly available database known as digital database for screening mammography (DDSM). The next step of the project is preprocessing of the mammogram images. In the image preprocessing step, the undesirable articles are wiped out from the mammogram images, which includes explanations about the image, marks, and the noise in the image. The preprocessing step enables the localization of the area for irregularity search.

**Fig. 1** Module description of proposed system

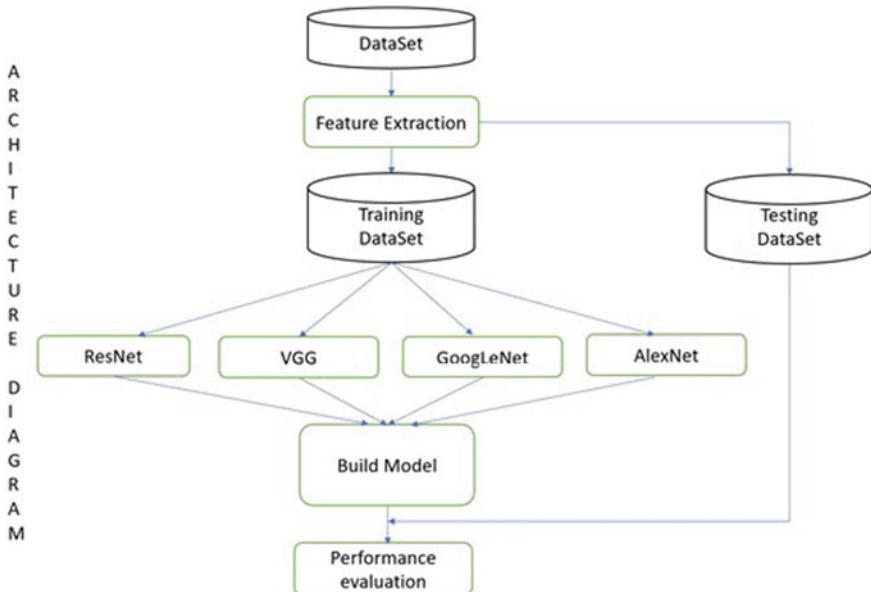


The strategy of image enhancements is utilized to improve the mammogram picture quality as far as improving the differentiation and expanding its comprehensibility. The picture improvement step encourages the framework to distinguish the mammographic injuries when the visibility is poor and the contrast by improving it. The significant objective of image enhancement of mammogram images is to improve the qualities of images with low contrast.

The fragmented zone is significant for the feature extraction process and strange tissue detection in the mammary gland, and it should be very much focused and exact. In this manner, the division cycle of the picture is essential to extricate a ROI that gives a precise estimation of the breast region with anomalies and ordinary regions. Picture segmentation step includes the essential step of isolating the breast region from the background and plans to isolate the bosom areas from different items.

After the image segmentation process, we pass the image into different CNN models and change the layered properties of these models. The CNN models will study themselves the characteristics of mammogram images, and they will extract the features from it. For example, if we take a ResNet CNN model, we pass the training mammogram images into the model. After giving the input, this pretrained model knows how to extract the features of an image. So, with help of transfer learning technology, we extract the features from a mammogram image.

The classification of image step is totally depending on other transitional steps, such as mammogram image segmentation and feature extraction of mammogram images. In this paper, I am going to use a SVM model also known as support vector machine, i.e., to achieve better accuracy in classification, we can attach the support vector machine (SVM) at the end of our CNN model. From the various literature studies, we found that a SVM model has better accuracy compared to other classification techniques.



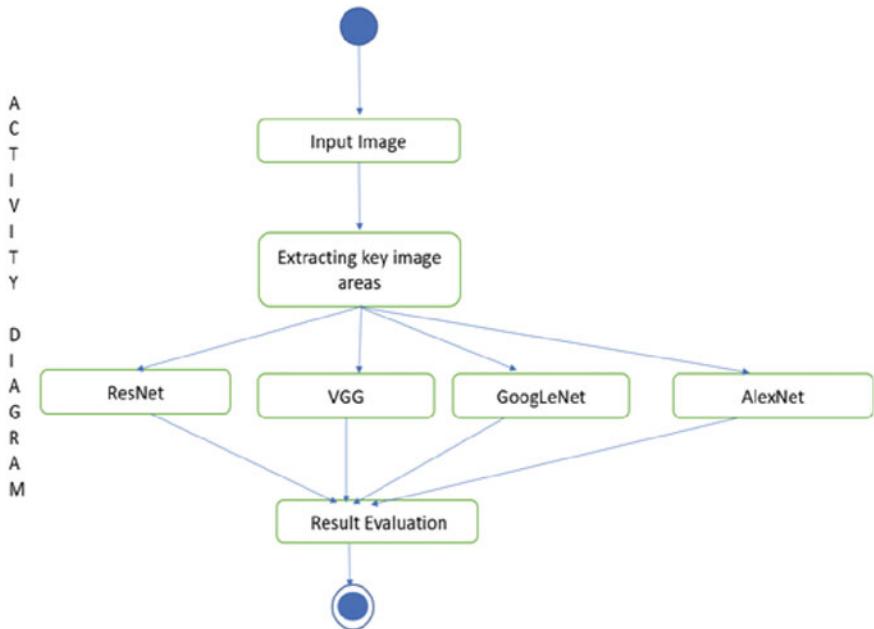
**Fig. 2** Architecture of our proposed system

Here, we are using transfer learning approach, so by changing the properties of CNN models, i.e., changing the properties of each layers we get various performance for the CNN models. Then we can evaluate the performance of each model, and it will help to perform an effective comparative study (Fig. 2).

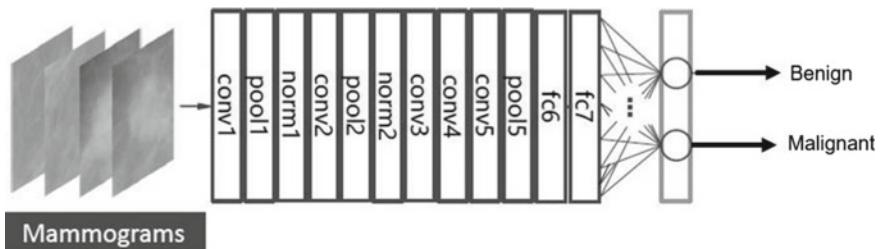
Here, we used breast mammogram image datasets from digital database for screening mammography (DDSM). After performing the above-mentioned preprocessing tasks, we split those images into training image datasets and testing image datasets. Using the preprocessed training dataset, we pass those images into CNN models like ResNet, VGG 16 and 19, GoogLeNet, and AlexNet. Along with this model, I connected an SVM for better classification of benign and malignant images. After training these models with training data, the models study the properties of images, and it will extract the features effectively. Then we can test the model using the test image dataset and perform the comparative study of each model.

Figure 3 shows the activity diagram of this project. Here, we can see that mammogram images as input, and from that, we are going to extract the key features from it and split it into training and testing images. These images are passed into the various CNN models and evaluate their performances.

Here in Fig. 4, it shows the basic architecture of our CNN model. We use a mammogram image dataset as input and pass it through convolutional layers. Next, the output of the convolutional layer takes as the input of the pooling layer; the number of these layers varies according to our models. In between the input and output layers of CNN models, hidden layers are there. The output layers are fully connected, and the final output is taken from it.



**Fig. 3** Activity diagram



**Fig. 4** Fine tuning of an CNN architecture

#### 4 Conclusion and Future Enhancement

In the present day and age, advanced image preprocessing has incredible significance in different zones of innovative works. Advanced image processing is utilized to handle computerized pictures and produce helpful qualities from the data, which would then be able to be utilized to settle on basic choices with high exactness. These procedures are additionally applied for application in the clinical field, explicitly in the region of identifying mammary gland cancer. In the present-day situation, breast malignancy is one of the significant reasons for death among ladies, and it is hard to forestall bosom disease as the principle reasons hidden bosom malignant

growth stays obscure. Notwithstanding, certain characteristics of bosom malignant growth, which contains masses and micro calcifications noticeable in the mammogram pictures, can be utilized for early detection and consequently are exceptionally advantageous for ladies who might be in danger of mammary gland cancer. Previously physicians analyzed the mammogram images and make conclusions according to that, but now with the help of deep learning (DL), we can find out whether it is benign or malignant within a short time span and more accurately. Here, we used different deep learning (DL) architecture models and evaluate their performance in the area of cancer prediction.

This project has a wide range of scope. In the future, we can implement any kind of medical image processing technique for various fields, and it will help the doctors or technicians to identify the actual problem, and they can finalize what all are the necessary treatment methods can be taken for that particular disease within a short span of time and more precisely. Now, we are using it in the area of mammogram image analysis; by making some modifications, we can extend this to the field of any medical image analysis.

## References

1. L. Zou, S. Yu , T. Meng, Z. Zhang, X. Liang, Y. Xie, A technical review of convolutional neural network-based mammographic breast cancer diagnosis, in *Hindawi Computational and Mathematical Methods in Medicine*, Vol. 2019, Article ID 6509357, 16p
2. C. Varma, O. Sawant, An alternative approach to detect breast cancer using digital image processing techniques, in *International Conference on Communication and Signal Processing*, Apr 3–5, 2018, India
3. K.J. Geras, R.M. Mann, L. Moy, Artificial intelligence for mammography and digital breast tomosynthesis: current concepts and future perspectives
4. Farahnaz Sadoughi, Zahra Kazemy, Farahnaz Hamedan, Leila Owji, Meysam Rahmani Katigari, Tahere Talebi Azadboni, “Artificial intelligence methods for the diagnosis of breast cancer by image processing: a review”.
5. M.M. Mehdy, P.Y. Ng, E.F. Shair, N.I. Md Saleh, C. Gomes, Artificial neural networks in image processing for early detection of breast cancer
6. S.J.S. Gardezi, A. Elazab, B. Lei, T. Wang, Breast cancer detection and diagnosis using mammographic data: systematic review. *Comput. Struct. Biotechnol. J.* **13**, 8–17 (2015)
7. S. Charan, M.J. Khan, K. Khurshid Breast cancer detection in mammograms using convolutional neural network, in *2018 International Conference on Computing, Mathematics and Engineering Technologies-iCoMET* (2018)
8. D.A. Ragab, M. Sharkas, S. Marshall, J. Ren. Breast cancer detection using deep convolutional neural networks and support vector machines
9. G.R. Nitta, T. Sravani, S. Nitta, B. Muthu, Dominant gray level-based K-means algorithm for MRI images. *Heal. Technol.* **10**(1), 281–287 (2019). <https://doi.org/10.1007/s12553-018-00293-1>
10. D.V. Pham, G.L. Nguyen, T.N. Nguyen, C.V. Pham, A.V. Nguyen, Multi-topic misinformation blocking with budget constraint on online social networks. *IEEE Access* **8**, 78879–78889 (2020). <https://doi.org/10.1109/ACCESS.2020.2989140>
11. P. Giri, breast cancer detection using image processing techniques
12. D. Abdelhafiz, C. Yang, R. Ammar, S. Nabavi, Deep convolutional neural networks for mammography: advances, challenges and applications

13. L. Shen, L.R. Margolies, J.H. Rothstein, E. Fluder, R. McBride, W. Sieh, Deep learning to improve breast cancer detection on screening mammography
14. N. Safdarian, M.R. Hediyezadeh, Detection and classification of breast cancer in mammography images using pattern recognition methods
15. D.P. Pati, S. Panda, Extraction of features from breast cancer mammogram image using some image processing techniques
16. A.-A. Nahid, Y. Kong, Involvement of machine learning for breast cancer image classification: a survey
17. C. Zhang, J. Zhao, J. Niu, D. LiID, New convolutional neural network model for screening and diagnosis of mammograms. Published in *BMC cancer* (2018)
18. M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, S. Mougiakakou, Lung pattern classification for interstitial lung diseases using a deep convolutional neural network. *IEEE Trans. Med. Imaging* **35**(5), 1207–1216 (2016)
19. K. Sirinukunwattana et al., Locality sensitive deep learning for detection and classification of nuclei in routine colon cancer histology images. *IEEE Trans. Med. Imaging* **35**(5), 1196–1206 (2016)
20. Z. Yan et al., Multi-instance deep learning: discover discriminative local anatomies for body part recognition. *IEEE Trans. Med. Imaging* **35**(5), 1332–1343 (2016)
21. G. Van Tulder, M. de Bruijne, Combining generative and discriminative representation learning for lung CT analysis with convolutional restricted boltzmann machines. *IEEE Trans. Med. Imaging* **35**(5), 1262–1272 (2016)
22. A. Payan, G. Montana, Predicting Alzheimer's disease: a neuroimaging study with 3D convolutional neural networks. *arXiv preprint arXiv:1502.02506* (2015)
23. E. Hosseini-Asl, G. Gimel'farb, A. El-Baz, Alzheimer's disease diagnostics by a deeply supervised adaptable 3D convolutional network. *arXiv preprint arXiv:1607.00556* (2016)
24. S.M. Anwar, M. Awais, A. Farooq, A deep CNN based multi-class classification of Alzheimer's disease using MRI. Presented at the IST (2017)
25. J. Ma, F. Wu, J. Zhu, D. Xu, D. Kong, A pre-trained convolutional neural network-based method for thyroid nodule diagnosis. *Ultrasonics* **73**, 221–230 (2017)
26. W. Sun, T.-L. B. Tseng, J. Zhang, W. Qian, Enhancing deep convolutional neural network scheme for breast cancer diagnosis with unlabeled data. *Comput. Med. Imaging Graph.* (2016)
27. H. Pratt, F. Coenen, D.M. Broadbent, S.P. Harding, Y. Zheng, Convolutional neural networks for diabetic retinopathy. *Procedia Comput. Sci.* **90**, 200–205 (2016)

# Media Bias Detection Using Sentimental Analysis and Clustering Algorithms



Sachin Rawat and G. Vadivu

**Abstract** Biased media or the coverage of slanted news can have strong impact on perception of the public on topics reported by media. In the recent years, researchers have developed many comprehensive models for describing media bias effectively. These methods of analysis are manual and therefore cumbersome. Whereas in computer science and especially NLP has developed fast, automated and scalable methods which systematically analyse bias in the news posted by media houses. Various models which are generally used for analysing bias in media using computer science models appear to be simpler as compared to models developed by social science researchers. But these computer science models do not answer the most important questions despite being superior in technology. Most of the methods used in this respect are based on supervised learning and problem is that there is not enough data to go around. Also, most of the projects generally classify news as biased or unbiased. They do not tell towards which political party or ideology news is biased. In case of Indian political news data, it is not available at all. In this project, we will try to use the latest machine learning techniques like sentiment analysis, bias score and clustering to analyse the bias in the political news articles and try to group them on the basis of their media houses to predict which media houses are biased to which political parties and how much. For it, we will first collect political news article of India from various media houses using web crawlers and then filter them so that we can get only English news and separate them on basis of political alignments. After that we try to predict whether they are biased or unbiased and if they are biased, then we will try to predict towards which political party on basis of bias scores of the news article. On basis of many news articles, we will try to form a report on, which media house is biased to which political party and which media house provides unbiased news in India and see how media houses polarize public opinions.

---

S. Rawat (✉) · G. Vadivu  
IT Department, SRMIST Kattankulathur, Chennai, India  
e-mail: [sr9416@srmist.edu.in](mailto:sr9416@srmist.edu.in)

G. Vadivu  
e-mail: [vadivug@srmist.edu.in](mailto:vadivug@srmist.edu.in)

**Keywords** Clustering · Sentiment analysis · DBSCAN · PCA · K-Means · VADER

## 1 Introduction

The clever and artistic use of language is very important in politics. The use of language by politicians and their media advisors has attracted considerable interest among scholars of political communication and rhetoric and computational linguistics. Many politicians while running for various offices give speeches and write books and manifestos for putting forward their ideas and agendas. However in during every election season, there are news of candidates back rolling on their promises and agendas through interviews and public speeches.

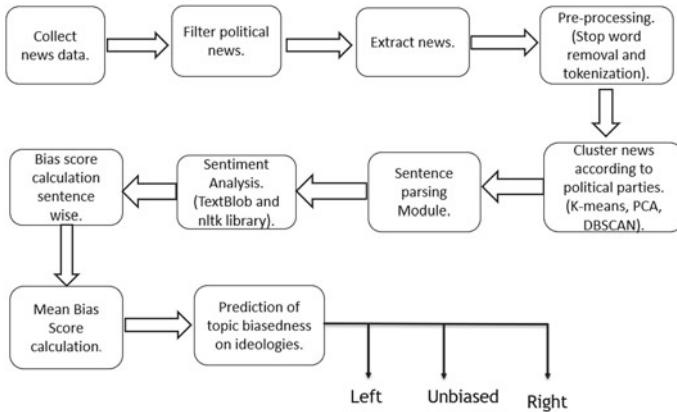
A more general observation by many experts is that many politicians tend to appear more of “centre” or “neutral” ideology candidate during election season to attract votes from all sections of society. This is done mainly due to some famous and effective old political theories and strategies which are collectively called as “median voter theorem”. It states that when be a set of voters that are more ideologically concentrated are replaced by a set of voters who are dispersed more widely across the ideological spectrum, then the politicians will present themselves as more “moderate” in order to attract enough voters to their cause in order to win the election.

Media houses provide these facilities to political parties to change their image and appear as more “centred”. These media houses nowadays not only try to clean the image of political party but also try to malign rival party’s image. This problem not only exists in India but all over the world.

Our project aims to calculate what news is biased and by what percentage towards each political party. We will categorize news as left biased (biased towards left ideology parties or their allies) and right biased (biased towards right ideology parties). Here, we will first cluster news using clustering methods on the basis of political party to which they belong (news are partitioned in two clusters on basis of political party to which they belong). The clustering algorithms used here can be *K*-means, PCA and DBSCAN. After that, we will use sentence parser to extract article sentence-wise. We will then preprocess this sentence to remove punctuations and stop words. After that, we will use sentiment analysis to find whether sentence has positive, negative, or neutral sentiment. After doing it for each sentence, we will calculate bias score for each news article as:-

$$\text{Bias Score} = (\text{No. of positive or negative or neutral sentences}) / (\text{Total no. of sentences in article})$$

To know bias percentage, we will first look to which political party the algorithm belongs. Positive sentiment means positive bias sentence, negative sentence means negatively biased sentence and neutral sentiment means unbiased sentence. After that, we will calculate percentage of each bias and then calculate overall bias on the



**Fig. 1** Architecture diagram

basis of bias percentage; i.e., article will be biased towards political party mentioned in article if positive bias percentage is highest and biased towards other party if negative bias percentage is highest. Article will be unbiased if neutral bias percentage is highest.

After that total bias percentage of each media house can be calculated on basis of total no of biased article towards a particular political ideology and total no of articles posted by them. For it, we will use five media houses and will try to collect as much political news articles as possible for each media outlet (Fig. 1).

## 2 Literature Survey

Biased media or coverage of slanted news can have strong impact on the perception of the topics which are being reported by the media houses [1]. The main objective of biased news detection is to provide “balanced reporting” and prevent spread of any misinformation. It also includes distortion of facts, lack of transparency in news, selective omission of certain facts and imbalanced reporting [2]. In the recent years, researchers have developed many comprehensive models for describing media bias effectively. These methods of analysis are manual and therefore cumbersome. Whereas in computer science and especially NLP has developed fast, automated and scalable methods which systematically analyse bias in the news of various media houses. Various models which are generally used for analysing bias in media using computer science models appear to be simpler as compared to models developed by social science researchers [3–5]. But these computer science models do not answer most important questions despite being superior in technology. “Computer science models” thus profit from the integration of models for “media bias” study in “social sciences” and automated models developed using computer science

algorithms. These models are generally automated and are used in natural language processing [1]. The scope of article defines if the bias is manifested between news sources (external) or inside a news source (internal). The type defines if the bias is present in the set of all news from a news source (generic) or in the news that refer to an entity (specific). We do not define internal generic bias because, for a specific news source, there is no fair point of comparison of its generic language profile. News sources that occasionally produce biased news articles do not represent a problem.

The problem arises when the sources tend to support or oppose systematically certain entities. If the bias is recurrent, it becomes an integral part of the language profile of the news source, when referring to that entities. Bias is not acceptable for news articles, as they are supposed to be grounded by facts; bias is acceptable in opinions, even if published as part of news articles [6].

There are many supervised algorithms like SVM, naive Bayes and many artificial neural networks like RNN, CNN, GRUs and LSTMs. Among these algorithms, each algorithm is found to be most accurate in different scenarios and languages [7]. For using RNNs and multilayer perceptron, Python's scikit library can be used. It not only provides MLP but also with functions to randomly split dataset in test and training data. This can later be packaged as a Google Chrome extension using flask server to detect and show the user that how much news is biased towards conservative party, how much to liberals and how much is unbiased. It shows this information as a pop-up in form of a percentage bar. The only problem with MLP is its quite low accuracy at about 70% as it is difficult to tune hyperparameters in it which include number of hidden layers, number of epochs, learning rate, etc. [8].

There are many semi-supervised algorithms also which can be used to for bias detection. Some of them are: random walk with restart (RWR), local consistency global consistency (LCGC) and absorbing random walk (ARW). Among these algorithms, ARW is found to be most accurate at 97%. It was most accurate than many supervised learning algorithms like SVM which was about 92% accurate [9].

The task of “event annotation” is very difficult for many reasons as many articles of news may refer to various events. Sentences which refer to the same event are often scattered throughout the article without any “sequential pattern”. In the area of text summarization, the clustering of similar sentence problem has been investigated. We are using clustering at sentence level [10]. “Text clustering algorithms” are often classified into several groups:  $k$ -means variations, “vector space models”, spectral algorithms, phrase-based methods, dimensionality reduction methods and generative algorithms. “ $K$ -means algorithm and its extensions” are most popular for “partitioned” and “hierarchical clustering”. “Vector space models” generally show better results on same topics, and number of clusters must be known in advance to use it. These methods have many disadvantages: like the effectiveness of these methods decrease on large datasets. It generally relies on random initialization. They are also very susceptible to noise and outliers. They also need to know the number of clusters required in advance. In contrast to it the dimensionality reduction methods like PCA which were developed originally for “computer vision applications” can also be used for document’s clustering.

Their only drawback is that they rely on “random initialization” which can produce different results for different runs on similar datasets. But they generally have very high performance and can also estimate counts of the clusters [11]. Two problems must be solved in order to group text blocks into clusters or articles. They are:

1. Finding best-describing features is the objectives of document in an article.
2. Finding “features” which is best in distinguishing document objective in an article from the object of other’s article.

The first feature set explains about the “intra-cluster similarity”, whereas second one explains about the “inter-cluster similarity”. Every article or cluster is defined by intra-cluster similarity or set of words it has. It is different from others as it does not have terms which other articles have or “inter-cluster dissimilarity”. The best method to check whether two blocks belong to the same article or not is to compare the words in them. If they have “same words” or “group of words” or “synonyms”, then they are referring to same topic. All the terms in article are not useful in explaining the contents of the articles equally. Their frequency must be considered for it. In “Information Retrieval”, weight is assigned to each word in the block. This value generally contains “quantification” for both “inter-cluster difference” and “intra-cluster similarity”. The first one is represented by the “rarity of the term” in the page, whereas second one is measured “frequency of the term” in the block. It also provides measure of “how well it describes the subject of the block”. These two values are combined together as one value for measure [12]. To search for the group of articles with high in-group similarity and low out of group similarity are based on paragraph vectors. This classic clustering problem can be approached in a myriad of ways, from simple hierarchical clustering to k-means to spectral clustering. Similarity, between articles can be obtained by pairwise cosine similarity. These can be constructed as a weighted adjacency matrix [13].

Use of the existing k-means method for news headlines clustering from corpus of dataset is done through various document preprocessing and text mining techniques. This idea can be explored further and can be used to develop a new system related to web data mining. After the document vector is constructed the process of clustering is carried out [14].

Bias news detection can also be developed into a mobile application which can later be used for auto-detecting the news bias when the user is reading a particular news in his mobile. This type of application is currently in development in Samsung RandD but due to the small datasets and testing implementations, it has not gone beyond first development phase and still requires a lot of work [15].

News can also be checked for bias using sentiment analysis. It includes separating news into different sentences using sentence POS tagger and then find sentiment for it, i.e. positive, negative or neutral. Subject of news can be find using POS and NER tags for identifying subject in the text. Then overall sentiment can be found by finding average of sentiment score for entire article. If it is above certain threshold then news are biased else unbiased. It does not tell towards which political party news is biased and tells only if news is biased or not [2]. Sentiment analysis is an area of research under natural language processing (NLP). It classifies text pieces based

on the expressions or opinions expressed. This sub-specialization of identifying and then classifying biased news is growing more and more in this era of fake news. There are many methods which are being used by researchers nowadays. There are many machine learning techniques and text embedding techniques which have grown leaps and bounds in the past few years. There are many sentence embedding techniques for embedding models are: GloVe, word2vec and fastText. This embedding uses supervised machine learning models. Some of them are: SVM, CNN, RNN and neural networks. Data is generally collected through AMT [6]. The main difference between fastText and other word embedding models (word2vec, GloVe) is that fastText treats each word as n gram model or composed of characters. The vector of word is also made of this n gram character model. Word2vec and GloVe like algorithms use words as the smallest unit for training. Therefore, fastText can generate better word embeddings for rare words and words out of vocabulary as compared to Word2vec and GloVe [16].

For collecting data, many corpuses can be used. One of the most famous technique is to use ideological corpus which includes many political writings by various authors whom are perceived to have a certain ideology. In addition to it, each part of the writing can be classified and labelled as per the ideology it reflects. Researchers can also infer ideological cues which are the terms strongly associated with the ideology. Sparse additive generative (SAGE) probabilistic model can be used to assign probability to text [17].

Data preprocessing is also an important step as it can be used to remove stop words, normalize text in one format and split news articles first into sentence and then into tokens using various embedding techniques. Weightage of each factor can be determined using TF-IDF algorithm [18]. There are many limitations to it as many sentences may not provide similar weightage to the bias in news and can be less relevant. Sentiment analysis can also be done at various levels like:

1. Document level: Sentiment analysis of entire document.
2. Sentence level: It can involve sentiment analysis of sentences on the basis of grammar as well as semantics.
3. Feature level: It is pinpoint approach where exact opinion about entity can be correctly extracted.

Many researches have used NPOV rules and lexicon to find the sentiment and bias of the news as it is one of the best lexicons available today for bias and sentiment analysis [15]. There is HMM model called CLIP which can be used to find towards which the given text is biased. It uses bigram and trigram lexicons and SAGE model to find towards which political party news is biased [17].

To minimize the processing, time and some researches have also proposed to use headlines for bias analysis as it will provide a brief introduction to the entire article and help us determine the mindset of the author and article as whole. It is very fast but less accurate than analysing entire article as sometimes article headline may be tacky and involve use of lots of abbreviations and idioms which become difficult to analyse. The most common supervised algorithms done for headlines analyses are SVM and multiclass naive Bayes algorithm [19].

Lexicon-based sentiment analysis methods generally use  $n$ -grams, negation handling, feature selection by mutual information can lead to better efficiency. It focuses on generalize methods for number of text categorization problem and improving. Generally, the information about the sentiment is conveyed by the adjectives or combination of adjectives with some other parts of speech. The information is captured by adding various features like bigrams or trigrams. [20]

Lexicon-based sentiment analysis generally uses a predefined list of words where each word is related or mapped to a particular emotion.

It uses three methods:

**Dictionary-based methods:** Here dictionary is used for finding “positive sentiment words” and “negative sentiment words”.

**Corpus-based methods:** It uses large corpus of words. It is based on “syntactic patterns” and other opinion words which can be found in the context.

“Sentiment analysis” can be applied at “document level”, “sentence level”, “word level” or “phrase level” [21].

Lexicon-based methods can be used for predicting sentiment on short texts. “Sentiment scores” can be calculated by analysing the articles which are collected by “VADER algorithm”. “VADER” is a very popular algorithm for “sentiment analysis” for the texts in social media. Other famous lexical-based methods are SentiStrength and Afinn. VADER has open-source code and can be applied for large datasets. “NLTK-VADER” is very famous method for “sentiment analysis”. It has the underlying implementation of rule-based models. This library is used in python and is used to determine the sentiment score or polarity score (PS) for the text given. Sentiment score determined by this method is used to calculate polarity of the text. Polarity scores generally range from  $-1$  which is strongly negative to  $+1$  which is strongly positive. This range has three regions which are positive, negative and neutral. PS values from  $-1$  to  $-0.33$  are negative region, PS values in range of  $-0.33$  to  $+0.33$  are neutral region, whereas PS values ranging from  $+0.33$  to  $+1$  are positive region [22].

### 3 Existing Model

The existing system model is generally constructed using supervised learning algorithms like SVM, neural networks and LSTMs. There were some models which also use semi-supervised learning algorithms also. Most of these models were made for the USA political news dataset, and dataset was annotated using AMT by the researchers only as annotated dataset is not easily available. Most of the existing systems only classify dataset as biased or unbiased. Most of them do not show towards which political party dataset is biased. There were few models which were constructed to show towards which political party news is biased and were developed as Google extensions and mobile applications. But they were in testing phase only.

These models were also event-based like US Presidential elections and Israel/Palestine conflict. There was only one model constructed for Indian political

news and that too was for Telugu news. In its dataset was collected and annotated by developers only. One of the models constructed is a sentiment analysis based and used polarity scores to find whether data is biased or not. It does not classify biased news on basis of political parties.

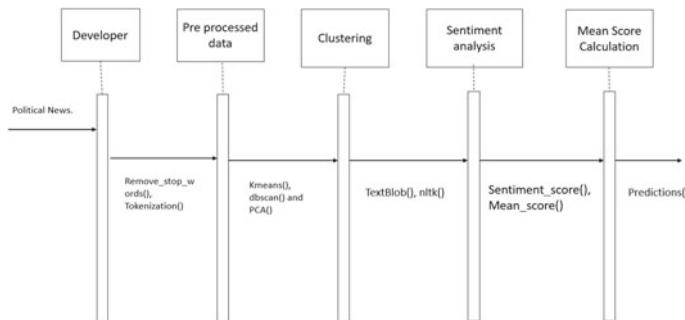
## 4 Proposed Model

The steps involved in construction of proposed model are:

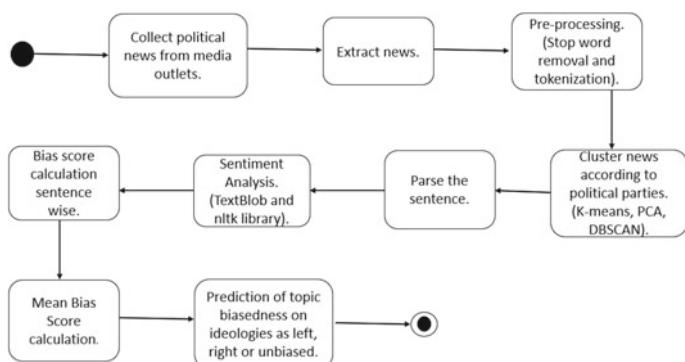
1. Collection of Indian political news using web crawlers.
2. Clustering of news article according to the political party. It depends on subject of article.
3. Clustering is done using K-means, PCA and DBSCAN algorithms.
4. Doing sentiment analysis of entire article to find whether it is positive, negative or neutral at sentence level.
5. Sentiment analysis is done using nltk corpus-based method called VADER and TextBlob which is rule-based method.
6. Calculate bias score for article using formula no of negative or positive sentences divided by total sentences.
7. If article is above certain threshold, then it is biased towards that political party if article is positive, towards other party if it is negative and unbiased if it is neutral.
8. Calculate no of biased articles according to political party and unbiased articles for each media outlet.
9. Find percentage of biased and unbiased articles for each media outlet.
10. Display all these results as a graph.
11. Take test data (political news articles) from the website of a media house and predict towards which political party the news is biased or if the news is biased.
12. Check test data accuracy to find if model is working or not.
13. Display accuracy of the models and find best algorithm for clustering as well as sentiment analysis (Figs. 2 and 3).

## 5 Conclusion

As opposed to many other projects, we are using unsupervised learning approach to cluster data into political parties. It will use Indian and US political news published in English by various media houses. Here using classical sentiment analysis method to find whether news has positive, negative or neutral tone towards a particular political party. It will help to understand towards which political party the news article is biased. These results can be then used to find how many articles of a particular media house are biased towards which political party and how many are unbiased. It



**Fig. 2** Sequence diagram



**Fig. 3** Activity diagram

will help to provide a broader picture on how much-biased news a particular media house is publishing.

In the near future, improving the accuracy of both clustering and sentiment analysis algorithms is by using a larger corpus. It can also be used to make a mobile application as well as Google Chrome extension for detecting biased political news for English political news. These works can be added in near future for making the project more useful to the general public use.

## References

1. F. Hamborg, K. Donnay, B. Gipp, Automated identification of media bias in news articles: An interdisciplinary literature review (2018)
2. S. Bharathi, A. Geetha, Determination of news biasedness using content sentiment analysis algorithm (2019)
3. S.-I. Chu, B.-H. Liu, N.-T. Nguyen, Secure AF relaying with efficient partial relay selection scheme. Int. J. Commun. Syst. **32**, e4105 (2019). <https://doi.org/10.1002/dac.4105>

4. M. Balaanand, N. Karthikeyan, S. Karthik, Envisioning social media information for big data using big vision schemes in wireless environment. *Wireless Pers. Commun.* **109**(2), 777–796 (2019). <https://doi.org/10.1007/s11277-019-06590-w>
5. A. Zahid, M.N. Khan, A.L. Khan, B. Nasir, Modelling, quantifying and visualizing media bias on Twitter (2020)
6. A.M. de Castro Ribeiro, A supervised approach to detect bias in news sources (2019)
7. M. Bellows, Exploration of classifying sentence bias in news articles with machine learning models (2018)
8. M. Vu, Political news bias detection using machine learning (2018)
9. P. Resnik, D.X. Zhou, Q. Mei, Classifying the political leaning of news articles and users from user votes (2011)
10. M. Naughton, J. Carthy, N. Kushmerick, Clustering sentences for discovering events in news articles (2006)
11. I. Blokh, V.N. Alexandrov, News clustering based on similarity analysis (2017)
12. M. Aiello, Textual article clustering in newspaper pages (2006)
13. M.T. Altunku, S.N. Yaliraki, M. Barahona, Content-driven, unsupervised clustering of news articles through multiscale graph partitioning (2018)
14. P. Lama, Clustering system based on text mining using the k-means algorithm—News headlines clustering
15. A.A. Patankar, J. Bose, H. Khanna, A bias aware news recommendation system (2018)
16. V. Minh, Political news bias detection using machine learning (2017)
17. Y. Sim, N.A. Smith, J.H. Gross, B.D.L. Acree, Measuring ideological proportions in political speeches (2013)
18. H.M.I. Lubbad, Bias detection of Palestinian/Israeli conflict in western media: A sentiment analysis experimental study (2018)
19. J.M. Sholar, N. Glaser, Predicting media bias in online news CS 229: Machine learning—Final project (2016)
20. U. Swati, C. Pranali, S. Pragati, Sentiment analysis of news articles using machine learning approach (2015)
21. S. Taj, Sentiment analysis of news articles: A lexicon-based approach (2019)
22. S. Aggarwal, T. Sinha, Y. Kukreti, S. Shikhar, Media bias detection and bias short-term impact assessment (2020)

# A Review on Video Summarization



R. V. Krishna Vamsi and Dhivya Subburaman

**Abstract** Automatic video summarization technique using natural language processing. The significance of automated video summarization is large within the new generation of big data. A video summarization helps in economical storage and additionally fast of enormous assortment of videos while not losing the necessary. Every video may be an assortment of many frames and every of those frames is truly images, and every second of ordinary video consists of twenty-four frames. The projected technique generates the summarized videos with the assistance of subtitles. We will conjointly see Machine learning techniques to develop video summarization however it had a drawback of the necessity of high-performance devices as the training period can be huge for data sets that contain videos. Even after we train using data sets that contain pictures, it takes some amount of time to process the dataset or to relinquish the acceptable results. So, it would be harder to use machine learning algorithms when put next to natural language processing.

**Keywords** Video summarization · Subtitles · Machine learning · Natural language processing

## 1 Introduction

Nowadays the utilization of consumer device like mobile phones, cameras, tablets are increasing day by day because of that capturing of videos has been exaggerated. Traffic Controller systems conjointly records immense quantity of video data and stores it within the servers. Some people are also installing personal home observance systems to monitor the protection of their surroundings. The corporate offices are

---

R. V. K. Vamsi

M. Tech (Big Data Analytics), Department of Information Technology, SRM Institute of Science and Technology, Kattankulathur, India

D. Subburaman (✉)

Department of Information Technology, SRM Institute of Science and Technology, Kattankulathur, India

e-mail: [dhivyas@srmist.edu.in](mailto:dhivyas@srmist.edu.in)

also installing the security cameras to manage their security. Due to increasing usage of Internet people tend to upload more videos. Recording videos of every and each instance of day-to-day life people tend to upload the videos in online platforms like YouTube, Flickr, Facebook and a few alternative online websites.

The average range of individuals who use YouTube are approximately 1.5 billion people and nearly YouTube gets 35 million visitors per day and for each minute 400 h of video has been uploaded to YouTube. The advance in digital media technology created recordings, collecting and maintaining of videos effortlessly. Approximately for every second 100 h of video has been uploaded into the web. As thousands of suggestions are provided for every topic. Browsing these immense volumes of videos and acquire the desired video is time overwhelming. These overabundances of videos make users to rely upon thumbnails, title, descriptions to perceive the desire one. In some cases, this metadata knowledge is inaccurate to obtain the desire information that the user wants to watch. So, there is increase in demand for video summarization.

In this aspect, video summarization has enormous extent. So, by using the video summarization we will summarize the video by removing the unnecessary content within the video. So, that the user can be able to see whatever the video in the summarized format and the user can also select the length of the summarization of the video. The video that is generated from the video summarization should representative the video and its domain. The summarized video should be made up of all the key frames in the video.

The summarized video would be vary from each user because of the time frame that what they are using to fix the length of the video. So, depend upon the time frame we need to summarize the video effectively as possible.

Focused on solve productive video summarization and also to overcome the drawbacks of the existing techniques. The summarization technique that using in this paper is based upon natural language processing. The video which consists of subtitles is being used by using subtitles we summarize the video by extracting the time frames of important subtitles. Some of the videos does not consist of subtitles videos like security camera footage or traffic control surveillance system it is complex to sum up the videos which does not consist of any subtitles. Summarization of videos utilizing the subtitles is the most effective way as we can summarize the video by using different analysed methods and using the subtitle gives the result faster. For summarization of subtitles, we can use various natural language processing algorithms which is very accurate in giving results.

We can also implement various algorithms to obtain the video summarization more accurately. Text summarization algorithm is used to filter the key elements in subtitles by using that particular key subtitle we can acquire the important frame from the video rank wise and sum up the video based upon that ranks of the video extracted from key elements of subtitles. By using ensemble technique that means combining the different algorithms and making the output more accurately.

## 2 Problem Statement

The main problem is that the people do not have much time to watch lengthy videos. So, summarization helps to solve that problem by decreasing the length of the video with main content in it.

So, using video summarization users can able to watch lengthy video within short time. Summarization can save their time. Some of the online video summarization tools are not even giving any output.

The online video summarization tools have different type of inputs which are difficult to understand for a normal person who want to summarize the video cannot able to input the parameters to get the output. Some of the online tools gives the output in the form of video.

## 3 Literature Survey

The paper [1] provides an Web service for a video summarization exploitation integrated deep learning based video analysis in this methodology Web application decomposes the video segments, evaluates the fitness of every segment to be enclosed within the video and choose the appropriate segments to summarize the video.

An effective video analysis methodology is required with the increasing availability to summarize the video. In [2], an unsupervised video summarization methodology with attentive conditional generative adversarial network method is used that means the method gives a high-level freighted frame options and anticipate the frame-level significance scores, whereas the distinct tries to differentiate allying weighted frame attribute and raw frame attribute and summarizes the video.

This paper [3] proposes unsupervised task of succeeding frame anticipation and a supervised task of scene begin recognition and suggest a loss function that directly emphasis on achieving the right stabilization between progression and divergence within the generated summary.

The paper [4] proposes a Cluster based mostly scene identification approach could be a two-step method identifying scenes and choosing representative content for every scene within which Global features are utilize to spot the scenes through cluster local features are used to summarize the identical scene.

In report conferred in paper [5], it uses a technique named Bag-of-visual-Texture's approach and colour information during this method the projected model decreases the number of motion picture by a component of 27, with solely inconsiderable degradation within the data content.

In report conferred in paper [6], it takes videos from online resources like open video.org. The approach takes numerous features under consideration like characteristic, consistency, fixed attention, temporal attention and standards which incorporates colourfulness, illumination, contradistinction, colour tone count, edge issuance

for choosing appropriate key frames and also the results are compared with the prevailing algorithms.

In paper [7] client quality recording videos that are encapsulated beneath uncontrolled state and have numerous contents has been used to summarize. To assist summarization numerous parameters are used like incorporating text representations, visual appearances and audio resonates. This report observes summarization in the client domain where most previous approach cannot be simply applied due to the demanding issues for content analysis.

The methodology that is used in [8] this paper demonstrates key frame extraction exploitation various related measure and Thepade's Sorted Ternary Block Truncation Code. The Block truncation programming is one amongst colour feature extraction strategies in Content-Based Video Retrieval (CBVR). The extended version of Block Truncation Coding (BTC) is Thepade Sorted Ternary Block Truncation Code and summarizes the video.

The paper [9] provides a technique of video summarization exploitation highlight identification and pairwise deep ranking method. Two-dimensional Convolutional Neural Network (CNN) is utilized to take advantage of spatial data wherever a three-dimensional Convolutional Neural Network (CNN) is utilized to take advantage of temporal data to produce highlight ranks for segments of the video and summarizes it by using that segments.

In paper [10] the projected model has an encoder-decoder anatomy. The encoder obtains a succession of video features through Convolutional Neural Network, and also the decoder predicts frame-level significances ranks from the wide and deep network consistent with that frame-level importance the video is summarized.

In paper [11] the technique used was video summarization via block scattered dictionary selection VSum and TVSum datasets are summarized using this technique. During this technique the evolution of sparse representation based mostly perspectives the block-scatters of major keyframes is taken into deliberation, by which the VS problem is formulated as a block scattered dictionary selection method. According to sparsity of the keyframe selection the video is summarized.

Frame cluster technique towards single video summarization is a projected model in [12] this methodology works on repetition concepts like massive number of videos in domains like trip guide, documented or chronicled, screenplay. This paper presents a novel key frame cluster method for producing very compressed summaries by combining all frames of alike concepts together regardless of their occurrence progression. In this methodology similar keyframes that are continuance within the video has been removed. Exploratory related outcomes uphold the potency of the projected method in producing compressed video summaries with frequent concepts.

Paper [13] uses a technique which is applied solely on one new public dataset named MVS1K, a completely unique query-aware method by constructing the different video summarization using a sparse programming framework, where the online pictures explored by a query are taken as the major preference details to disclose the query objective. To produce an easy summarization, this paper conjointly expands an event-keyframe information structure to present keyframes in category of

particular events associated with the query by utilizing an unsupervised multi-graph fusion methodology. By exploitation of these strategies, it summarizes the video.

A Video Summarization Approach Based on Machine Learning [14] in this proposed methodology an approach based on machine learning is developed for video summary prediction. Various novel options are taken out to characterize video border, in conjunction with cut, decline in, decline out and disintegrate to ease the acknowledgement of content description and domain regulations of a video. The outputs have shown that the method can precisely predict the transitions in a video order and would be an empirical solution for automated video segmentation and video summarization.

This paper [15] mainly proposes a methodology that this technique generally works on accumulating a new scrutiny video dataset for the objective estimation. This paper says that it gives up to the minute performance results for object motion-level and frame-level work. A networked motion-AE model is used for key item motion-based video summarization. Newly gathered scrutiny dataset and people datasets have demonstrated the usefulness of the projected method.

In this paper [16] a methodology is described it collects the leverages an organized minimum-risk differentiator and systematic submodular inference. To check the precession of the forecasted summaries this paper utilizes a newly suggested measure (V-JAUNE) that examines both the content and frame sequence of the actual video. This paper show that the suggested method outputs more precised summaries than the compared minimum-risk and syntactic methods.

This paper proposed a method [17]. It is developed using a patch form technique and a assemble approach. The initial step suggests a Definite Frame Patch implication for choosing a collection of superior candidate frames, and also the subsequent step suggests a unique appearance-based mostly linear clustering to refine them for definite ones. This paper validates the results over two public accessible datasets.

This paper [18] is based on high-level features. The proposed system could be a domain adaptive video summary structure supports top ranking features in such a way that the summarization video will encapsulate the key contents by guaranteed minimum number of frames. One of the top ranking features extraction is native binary arrangement. Major frames can be taken out once after finding the Euclidean space in the middle of the native binary order shape in numerous ways. The major frames are categorized utilizing the k-means clustering algorithm and also by means of thresholding.

In [19] this paper describes about the video summary model built on using the Priority curve algorithm architecture and various image processing algorithms. The proposed model is applied on online soccer video. The paper executed the precedence curve algorithm and differentiated it with other summarizing methods in the literature survey with respect to both performance and the output grade.

In this paper, the proposed model [20] describes the methodology major frame withdrawal and orthogonal alters such as Cosine, Slant and Kekre transforms are the techniques used in this paper. The key frame withdrawal is done using various percentage values segmented energy coefficients of altered video frames. Here, the

integrity estimate is defined to check accomplishments of each of the amalgamation for every altered video clip. Experimental results shows that the video content summary of orthogonal altered video gives better completeness.

The proposed methodology [21] in this paper is, the effective content in every video frame is taken. The combination of multimedia system associated with nerve cell signals provides an interlink that connects the digital illustration of multimedia system with the observer's perceptions and then divide-and-conquer based framework is used for an effective summary of big video data. The projected attention methodology gives a precise reflection of the user liking and ease the extraction of extremely affective and customized summaries.

This paper states the [22–24] designed attention model in two-stage hierachic formation for building numerous detailed maps, novel framework named Hierarchical-Multi-Attention Network (H-MAN) which contains the frame-level reconstruction model and multi-head attention model. By, using these techniques the video summarized and the paper says that the desired model of quantitative and qualitative results demonstrates the performance of the model.

In this paper, a methodology is proposed based on [25] clustering effectively by adapting the hierachic clustering data model to temporal ordered clusters. Utilizing a new multi-stage hierachic clustering methodology, the latest image feature outputs from normalising colour channels for frames and then creating 2D histograms of chromaticity as images and compacting these. Because this method efficiently decreases a video to the same lighting circumstances, so that it can recalculate a universal foundation on which to forecast the video frame feature vectors and summarize the video.

In the paper a methodology is projected based on the reinforcement learning [26] The training set is especially sub divided into two elements, video cut by movement parsing and video summary supported reinforcement learning. In the initial part, a consecutive multiple occurrence learning model is trained with debile explicated data to resolve the issue of full explicated time overwhelming and weak explicated ambiguity. In the next part, we tend to outline a deep continual neural network-based video summary. Experimental results show that the derived major key frames might be estimated by the classification precision. Experiments and differentiation with successive strategies demonstrate the advantage of the projected methodology.

Hierarchical Video Summarization Extraction Algorithm in Compressed Domain [27] we can able to obtain the reduced data Information System from DC coefficients and Rough sets theory. Since the main accommodated all the knowledge in video succession, and at a same time it expelled unessential video frame, therefore it may be observed as the efficient summary illustration. Hierarchical video summary illustration algorithm, was projected for video analysis in compacted domain. Experimental results shows that the representation becomes more scientific and efficient than preceding methods.

The paper [28] proposes a technique of content explanation-based video summary for video journal a new video summary technique for helping users to generate a

video journal post. Experimental results shows that the projected model is acceptable to generate video journal posts compared with typical methodologies for video summarization.

The paper [29] provides a method of automated video summary depended on MPEG-7 interpretations. Enlarging demand to expand systems able to automatically summarize audio visual details. This paper suggests a completely unique query-based concise formation mechanism using a connexion criterion and a limitation schema enforced within the context and summarizes the video.

A Framework for Scalable Summarization of Video [30] proposes expandable summary as a technique to easily modify the summary, according to the prerequisites in each case, onwards with an acceptable framework and the paper also uses another technique of unique continual categorization procedure in which each summarization is the output of the extended of the previous one, balancing detailed coverage and visual pleasantness and the video is summarized using bitstream extraction.

The paper [31] provides a way of Video Summarization Using R-Sequences. The issue of taking out a hard and static range of typical frames to summarizing a given digital video. To resolve it, formulated associate algorithm referred to as content-based accommodative categorization. In our algorithm, shot border identification is not required. Video frames are considered as points within the multi-dimensional feature space similar to a low-level feature like colour, movement, appearance and texture and therefore the video is summarized.

The paper [32] provides a technique of Video summarization by Contourlet remodel and structural similarity. Contourlet alter are analysed orderly to find the frame changes. Finally, Renyi Entropy are often accustomed to derive most applicable frames from categorization to output the full motion summarized video.

## 4 Related Work

Figure 1. shows the planning of the proposed system. The Fig. 1 tells regarding the step by step method of the project that however it is getting to be implemented. The project will be conducted on the subsequent phases.

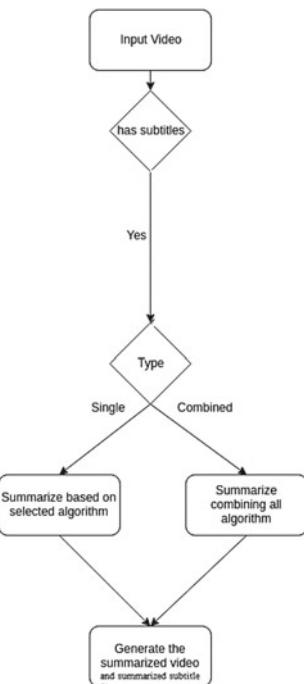
**Step 1:** Input Video in this phase we need to upload a video from local storage and the subsequent file of the subtitles of the video. If subtitles of the video is not present we can get the subtitle file from the Internet and we need to upload it.

**Step 2:** At this phase it checks the subtitles file is present or not if present then only it goes to the next phase.

**Step 3:** At this phase of we need to decide that what type of algorithms we are going to implement to give the summarization.

**Step 4:** At this phase we can able decide a single algorithm based summarization or combined algorithm based summarization. If it is a single based summarization it can summarizes the content based on selected algorithm or if it is a combined algorithm based selection all the algorithms are combined to summarize the video.

**Fig. 1** Architecture for the proposed system



**Step 5:** At this phase the generated video summarization is ready including the text document of the summarized subtitle file.

## 5 Conclusion

Automated video summarization helps people to save their time by removing unnecessary video clips from the video. A best video summarization needs to be representation of the whole video and need to maintain the semantic description of the entire video. The projected solution aims to utilize natural language processing methods. The video is summarized by using the subtitles of that particular video. Natural language processing algorithms are used to find the best subtitles and summarizes the video based on the time stamp of that particular subtitles.

## References

1. Unsupervised video summarization with attentive conditional generative adversarial networks, in *MM '19: Proceedings of the 27th ACM International Conference on Multimedia* (Oct 2019), pp. 2296–2304

2. S. Lal, S. Duggal, I. Sreedevi, Online video summarization: Predicting future to better summarize present, in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)* (Waikoloa Village, HI, USA, 2019), pp. 471–480. <https://doi.org/10.1109/WACV.2019.00056>
3. G. Guan, Z. Wang, K. Yu, S. Mei, M. He, D. Feng, School of Information Technologies, The University of Sydney, Australia. School of Electronics and Information, Northwestern Polytechnical University, China
4. J. Carvajal, C. McCool, C. Sanderson, Summarisation of short-term and long-term videos using texture and colour, vol. 1, pp. 769–775, NICTA, GPO Box 2434, Brisbane, QLD 4001, Australia
5. M. Srinivas, M.M. Pai, An improved algorithm for video summarization—A rank based approach
6. W. Jiang, C. Cotton, A.C. Loui, Automatic consumer video summarization by audio and visual analysis
7. S.D. Thepade, P.H. Patil, Novel visual content summarization in videos using keyframe extraction with Thepade's sorted ternary block truncation coding and assorted similarity measures, in *2015 International Conference on Communication, Information & Computing Technology (ICCICT)* (Mumbai, 2015), pp. 1–5. <https://doi.org/10.1109/ICCICT.2015.7045726>
8. M. Sridevi, M. Kharde, Video summarization using highlight detection and pairwise deep ranking model
9. J. Zhou, L. Lu, Wide and deep learning for video summarization via attention mechanism and independently recurrent neural network, in *2020 Data Compression Conference (DCC)* (Snowbird, UT, USA, 2020), pp. 407–407 <https://doi.org/10.1109/DCC47342.2020.00074>
10. M. Maa, S. Mei, S. Wan, J. Hou, Z. Wang, D.D. Feng, Video summarization via block sparse dictionary selection
11. P.R. Sachan, Keshavani, Frame clustering technique towards single video summarization, in *2016 Second International Conference on Cognitive Computing and Information Processing (CCIP)* (Mysore, 2016), pp. 1–5. <https://doi.org/10.1109/CCIP.2016.7802877>
12. Z. Jia, Y. Maa, Y. Panga, X. Lib, Query-aware sparse coding for web multi-video summarization
13. W. Ren, Y. Zhu, A video summarization approach based on Machine Learning, in *2008 International Conference on Intelligent Information Hiding and Multimedia Signal Processing* (Harbin, 2008), pp. 450–453. <https://doi.org/10.1109/IIH-MSP.2008.296>
14. Y. Zhang, X. Liang, D. Zhang, M. Tana, E.P. Xing, Unsupervised object-level video summarization with online motion auto-encoder
15. F. Hussein, M. Piccardi, Minimum-risk structured learning of video summarization, in *2017 IEEE International Symposium on Multimedia (ISM)* (Taichung, 2017), pp. 248–251 <https://doi.org/10.1109/ISM.2017.41>
16. S. Kannappan, Y. Liua, B. Tiddeman, DFP-ALC: Automatic video summarization using distinct frame patch index and appearance based linear clustering
17. N.R. Aiswarya, P.S. Smitha, A review on domain adaptive video summarization algorithm, in *2017 International Conference on Networks & Advances in Computational Technologies (NetACT)* (Thiruvananthapuram, 2017), pp. 412–415. <https://doi.org/10.1109/NETACT.2017.8076806>
18. M. Albanese, M. Fayzullin, A. Picariello, V.S. Subrahmanian. A. Tonge, S. D. Thepade, The priority curve algorithm for video summarization. Key frame extraction for video content summarization using orthogonal transforms and fractional energy coefficients, in *2015 International Conference on Information Processing (ICIP)* (Pune, 2015), pp. 642–646. <https://doi.org/10.1109/INFOP.2015.7489462>
19. I. Mahmood, M. Sajjad, S. Rho, S.W. Baik, Divide-and-conquer based summarization framework for extracting affective video content
20. Y. Liu, Y. Li, F. Yang, S. Chen, Y. F. Wang, Learning hierarchical self-attention for video summarization, in *2019 IEEE International Conference on Image Processing (ICIP)* (Taipei, Taiwan, 2019), pp. 3377–3381. <https://doi.org/10.1109/ICIP.2019.8803639>
21. R.S. Ram, S.A. Prakash, M. Balaanand, C.B. Sivaparthipan, Colour and orientation of pixel based video retrieval using IHBM similarity measure. *Multimedia Tools and Applications* **79**(15–16), 10199–10214 (2019). <https://doi.org/10.1007/s11042-019-07805-9>

22. B.H. Liu, N.T. Nguyen, V.T. Pham, An efficient method for sweep coverage with minimum mobile sensor, in *2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing* (Kitakyushu, Japan, 2014), pp. 289–292. <https://doi.org/10.1109/IIH-MSP.2014.78>
23. M.S. Drew, J. Au, Clustering of compressed illumination-invariant chromaticity signatures for efficient video summarization
24. J. Lei, Q. Luan, X. Song, X. Liu, D. Tao, M. Song, Action Parsing-Driven Video Summarization Based on Reinforcement Learning. *IEEE Trans. Circuits Syst. Video Technol.* **29**(7), 2126–2137 (2019). <https://doi.org/10.1109/TCSVT.2018.2860797>
25. L. Xiang-wei, Z. Li-dong, Z. Kai, Hierarchical video summarization extraction algorithm in compressed domain
26. M. Otani, Y. Nakashima, T. Sato, N. Yokoya, Textual description-based video summarization for video blogs, in *2015 IEEE International Conference on Multimedia and Expo (ICME)* (Turin, 2015), pp. 1–6. <https://doi.org/10.1109/ICME.2015.7177493>
27. P.M. Fonseca, P. Fernando, Automatic video summarization based on MPEG-7 descriptions
28. L. Herranz, J.M. Martínez, A framework for scalable summarization of video. *IEEE Trans. Circuits Syst. Video Technol.* **20**(9), 1265–1270 (2010). <https://doi.org/10.1109/TCSVT.2010.2057020>
29. X. Sun, M.S. Kankanhalli, Video summarization using R-sequences
30. R. Hari, M. Wilscy, Video summarization by contourlet transform and structural similarity, in *2011 IEEE Recent Advances in Intelligent Computational Systems* (Trivandrum, Kerala, 2011), pp. 178–182 <https://doi.org/10.1109/RAICS.2011.6069297>

# Prediction and Grading Approach for Cataract Diagnosis Using Deep Convolutional Neural Network



**P. Nithyakani, R. Kheerthana, A. Shrikrishna, S. Selva Ganesan, and Anurag Wadhwa**

**Abstract** Cataract has become the most common cause of vision loss and blindness due to the denaturation of protein in the crystalline lens. Cataract affects the iris pattern and structure and minimizes the focus of light into the retina. Using a slit-lamp microscopic set of images, clinicians grade the stages of cataract by comparing its appearance. This paper intends to research the execution and proficiency by utilizing Deep Convolutional Neural Network (DCNN) to recognize and graduate cataract naturally with the help of preprocessed cataract images automatically by extracting the higher-order features. This work will help the ophthalmologists and clinicians to classify and predict the cataract with less time and high accuracy.

**Keywords** Cataract detection · Cataract grading · Deep Convolutional Neural Networks · Feature Maps · Feature Extraction · Image color analysis

## 1 Introduction

A cataract is a very serious ophthalmic disease, and if not curbed at the right point of time can turn into blindness. According to the World Health Organization, cataract is the reason for 51% of world blindness, which represents about 20 million people (2010). A cataract is of three types: Nuclear Cataract, Posterior subcapsular and Cortical Cataract. There can be many causes for the cataract like overproduction of

---

P. Nithyakani · R. Kheerthana (✉) · A. Shrikrishna · S. S. Ganesan · A. Wadhwa  
SRM Institute of Science and Technology, Chennai, India  
e-mail: [kr4373@srmist.edu.in](mailto:kr4373@srmist.edu.in)

P. Nithyakani  
e-mail: [nithyakp@srmuniv.ac.in](mailto:nithyakp@srmuniv.ac.in)

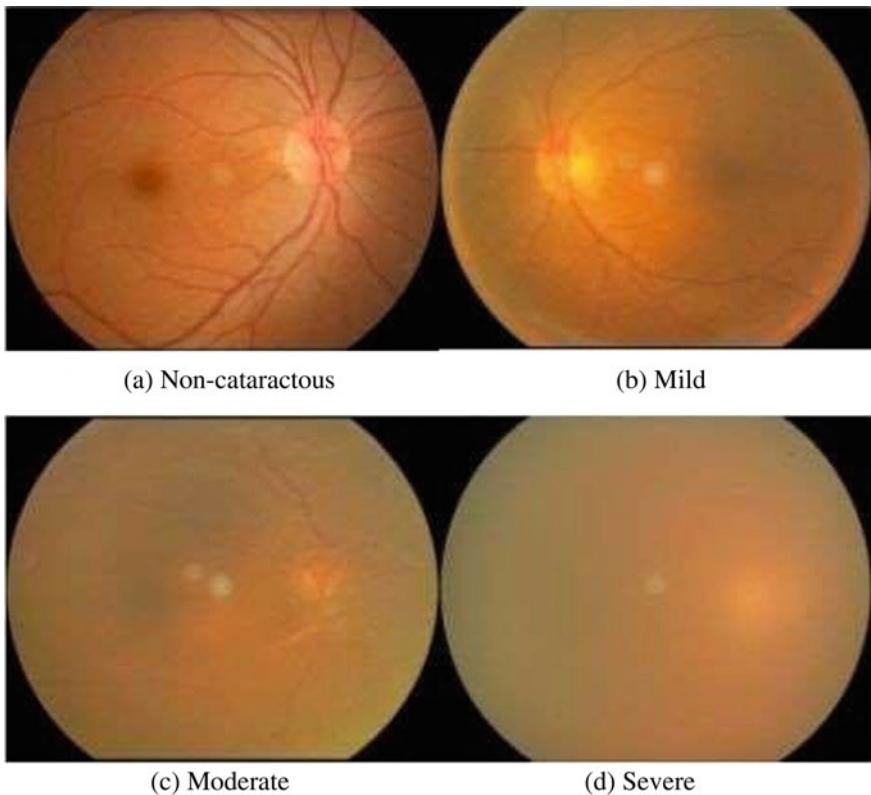
A. Shrikrishna  
e-mail: [sa2034@srmist.edu.in](mailto:sa2034@srmist.edu.in)

S. S. Ganesan  
e-mail: [ss5788@srmist.edu.in](mailto:ss5788@srmist.edu.in)

oxidants; smoking; ultraviolet radiation; etc. The risk factors of Cataract are older age; smoking; obesity; high blood pressure; heavy alcohol use; etc.

Various researches are going on the early detection of the cataract. Cataract detection and grading can be done by four methods: iris image projection; light focus method; slit-lamp examination; ophthalmoscopic transillumination. Prevention of cataracts is: wearing sunglasses that protect the eye from UVB rays, have regular eye exams and maintain diabetes and medical conditions. Surgery is the treatment for the cataract, various kind of surgeries are performed nowadays to cure cataract such as phacoemulsification removes the pieces using UV rays, extracapsular surgery removes the eye cloud portion. Usually, surgery to remove the cataract is safe. Symptoms of cataract are blurry vision; trouble seeing at night and seeing color in faded appearance. A cataract is categorized into four classes: Non-cataractous, Mild, Moderate and Severe.

Figure 1 represents the different levels of cataract disease in standard retinal fundus images. Non-cataractous (a) represent an eye with no cataract and we can see the optic disk and all the small and large blood vessels clearly. Mild (b) is that stage



**Fig. 1** Levels of cataracts

of cataract in which fewer blood vessels are visible. Moderate (c) large blood vessels and optic disk are only visible when the cataract was started. In Severe cataract (d) almost nothing is visible. Thus, we can predict and grade the image based on the optic disk and blood vessels in the retinal fundus images.

The paper is further organized as follows: Sect. 2 gives an idea about the previous papers published on this topic, Sect. 3 explains the System Architecture, Sect. 4 provides the Methodology used for classifying the cataract image based on Deep Convolutional Neural Network, Sect. 5 comprises of the experimental representations conducted and Sect. 6 gives the conclusion for the paper.

## 2 Related Work

Linglin Zhang et al. published a paper that studied the detection and grading task with supervised learning along with Deep Convolutional Neural Network (DCNN) [1]. This method has achieved accuracy and time efficiency in state of the art [1]. In DCNN classification, G-Filter is used to eliminate the local uneven illumination and eye reflection [1]. High level information is characterized and extracted efficiently [1]. Combination of feature extraction and classifier provide the higher level of intelligence [1]. This methodology has proven the need of the practical importance in early stage of cataract and its diagnosis and this can be applied to any eye diseases [1]. A paper by Xiyang Liu et al. extracts the features in depth using the framework of CAD for classification and grading of slit-lamp images [2]. Initially the images of pediatric cataracts are analyzed for the complexity and grading was applied in terms of morphology [2]. Region of interest was identified with candy detection algorithm and though transformation and ROI was given as an input the convolutional neural network for further identification [2]. The overall performance of this methodology has been significantly better than the traditional feature extraction methods according to the quantitative measures results [2]. This has been used for the software development to analyze the patients with eye problems [2].

Amol B. Jagadale and D. V. Jadhav proposed the detection of cataract using the circular though transform technique to extract the features which correlates the region of interest from the pupil effectively [3]. This method was helpful in detecting the cataract in early stage and its severity with the percentage and provides the best grading for cataract recovery [3].

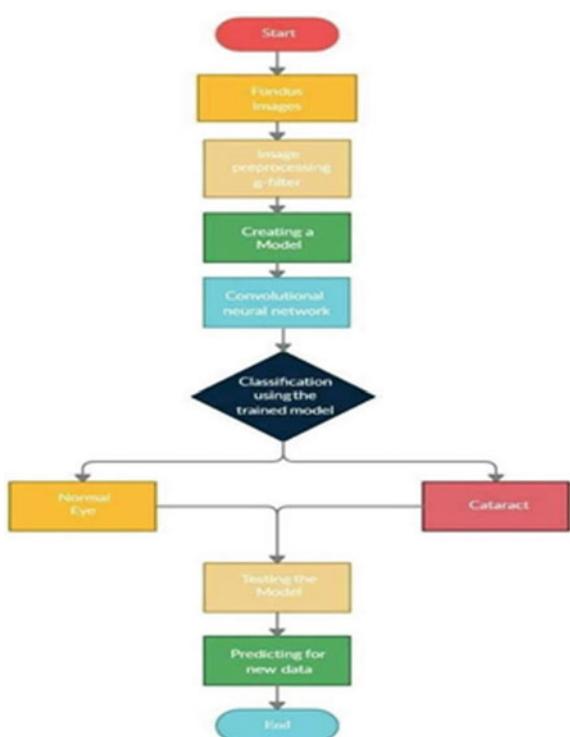
## 3 System Architecture

The model takes a retinal fundus image as input. Batch Normalization technique trains neural networks that standardize the input images to a layer batch-wise thus stabilizing the method of learning. This decreases the number of training epochs

that are required to train networks [4]. 2D Convolutional Layer creates a convolution kernel that coevolves with the layer input and produces a tensor of outputs. Max Pooling down-samples an input image to reduce its dimensions. It allows making assumptions about features contained within the binned sub-regions [5]. The featured map obtained is pooled then flattened. The pooled feature map matrix is transformed into one column by flattening which is further fed to the neural network for processing. In this, the images that are taken as the dataset is not even in size so all the images have been changed to the same size.

The undistributed illumination of the fundus image is being removed. We eliminate this uneven illumination because it was causing difficulty in detecting and grading the cataract precisely [6–10]. The image is being processed by the subsequent layers of convolutional neural networks. After preprocessing the fundus image, the data image is obtained. This image is used for creating the model using CNN and then classification of the trained model. This results whether the image is the one with cataract eye or not. Then prediction for new data is being done iterative. Figure 2 depicts the proposed architecture of cataract diagnosis.

**Fig. 2** Proposed architecture



## 4 Methodology

Cataract can be classified from the four components: Preprocessing; Deep Convolutional Neural Network which further includes feature extraction and feature selection and classifier.

### a. Preprocessing

In this method of preprocessing, fundus image is unified and the patient's personal information is been erased. The fundus image has been captured with a size of 3080\*2464 pixels using MATLAB functions such as interception of ellipse and rectangle cutting [11]. The undistributed illumination of the fundus image is being removed. We eliminated this uneven illumination because it was causing difficulty in detecting and grading the cataract precisely. To obtain the green channel image, fundus image which is in RGB color format has to be converted to Green color format. Figure 3 represents the change of fundus image to R-channel, G- channel and B- channel. In this method of preprocessing, fundus image is unified and the patient's personal information is been erased. The fundus image has been captured with a size of 3080\*2464 pixels using MATLAB functions such as interception of ellipse and rectangle cutting. The undistributed illumination of the fundus image is being removed. We eliminated this uneven illumination because it was causing difficulty in detecting and grading the cataract precisely.

To obtain the green channel image, fundus image which is in RGB color format has to be converted to Green color format. Figure 3 represents the change of fundus image to R- channel, G- channel and B- channel.

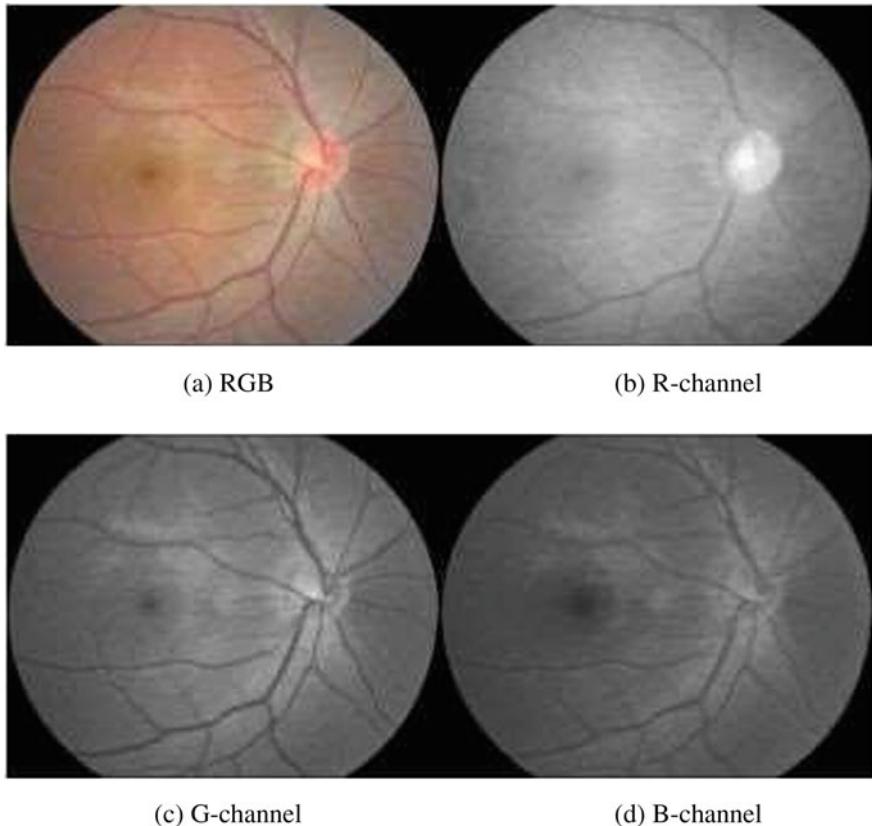
### b. Deep Convolutional Neural Networks

A convolutional neural network is a type of artificial neural network which is used widely in image recognition and processing that is designed to process pixel data. It performs both generative and descriptive task that incorporates image and video recognition, along with recommender system and natural language process (NLP) [12, 13].

A CNN uses a multilayer perceptron that has been designed for reduced processing requirements. The evacuation of limitations and increase in efficiency for image processing results in a system that is more effective, simpler to train limited for image processing and natural language processing.

Deep Convolutional Neural Networks is the predominant type of convolutional neural networks used for multidimensional signal processing. In general, one stage of CNN consists of three volumes: input maps, feature maps and pooled feature maps. All maps within a volume are of the same size. A map is a 2D array whose size varies from volume to volume. Each stage of CNN is convolution, from which the neural nets drive their names.

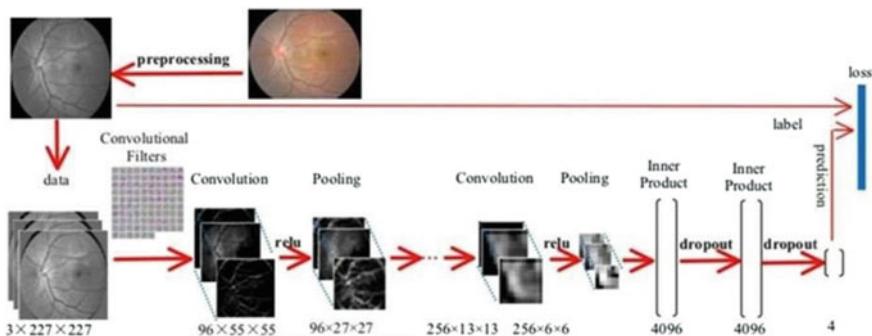
We have used a total of eight layers for our problem, in which initial five are convolutional layers and the remaining three are fully connected layers. The output of the last fully connected layer is passed to a four-way SoftMax. Each layer has



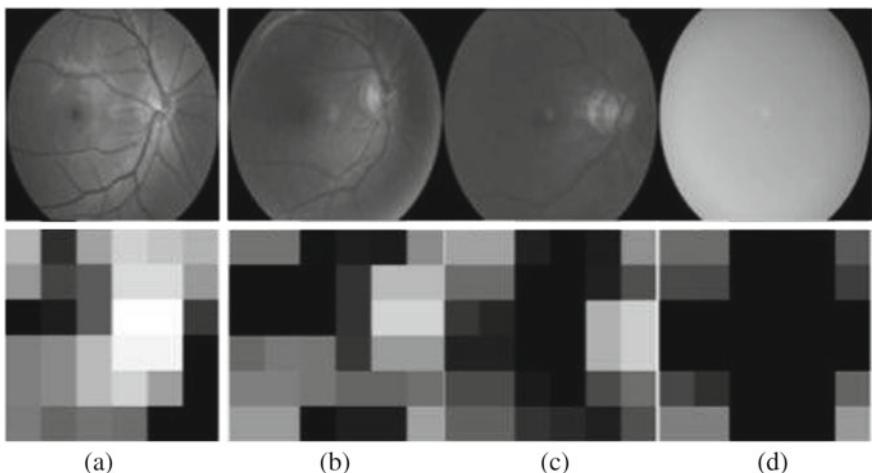
**Fig. 3** Representation of fundus image in R-channel, G-channel, B-channel

several 2D feature maps. Feature map often captures the specific aspect of images such as color, object category or attributes. After preprocessing the fundus image, the data images are obtained. The convolutional filters are applied and the convoluted images that are obtained are further rectified which is known as pooling. Pool5 layer consists of 256 feature maps where the resolution of each feature map is  $6 \times 6$ . Figure 4 describes the Deep Convolutional Neural Network-based classifier for cataract detection and grading. The image that we receive from pooling is further analyzed to the inner product and thus after dropout prediction is made out.

Figure 5 represents the semantic information taken by the feature map of the pool5 layers using Deep Convolutional Neural Networks. It represents the G-channel retinal fundus image of different classes. The high activation cells localize the blood vessels. The contrast between the blood vessels and the background is expressed by high activation cells. The stronger contrast between blood vessels and background is represented by a higher pixel value of high activation cells.



**Fig. 4** Deep convolution neural network



**Fig. 5** Semantic information of pool5 layers

Change in the view of the fundus image will not affect the visibility of the count of blood vessels and the appearance of the background. So, cataract detection and grading can be done with the help of feature representation derived from Deep Convolutional Neural Networks.

**Table 1** Confusion matrix

Confusion matrix		Predicted	
		cataract	Non_cataractous
Actual	cataract	TP	FN
	Non_cataractous	FP	TN

## 5 Experiments and Results

### 5.1 Setup of Experiments

In this, we will discuss the Database, Evaluation Criteria and Implementation. We got the DRIONS database which is an open-source dataset. For cataract detection and grading, we used formulas for Accuracy, Sensitivity and Specificity Table 1 represents the confusion matrix. Accuracy judges the whole sample in which it checks both whether the positive sample is positive and negative sample is negative. Sensitivity only checks if the positive sample is positive sample whereas Specificity checks if the negative sample is negative sample

$$\text{Accuracy} = (\text{TruePositive} + \text{TrueNegative}) / \left( \begin{array}{l} \text{TruePositive} + \text{TrueNegative} \\ +\text{FalsePositive} + \text{FalseNegative} \end{array} \right)$$

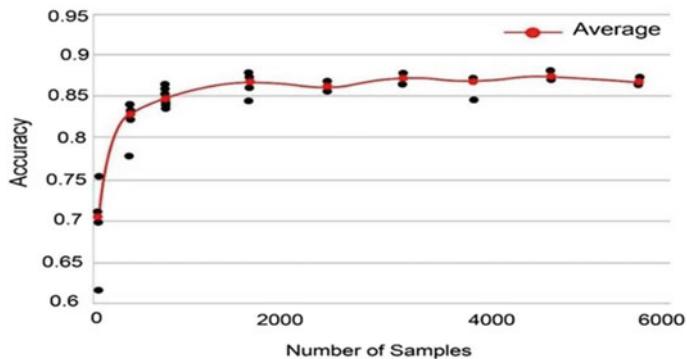
$$\text{Sensitive} = \text{TruePositive} / (\text{TruePositive} + \text{FalseNegative})$$

$$\text{Specificity} = \text{TrueNegative} / (\text{FalsePositive} + \text{TrueNegative})$$
(1)

The hardware configuration setup used for classifying retina fundus image with DCNN is 8 GB RAM, Intel Core i5 Processor and Windows 10 as the operating system. The fundus image was processed using the toolbox of MATLAB (image processing toolbox). The automatic classification system was implemented with help of Fast Feature Embedding Convolution Architecture.

### 5.2 Experimental Result and Analysis

The dataset has been divided into 80% of training data and 20% of testing data. The database name consists of two types which are the Testing Database and the Training Database. In the testing database first, the cataract percentage is around 89.86 and the non-cataractous percentage is around 87.52, hence the accuracy obtained in this Testing database is 89.52 and in the Training database where 20% of the data is only used, the cataract percentage is around 94.54 and the non-cataractous percentage is around 92.03, thus the accuracy obtained in this Training database is 93.02 which is higher compared to the testing database.



**Fig. 6** Performance of cataract diagnosis

Next, we again move back to the Testing database, here the specificity can also be termed SE and the sensitivity can also be termed as Sp. The specificity percentage of the non-cataract fundus is 92.53 and the sensitivity percentage is 72.25, the SE percentage of Mild is 79.21 and the SP percentage is around 88.84. The Specificity percentage and the Sensitivity percentage of the Moderate are 43.25 and 86.60, respectively. The Severe percentage of specificity is 74.28 and the severe percentage of sensitivity is 81.64. Hence, the overall accuracy which has been obtained in this database is around 83.65.

In the Training database, the specificity percentage and sensitive percentage of non- cataract Fundus is 97.32 and 75.98, respectively. The mild percentage of specificity is 85.38 and the mild percentage of sensitivity is 88.73. The specificity percentage of Moderate is 55.89 and the sensitivity percentage of moderate is 91.92. The SE percentage of Severe is 83.62 and the SP percentage is 86.54. Thus, the overall accuracy which is obtained in the Training database is 85.86 which is higher than the accuracy of the testing database. Figure 6 represents the performance of cataract diagnosis.

## 6 Conclusion

This paper describes the cataract detection and grading using Deep Convolutional Neural Networks was built in such a way that, convolution was done in first five layers and fully connection of network was performed in next three layers. As a result, cataract was detected and graded using the G-channel fundus image, which gives the differentiation of the blood vessels visibility and its background. The feature extraction process was combined with the classification. Thus, this approach is very useful in detecting and grading cataract at an early stage and can be used to classify other eye diseases.

## References

1. L. Zhang, Automatic cataract detection and grading using deep convolutional neural network, in *2017 IEEE 14th International Conference on Networking, Sensing and Control (ICNSC)*, pp. 60–65 (2017)
2. X. Liu, J. Jiang, K. Zhang, (2017). <https://doi.org/10.1371/journal.pone.0168606>
3. A.B. Jagadale, D.V. Jadhav, Early detection and categorization of cataract using slit-lamp images by hough circular transform, in *2016 International Conference on Communication and Signal Processing (ICCP)*, pp. 232–0235 (2016)
4. Y.Q. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, Caffe: Convolutional architecture for fast feature embedding, in *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 675–678 (2014)
5. D.C. Ciresan, U. Meier, J. Masci, L.M. Gambardella, J. Schmidhuber, Flexible, high performance convolutional neural networks for image classification, in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, vol. 22, pp. 1237–1242 (2011)
6. A. Krizhevsky, I. Sutskever, E. Hinton, ImageNet classification with deep convolutional neural networks. *Adv. Neural Inform. Process. Syst.* 1097–1105 (2012)
7. M. Anbarasan, B. Muthu, C. Sivaparthipan, R. Sundarasekar, S. Kadry, S. Krishnamoorthy, A.A. Dasel, Detection of flood disaster system based on IoT, big data and convolutional deep neural network. *Comput. Commun.* **150**, 150–157 (2020). <https://doi.org/10.1016/j.comcom.2019.11.022>
8. D. Vu, T. Nguyen, T.V. Nguyen, T.N. Nguyen, F. Massacci, P.H. Phung, A convolutional transformation network for malware classification, in *2019 6th NAFOSTED Conference on Information and Computer Science (NICS)*, (Hanoi, Vietnam, 2019), pp. 234–239. <https://doi.org/10.1109/NICS48868.2019.9023876>
9. W.M. Fan, R.F. Shen, Q.Y. Zhang, J.J. Yang, J.Q. Li, Principal component analysis based cataract grading and classification, in *2015 IEEE 17th International Conference on E- Health Networking*, pp. 459–462 (2015)
10. L.Y. Guo, J.J. Yang, L.H. Peng, J.Q. Li, Q.F. Liang, A computer-aided healthcare system for cataract classification and grading based on fundus image analysis. *Comput. Ind.* **69**, 72–80 (2015)
11. L.C. Huang, H.C. Chu, C.Y. Lien, C.H. Hsiao, T. Kao, Privacy preservation and information security protection for patients' portable electronic health records. *Comput. Biol. Med.* **39**, 743–750 (2009)
12. Y. Liang, L. He, C. Fan, F. Wang, W. Li, Preprocessing study of retinal image based on component extraction, in *Proceedings of 2008 IEEE International Symposium on IT in Medicine and Education*, pp. 670–672 (2008)
13. H.Q. Li, J.H. Lim, J. Liu, P. Mitchell, A.G. Tan, J.J. Wang, A computer- aided diagnosis system of nuclear cataract. *IEEE Trans. Biomed. Eng.* **57**, 1690–1698 (2010)

# A Technical Review and Framework Design for Influence Extraction from Social Networks



Akash Saini and K. Sornalakshmi

**Abstract** Now-a-days, social networks are becoming more than just a network, as it is used to identify how information flows among the nodes or how people communicate. This results in the huge increase in data which can be used to depict people behavior among one another. It has been an interesting practice to mine information from these networks but in a formalized manner. Analyzing such network can result in interesting facts about people's behavior and also some facts that are not observable normally. In social networks nodes or entities sometimes get influenced by each other for various reasons. For example, in social circles, people many times get influenced by others lifestyle. In this paper, we aim to survey the existing techniques to measure and visualize huge networks and find expert nodes using algorithms such as PageRank. Here, we will be modeling the topic based social influence on huge networks to differentiate between influencing nodes based on different topics. The idea here is to design a framework for developing a high-performance implementation of the graph algorithms for influencing node extraction.

## 1 Introduction

With the growth and fast popularity of these social techniques and social media applications, like instant messengers, data sharing sites, blogs sites, wikis, micro bloggers, social networks, collaboration networks and many others, but there is small thing that we are not certain of is that social influencing is slowly becoming a widespread, complex and understated task that directs the undercurrents of all social networks. Therefore, there is a very clear requirement for tools and techniques to analyze and measure the social influence for each node.

---

A. Saini · K. Sornalakshmi (✉)

Department of Data Science and Business Systems, School of Computing, Faculty of Engineering and Technology, SRM Institute of Science and Technology, SRM Nagar, Kattankulathur, Kanchipuram, Chennai, TN 603 203, India

e-mail: [sornalak@srmist.edu.in](mailto:sornalak@srmist.edu.in)

Social network analytics mainly aim to capture the overall behavior of all nodes in the network and such behavior is very crucial to identify because these information are not easy to identify and once identified can be very helpful in understanding how network operates.

Social network analysis frequently focuses on micro models like degree measures, scores, small world effects, preferential attachments, clustering coefficient, communalities, etc. Recently, social influence analytics study has started to gather additional and more consideration due to many significant and high-level applications. Though, major works in this zone present some of the more important and qualitative findings about social influences.

Social networks analysis suggests that one person can influence many numbers of people through their work or by their lifestyle, social network analysis aims to capture this influence in networks and try to model them in a way which can be quantified in a number of ways which can be used to implement on various applications. Suppose a person in a group of people have some skills, so it is likely that everyone else in the group may try to adapt those skills; this is called social influencing. Social influencing is becoming an important thing in the world as it can be used in marketing of various products.

Due to this reason only, social networking is becoming popular nowadays. So, if we know that a particular person is popular among the crowd for any skills or their or any other aspect so it is easy for him to endorse any product or service related to his skills as lot of people already follows him. This is a new way of marketing at getting more and more popular every day.

## 2 Influential Node Finding in Social Networks

The objective of social influence investigation is to infer the theme based social influences dependent on the info organization and point conveyance on every hub. First, we present some phrasing, and afterward characterize the social influence investigation issue.

**Topic distribution:** In social groups, a client generally has interests on numerous points. Officially, every hub is related with a vector of T-dimensional theme dissemination. Every component is the probability (rank) of the hub on a subject.

**Topic based social influences:** Topic based social influence implies how much a hub influence comparing hub dependent on a specific subject. Assume a hub A influence hub B with certain number, it isn't fundamental that hub B additionally influence hub A with same number.

The main problem is finding the expert node from the whole network that is based on a particular topic. Suppose we want to find expert node from a network than we can simply use any algorithm to find it. But if a node we need to find should be from a particular group then in that case first we need to identify the groups and then the expert nodes. So, finding these groups and then parallelly finding the expert nodes is a trivial task.

**Table 1** Social network analysis techniques

Technique	Purpose	Importance
Latent Dirichlet allocation	To find important word from given text file	It works based on the probability of occurrence of word
Affinity propagation	To find out group or clusters from big networks	It finds clusters or group through message passing technique
Map-reduce	Perform parallel execution in big data	There are no better framework for parallel execution to save complexity
PageRank	To find out important node from network	It finds expert nodes based on their relationship with other nodes

The issue here is that finding a master hub from an organization is certifiably not a troublesome undertaking; however, finding a hub dependent on the specific point is troublesome as their high-multifaceted nature preparing is included dependent on the strategies accessible.

The popular techniques used in social networks information extraction are summarized in Table 1.

### 3 Social Networks Overview

In the paper [1], they study social influence analysis in social networking big data from three alternate points of view: assets of social impact, connection among social influencers and big data, and engineering of social influence analysis. They have also discussed about the opportunities in social networks and also discussed some of the underlying problems. With the current increasing trend of big data and the data itself, many new opportunities are yet to be discovered as many industries and commercial will take up this technology. As we will continue with the in depth work in this field many new challenges will be faced in future research. In this article, they have contributed toward the theoretical advances of social networking techniques.

In the paper [2], they have examined a novel issue of topic based social influence investigation. Here, they suggest a Topical Affinity Propagation (TAP), a new way to describe the data using graphic probabilistic prototype. To manage the proficient issue, they extant another way for preparing the TFG model. A distributed learning framework has been executed under the map-reduce software design model. Trial results on three unique sorts of datasets exhibit that the planned approach can adequately find the topic based social impacts. The distributed learning calculation likewise had a decent execution time. We analyze the projected way to deal with master node finding. Analyzes demonstrate that the founded topic based impacts by the given approach can increase the performance of master node finding.

In paper [3], the OSN fame has given a one-of-a-kind occasion to examining and understanding the social collaboration and correspondence among its clients. They examined the state-of-art important client identification algorithm in OSNs. The current approval approaches utilized for assessing the performance of various influential clients' recognizable proof calculations are additionally checked on. Moreover, they introduced taxonomy of influential user detection in this paper and also mentioned some of the challenges related to it. Future work can include finding the solution of the challenges specified in this paper.

In the paper [4], they ordered and examined different automatic techniques for social network extraction. These strategies have an edge over the manual techniques where removing credits included a ton of communication with the profile owner, e.g., surveys and meetings. They additionally proposed an overall structure that can be utilized for building a social network extraction framework. But in many cases, a few difficulties like data sampling, mix of various kinds of web mining strategies, data portrayal on the web are still to be addressed appropriately. It is important to plan effective approaches and methods for information extraction from the web since it gets hard for end clients to discover useful information as a result of various issues. Data representation is one of them.

## 4 Important Techniques in Social Networks Analysis

In this paper [5], authors research the accompanying neighbors concerning social networks: how to misuse our uncommon abundance of information and how we can dig social networks for various purposes, for example, promoting work; social network as a specific type of influence, i.e., the way that individuals concede to wording and this present marvel's suggestions for the manner in which we manufacture ontologies and the semantic web; social organizations as something we can find from information; the utilization of social organization data to offer an abundance of new applications, for example, better proposals for eateries, reliable email senders, or (possibly) arranged meetings; examination of the lavishness and trouble of collecting FOAF (companion of-a-companion) data; and by taking a gander at how data handling is bound to social setting, the subsequent ways that network geography's definition decides its results.

The paper [6] they provide a comparison of their algorithm with the well-known PageRank algorithm. PageRank algorithm ranks webpages based on the importance of the webpage they are directly linked to, basically it is a variant of eigenvector centrality and is different than closeness centrality. And in PageRank algorithm as vertices linked to are calculated which requires to calculate PageRank for all the vertices in the network even if only top-k are required. However, for closeness centrality, their algorithm does not demand to calculate closeness for all the nodes which reduces the time complexity to great level.

In the paper [7], they introduced an enormous scope estimation study for the disconnected device-to-device (D2D) content in Mobile Social Networks (MSNs)

with information gathered from Xender. They did complete investigation from the viewpoint of social networks including basic properties of sharing clusters, qualities of social diagrams, theme elements, and course trees. Furthermore, we exceptionally looked at and examined enormous, medium and blended groups adding to a bunch of significant results and rules, which could be of extraordinary use to improve the nature of Xender like D2D sharing services. Moreover, the estimation results and techniques have promising exploration incentive to advance future investigations on the examination and mining of disconnected device-to-device datasets with promising exploration values.

In the paper [8–10], they proposed a novel calculation to gain proficiency with the distilled chart and select the best agent that includes at the same time for ghastly grouping on social network information. In particular, we iteratively locate the most delegate feature subset w.r.t. the diagram and afterward update the chart by utilizing the same feature. We contrasted our algorithm with other state-of-art on both engineered and genuine informational datasets, and the trial results showed the better viability and proficiency as compared to baseline. The proposed technique can be additionally stretched out to gain proficiency with the diagrams in semi-supervised learning and supervised learning settings. The proposed calculation is centered around the situation of social network information investigation, which is a normal supervised learning setting. Be that as it may, the proposed strategy can be stretched out to chart based semi-supervised learning and supervised learning situations.

In the paper [11], the authors have trailed an alternate track by offering another strategy ComTector for the community detection in enormous scope social networks. Straightforwardly based on the similar nature of the communities in real world, ComTector can collect significant outcomes on network whose network edifices are known previously. The technique comprises of three basic advances. For the initial step, the implementation embraces an altogether effective calculation to count all maximal cliques in the given network. The cover of a few maximal factions compares to different connections and associations every individual may take an interest in straightforwardly. The social affair of maximal cliques frames the bits of each expected network.

For the subsequent advance, they utilize an agglomerative strategy to iteratively enhance the left vertices to their nearest neighbor's dependency on the overall degree matrix. For the third stage, the initially observed clustering results will be changed by combining sets of fragmentary networks to accomplish a superior network modularity. The final obtained network structures along with different segments comprise a definitive segment of the network.

According to paper [12] as consistently increasingly, a greater number of individuals are joining the social networks so discovery of influential nodes in such a network is definitely not a simple undertaking. Subsequently, this paper proposed a method to recognize the most powerful node that depend on network measures that incorporates degree, betweenness centrality, closeness centrality, eigenvector centrality, page rank and clustering coefficient. The proposed strategy has been tried utilizing an experimental data. The trial results and an itemized quantitative examination show

that this proposed method is more productive path for recognizing influential nodes in an enormous social network.

## 5 Topic Modeling Techniques

In the paper [13], they experience an information assortment with both abundant textual data and a network structure. Statistical topic models extricate intelligent topics from the content, while generally overlooking the organization structure. Social network examination will in general focus topologic network structure, while leaving aside the textual data. In this work, we officially define the significant tasks of topic modeling along with network structure. They suggest an overall arrangement of text mining with network structure, which upgrades the probability of topic generation and the topic perfection on the chart in a unified way. They suggest a regulation system for measurable topic models, by a symphonious regularizer dependent on the network structure. The overall structure permits discretionary decisions of the topic model and the chart based regularizer. They show that with solid decisions, and the model can be functional to handle real content mining issues, for example, creator topic examination, effective network revelation, and spatial subject investigation.

In this paper [14], we have portrayed Latent Dirichlet allocation, a flexible generative probabilistic model for assortments of discrete information. LDA depends on a basic exchangeability supposition for the words and points in a record; it is consequently acknowledged by a clear utilization of de Finetti's representation hypothesis. We can see LDA as a dimensionality decrease strategy, in the soul of LSI, yet with legitimate basic generative probabilistic semantics that bode well for the sort of information that it demonstrates.

In this paper [15], a favorable position of affinity propagation is that the update rules are deterministic, very straightforward, and can be inferred as an occurrence of the sum product algorithm in a factor diagram. Utilizing testing applications, they indicated that affinity propagation acquires better results (as far as percentile log-probability, visual nature of picture division and sensitivity-to-specificity) than different strategies, including K medoids, spectral grouping, Gaussian mixture modeling and progressive clustering.

As far as anyone is concerned, affinity propagation is the first algorithm to join points of interest of pair-wise clustering techniques that utilize bottom-up evidence and model-based techniques that try to fit top-down worldwide models to the information.

In this paper [16], we tackle the issue of finding thick sub graphs of the web-graphs. They propose an effective heuristic technique that is experimentally shown to be able to find about 80% of communities having around 20 fans/focuses, even at medium thickness (above half).

The adequacy increments and approaches 100% for bigger also, denser networks. For people group of under 20 fans/focuses (state 10 fans also, 10 focuses) our calculation is as yet ready to distinguish a sizable portion of the communities present (about

35%) at whatever point these are at any rate 75% thick. This technique is successful for a medium scope of network size/thickness which isn't all around distinguished by the current innovation. One can cover the entire range of networks by applying first our technique to identify huge and medium size networks, at that point, on the remaining diagram, the Trawling calculation to locate the more modest networks left. The effectiveness of the Trawling calculation is probably going to be supported by its application to a lingering chart cleaned of bigger networks that will in general be re-found a few times.

In this paper [17], they have defined another kind of web network that can be proficiently determined in a maximum flow framework. We have likewise presented a maximum flow framework web crawler that can estimated a network by coordinating an engaged web crawler along connect ways that are profoundly applicable. Their work found networks are extremely durable as in individuals from the network are more firmly coupled to one another than to non-individuals. The EM approach that we use steadily improves the crawl results by re-cultivating our crawler with profoundly applicable locales.

In this paper [18], they propose a unified labeling way to deal with specialist profiling. About a half million scientist profiles have been extricated into the framework. The framework has likewise coordinated more than 1,000,000 papers. We propose a probabilistic system to manage the name uncertainty issue in the coordination. We further propose a unified topic model simultaneously model the various sorts of data in the academic network. The displaying results have been applied to skill search and affiliation search. We lead tests for assessing every one of the proposed approaches. Exploratory outcomes show that the proposed techniques can accomplish better performance. There are numerous likely future headings of this work. It is intriguing to additionally explore new extraction models for improving the exactness of profile extraction. It would be likewise intriguing to research how to decide the genuine individual number  $k$  for name disambiguation. At present, the number is provided physically, which isn't reasonable for all creator names. Also, broadening the topic model with connect data (e.g., reference data) or time data is a promising heading.

## 6 Proposed Approach

In social network, we know that there are number of people and each has their own skills or interest, and they may influence others or maybe get influenced by others on various topics. Our aim here is to find people in a group which have influence over other people, i.e., people which are most popular in network based on their skills or topics.

To find such people, various methods need to apply on the network. First, we need to identify how much a person influence another person on a particular topic. As we get the topic level information for all nodes in the network then we are ready to fetch

sub networks based on each topic. Each sub network will contain nodes which have certain level of influence on specific topic.

Once we have these sub-graphs, we are ready to fetch expert node from each graph which means we will have expert node for each topic. For a particular topic, we will have an expert node.

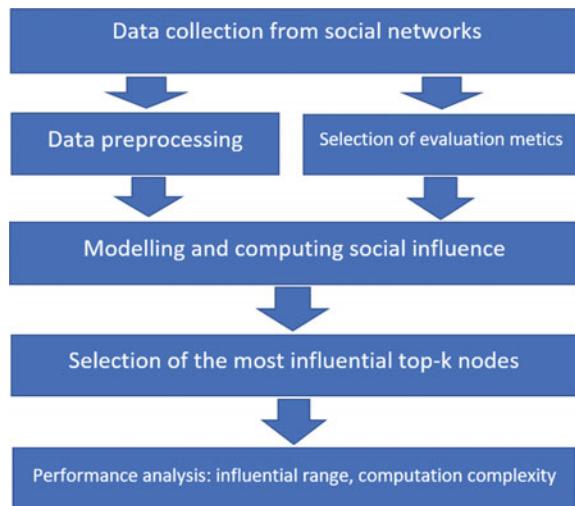
Now to extract topic level information for each node we will have to use some statistical modeling technique which will extract topic information from the data. For topic modeling various approach like Latent Dirichlet Allocation (LDA) is used.

Once topic information is fetched, next step is fetching sub-graph from whole network structure. To do this, we are going to use topical affinity propagation (TAP) which will take the whole network and topic level information as input and will return the sub graphs based on each network.

Now, we have  $n$  numbers of sub networks based on  $n$  numbers of topics, and now the part comes where we have to fetch most influential nodes from each of these sub graphs. To do this, we will use PageRank algorithm. Now if we apply PageRank one by one on each sub graph so the time taken will be very much so, to counter this problem we will apply PageRank in map-reduce manner so that all the sub graphs can be processed in parallel with the PageRank algorithm, and expert nodes can be fetched in one time.

Our approach will perform in a better way and will reduce the complexity and execution time for expert identification. map-reduce will help in reducing the execution time and PageRank along with influence scores will search expert nodes in an efficient way. This is shown in Fig. 1.

**Fig. 1** Process flow for influential node extraction



## 7 Conclusion and Future Work

Social networks play a major role in many domains relation to information processing today. The applications range from propagandas, marketing to more sophisticated applications like opinion influencing. The technique of influence extraction in social networks plays a key role in such applications. In this paper, we have surveyed the basic techniques used in social network analysis and the techniques used for influence modeling. The in depth analysis helps to understand the role of influencing nodes in social networks. In future, we plan to propose a hybrid model for influencing node extraction topic and location wise and the influence network using graph theory techniques.

## References

1. S. Peng, G. Wang, D. Xie, Social influence analysis in social networking big data: opportunities and challenges. *IEEE Network* **31**(1), 11–17 (2017). <https://doi.org/10.1109/MNET.2016.1500104NM>
2. M.A. Al-Garadi, K.D. Varathan, S D. Ravana, E. Ahmed, G. Mujtaba, M.U.S. Khan, S.U. Khan, Analysis of online social network connections for identification of influential users: survey and open research issues. *ACM Comput. Surv.* **51**, 1–37 (2018). <https://doi.org/10.1145/3155897>
3. T. Arif, R. Ali, M. Asger, Social network extraction: a review of automatic techniques. *Int. J. Comput. Appl.* **95**, 975–8887 (2014). <https://doi.org/10.5120/16558-3964>
4. S. Staab et al., Social networks applied. *IEEE Intell. Syst.* **20**(1), 80–93 (Jan–Feb 2005). <https://doi.org/10.1109/MIS.2005.16>
5. K. Okamoto, W. Chen, X.-Y. Li, Ranking of closeness centrality for large-scale social networks. *Frontiers Algorithmics* **5059**, 186–195 (2008). [https://doi.org/10.1007/978-3-540-69311-6\\_21](https://doi.org/10.1007/978-3-540-69311-6_21)
6. H. Wang, S. Wang, Y. Zhang, X. Wang, K. Li, T. Jiang, Measurement and analytics on social groups of device-to-device sharing in mobile social networks, in *2017 IEEE International Conference on Communications (ICC)*, (Paris, France, 2017), pp. 1–6. <https://doi.org/10.1109/ICC.2017.7997038>
7. W. Liu, D. Gong, M. Tan, J. Q. Shi, Y. Yang, A. G. Hauptmann, Learning distilled graph for large-scale social network data clustering. *IEEE Trans. Knowl. Data Eng.* **32**(7), 1393–1404, 1 July 2020. <https://doi.org/10.1109/TKDE.2019.2904068>
8. M. Balaanand, N. Karthikeyan, S. Karthik, Envisioning social media information for big data using big vision schemes in wireless environment. *Wireless Pers. Commun.* **109**(2), 777–796 (2019). <https://doi.org/10.1007/s11277-019-06590-w>
9. D.V. Pham, G.L. Nguyen, T.N. Nguyen, C.V. Pham, A.V. Nguyen, Multi-topic misinformation blocking with budget constraint on online social networks. *IEEE Access* **8**, 78879–78889 (2020). <https://doi.org/10.1109/ACCESS.2020.2989140>
10. N. du, B. Wu, X. Pei, B. Wang, L. Xu, Community detection in large-scale social networks. Joint Ninth WebKDD and First SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis. 16–25 (2007). <https://doi.org/10.1145/1348549.1348552>
11. Farooq, G. J. Joyia, M. Uzair and U. Akram, “Detection of influential nodes using social networks analysis based on network metrics,” 2018 International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), Sukkur, Pakistan, 2018, pp. 1–6, doi:<https://doi.org/10.1109/ICOMET.2018.8346372>.
12. Mei, Qiaozhu & Cai, Deng & Zhang, Duo & Zhai, ChengXiang. (2008). Topic modeling with network regularization. Proceeding of the 17th International Conference on World Wide Web 2008, WWW'08. 101–110. <https://doi.org/10.1145/1367497.1367512>.

13. D. Blei, A. Ng, M. Jordan, Latent Dirichlet Allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003). <https://doi.org/10.1162/jmlr.2003.3.4-5.993>
14. Frey, Brendan & Dueck, Delbert. (2005). Mixture Modeling by Affinity Propagation, 15. IEEE International Conference.
16. Dourisboure, Yon & Geraci, Filippo & Pellegrini, Marco. (2009). Extraction and Classification of Dense Implicit Communities in the Web Graph. TWEB. 3. <https://doi.org/10.1145/1513876.1513879>
17. Flake, Gary & Lawrence, Steve & Giles, C.. (2002). Efficient Identification of Web Communities. <https://doi.org/10.1145/347090.347121>.
18. Tang, Jie & Zhang, Jing & Yao, Limin & Li, Juanzi & Zhang, li & Su, Zhong. (2008). ArnetMiner: Extraction and Mining of Academic Social Networks. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 990-998. <https://doi.org/10.1145/1401890.1402008>

# Sentiment Diagnosis in Text Using Convolutional Neural Network



C. Sindhu, Shilpi Adak, and Soumya Celina Tigga

**Abstract** Web 2.0 has empowered people to voice their opinions, share their familiarities and views on various products, events, services, etc. Nowadays, people are turning toward social networking sites, forums, blogs and e-commerce websites to write opinions and reviews. The opinion mining process is very crucial to the recommender system which predicts product preferences of the user and voice of customer material as text can contain related features or aspects of the item the user's opinion about them. This process requires the use of natural language processing, text analysis, computational linguistics, biometrics and possibly machine learning for identifying the subjective and objective part of the text. Knowledge-based technique, hybrid, and statistical approaches are the major methods used for sentiment analysis. Depending on the type of text used and the type of result we desire, the methods will differ. Previously, methods like lexicons in corpus or dictionary-based approach, part of speech tagger (POS tagger), and even machine learning techniques have been used. Our work focuses on using a deep learning model of Convolution Neural Network (CNN) built on PyTorch. The model uses a vector model of GloVe and a dataset that contains reviews of various movies in two categories, namely plot and quotes.

**Keywords** Opinion mining · Sentiment analysis · Sentiment classification · Subjectivity analysis · Convolution neural network · Facts and the opinions

## 1 Introduction

In the past one decade, we have all witnessed an unprecedented increase in the online goings-on of people across the globe. Social media platforms have led to flooding of content on the internet, and they are no more just a platform where people talk to each other. People tend to share their opinions, views and experiences about various things with each other online. With the emergence of social media, millions of people

---

C. Sindhu (✉) · S. Adak · S. C. Tigga

Department of CSE, SRM Institute of Science and Technology, Kattankulathur, India

have started using this new tool to share their views. Nowadays, people can easily share their point of view by using micro blogging platforms like Twitter. People also have the liberty to share their reviews on the products they have purchased from various E Commerce websites. The data generated by the users can be useful for treating and reviewing the human behavior [1] and taking various decisions. The companies can know what their competitors are doing and plan better marketing strategies according to the responses they receive on social media. But the process is not easy. People can have varied opinions about the same product and can have different ways of expressing their opinions. Thus, it is very crucial to cleanse the reviews [2]. Due to the large amount of data available, it is humanly impossible to process it manually; to automate this tedious process, sentiment analysis came into play. Sentiment analysis deals with identifying subjective data and further analyzing the emotion attached to it: positive, neutral or negative [3].

Sentimental analysis is a two-step process [4]. The first step is opinion mining, followed by sentiment classification. The reviews can be divided into two parts: subjective and objective. Reviews that do not contain opinionated text comprise the objective part where reviews that contain personal point of views of people toward objects or entities makeup the subject of part. The subjective part can further be classified into positive, negative or neutral. According to Liu [5], an objective review provides factual information, while a subjective review has personal feelings, beliefs and point of views. The process of deciding whether a sentence is objective or subjective is called classification of subjectivity or opinion mining [6]. A sentence may contain several entities, but finding the individual to which the sentiment is directed is mandatory. It identifies the polarity (positive, negative or neutral) and the degree of the sentiment.

Because of the sheer volume of data, it is impossible to manually label the sentences as subjective or objective. Hence, sentiment analysis methods are used to classify different dimensions of sentiments in the reviews. Subjectivity classification is one of the most significant tasks of sentiment analysis. It is about classifying sentences into subjective and objective classes [7]. Subjective sentences show the writer's opinions, evaluations, emotions, beliefs, and judgments [7, 8].

There are various challenges in sentiment analysis. Some of them are discussed briefly below. Table 1 lists out few of the challenges and the papers in which they have been addressed, and the dataset that was used in the respective work.

**Table 1** Challenges addressed and dataset used in references

Challenges dataset paper
Subjectivity detection
Movie review, twitter data [9]
Word ambiguity Amazon reviews [10]
Entity identification [11]
Twitter data [12]
Spam detection Yelp and Amazon [13, 14]
Internet slangs American song (60 words) [15]

1. Subjectivity detection [4]: Subjective statements need to be separated from objective ones for pre-processing. Example: I bought a pair of Nike shoes. The shoes are amazing. Here, the first sentence is factual, whereas the second sentence has positive emotions attached to it.
2. Identifying the word meaning and removing the ambiguity [16]: One word can have multiple senses and it's very challenging to find out the context in which the word has been used. Example: The customer service of this bank is good. In this sentence, there arises an ambiguity whether the word bank refers to river bank or money bank.
3. Entity identification [17]: When there are multiple entities in a review, Example: Chris Craft is better looking than Limestone. There are two brand names in this example, and it is difficult to define the aim of the attitude.
4. Spam detection [17]: Examples: Competitors try to bring down each other by posting bad reviews from fake accounts. So it is very vital to filter out the spam reviews before analyzing the reviews.
5. Internet Slangs [8]: Example: The phone is really gud, and the camera quality is lit. Here, gud is written instead of good.

The aim of the current work is to propose a method for identifying subjective sentences for user opinions. This paper has been decomposed into four parts. Section two deals with the research work in the field of classification of subjectivity that has already been done [18]. Section three consists of the methodology. Section four comprises of the results and section five has the conclusion that provides the possible, and relevant areas that can be worked out further.

## 2 Related Work

The purpose of subjectivity classification is to distinguish between factual and opinionated responses present in customer reviews. This categorization can be done both at document and sentence levels [19]. To classify documents containing subjective texts from large collections of reviews for further processing, document level subjectivity classification is required. Mostly, supervised learning is used for subjectivity classification [5].

A subjectivity lexicon was created in [19] to distinguish between objective and subjective terms without marking the words as positive or negative; a genetic algorithm was introduced that is inspired by the natural selection process in which the fittest are chosen for reproduction in order to create offspring with better fitness. In genetic algorithm, solution is given in the form of chromosomes [20]. The fitness of chromosomes is calculated using a fitness function. Two datasets were used, Stanford Twitter Sentiment (STS) and Sanders that consisted of tweets people posted on Twitter. A tenfold cross-validation technique was used, and the lexicons were created based on the training dataset (ninefold) then tested on the test data (onefold) [21].

The method proposed in [22] uses POS tagger for every record along with WordNet and SentiWordNet [23–25]. SentiWordNet [26] is a lexical, opinion mining resource.

SentiWordNet assigns to each synset of WordNet [27] three sentiment numerical scores that describes how positive, negative and objective the synsets are. These three scores are in the range of 0.0 to 1.0 with the sum of each synset being 1.0.

In [28], a minimum cut algorithm was used for subjectivity detection. MinCut is a simple randomized algorithm to determine the minimum cut in a graph: a subset of vertices  $S$  in which the set of edges leaving  $S$ , denoted  $E(S, S)$  has minimum size among all subsets. The CutSenti algorithm is used to form two adjoint sets. In this algorithm instead of taking into consideration, the number of sentences between ( $s_i, s_j$ ) importance is given to the content of the sentences. The association value lies in the range 0.0–1.0. 1 means the two sentences are closely related (both are subjective or objective) and 0 means either of them is subjective and negative. It is observed that CutSenti outperforms MinCut.

In [29], an attempt was made via a Hidden Markov model to detect sentence-level subjectivity. The feature extraction algorithm calculates a feature vector based on the statistical occurrences of words without any lexical details except tokenization. Therefore, this model can be applied to any language, i.e., no lexical, grammatical, syntactic analysis in the classification process is used [30]. The model can therefore be extended to any language. A hidden Markov model allowing transition from one emission state to any other emission state is known as ergodic HMM whereas the model that allows transition from one state to the follower state is known as left right HMM. In movie reviews [Pang/Lee ACL 2004], a subjectivity dataset 1v.0 consisting of 5000 subjective and 5000 objective processed sentences were used [31].

A robust pre-processing method and an algorithm for extracting features from Reviews/Blogs is suggested in [32] and the proposed method is unmonitored, automated and independent of the domain. On a real life dataset that is collected from several reviewing websites such as CNET, Amazon, etc., the efficacy of the proposed solution is seen.

The deep learning approach in sentiment diagnosis became quite famous in previous decade [33]. It was due to ease in model designing, advancements in parallel computing hardware and formation of algorithms for neural networks with a large number of layers [34–36]. In order to follow the deep learning approach in sentiment analysis, we need to convert text into numeric values so that they can be given as input to neural networks. This can be done by capturing the relationship between the word and the context and positioning them in a  $n$  dimensional space. This relationship can be captured using dense vectors. There are several ways of coding words in dense vector format using word2vec [37] and GloVe [38]. Word embedding is a text representation technique where different terms have a common real valued vector representation with a similar context.

Another well-known model that learns vectors and phrases from their co-occurrence data is GlobalVector (GloVe), i.e., how often they appear together in large text businesses. GloVe is a model focused on counts, while word2vec is a predictive model that improves predictive ability. GloVe is basically a log bilinear model with a weighted minimum square objective [39, 40]. The model is based on a very simple concept that ratios of probabilities of word-word co-occurrence have

the potential to encode some type of meaning that can be encoded as vector differences. The training goal, therefore, is to learn word vectors in such a way that their dot product equals the logarithm of the probability of co-occurrence of the terms [41–45]. Almost all unsupervised methods for learning word representations use the statistics of word occurrences in a corpus as the primary source of information, but the question remains as to how we can derive meaning from the statistics and how the resulting vectors can represent that meaning. They are a key advancement that has led to the great success of models of the neural network on a number of difficult issues in the processing of natural language.

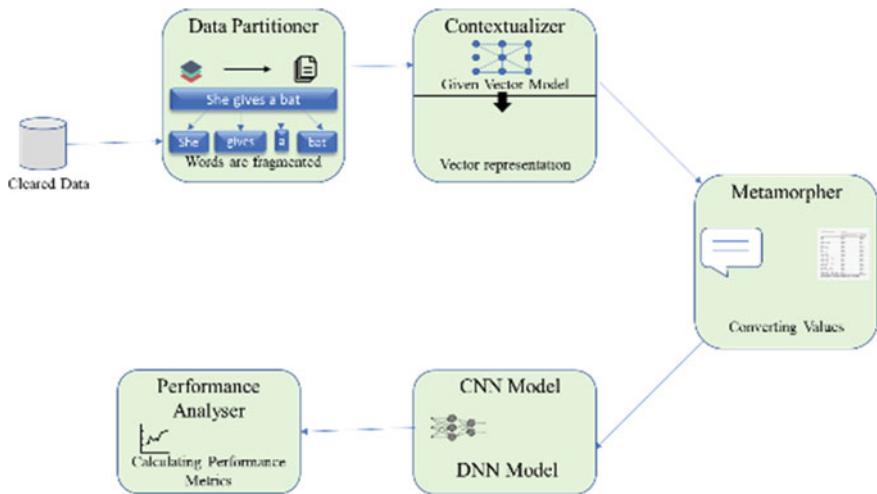
Convolutionary neural networks or CNNs form the basis of numerous modern computer vision systems. Tasks like classification of images, object recognition, semantic segmentation, etc., can be successfully handled by CNNs. It appears to be perplexing at first glance to use the same methodology for a task as distinct as sentiment diagnosis.

CNN (Convolution Neural Networks) is a class of deep neural networks, in deep learning which comes under artificial intelligence and machine learning as a subset. It is used for various purposes including data mining, image classification, etc. CNN or rather any neural network, essentially simulates the neural network of the human body [46]. These neural networks consist of multiple neuron layers which are completely connected to each other. Though, this makes them vulnerable to overfitting of data. CNN models can be trained to improve efficiency. But this improvement is limited to a certain extent. The framework of most CNN models include the input, output, and neuron layers which may or may not include hidden layers also. CNN model relies heavily on the use of tensors [47]. Now, the CNN has advanced so much that GPUs (Graphics Processing Units) are required for some implementations. CNN is the successor of the MLP (Multilayer Perceptron). The main benefit of using CNN over the previously popular MLP was the 3-dimensional volume of neurons. Other than this, the pooling ability, feature of sharing weights and also the local connectivity proved to be a plus.

### 3 Methodology

In our work, we will be using a pre-trained word embedding GloVe to get the word vectors. In order to obtain the word vectors, the co-occurrence of the words, semantic and syntactic similarity and context are taken into consideration. In this section, we discuss the architecture and functional details of the classification method used in our work. The given Fig. 1 presents the architecture of the proposed system including following modules: Data Partitioner, Contextualizer, Metamorpher, CNN model, and Performance Analyzer.

The following subsections address each of the components of the architecture in more detail.



**Fig. 1** Architecture diagram

### 3.1 Data Partitioner

We need to split the sentences into words and the output of the word tokenization is further analyzed. This module is used to separate the text data word by word. We have used the subjectivity dataset [48] that has 5000 subjective and 5000 objective processed reviews. Dividing the dataset is an excellent practice. We have divided the dataset into three parts: training, validating, and testing. The training dataset is used to teach the model about the parameters, to set hyper-parameters such as number of epochs or the learning rate we use the validation dataset and lastly, we measure the metric over the test dataset.

Each review in the dataset includes multiple words and sentences. This is the pre-processing module which prepares the data in a format which fits the requirements of the next module. We use the dataset to train the module using a sklearn model. Here, the text is split into words and stored in a vector separately. This is the first step of pre-processing.

### 3.2 Contextualizer

In this module, we use a GloVe vector module to remove the ambiguity of words and get their exact meaning by comparing it in the context of the sentences. A word can mean different things based on the context of the word in which it is used. For instance, "I didn't know the answer to the last question and I asked him again, but he didn't answer me," both these sentences use the word "answer." In both the sentences, the word holds different meanings; former is a noun and latter is a verb. This module

helps the model to identify the same. This GloVe vector holds different numerical values (float) to represent the different meanings of the words, known as weights. This process is known as embedding [9]. Word embedding is a method of modeling that portrays words or phrases as vectors. For words with a common meaning, the words are grouped together to achieve a similar representation. The word embedding learns the relationship between the terms for the representation to be built. Different strategies such as co-occurrence matrix, probabilistic modeling, are accomplished by neural networks. Word embedding can be classified into two types: Frequency based embedding and pre-trained word embedding. Count vector, Co-occurrence vector, HashingVectorizer, TF-IDF are examples of frequency based embedding and Word2Vec, GloVe, BERT, fastText are examples of pre-trained word embedding.

GloVe's key theme is to capture the meaning of a word that incorporates the whole corpus with its overall structure and to derive relationship between them from global statistics. Using the word frequency and co-occurrence counts of terms, most of the unsupervised algorithms are educated. By training a model based on global co-occurrence counts of terms, global statistics and using mean square errors as the loss function, GloVe generates word embedding [10]. Word relationships and similarities are maintained by the created word integration with such a model. For a given sentence, a co-occurrence matrix tells us how often a given pair of words appear together. In the matrix, each element is the count of the pair of words that occur together.

In our work, we use the glove.6B.300d.txt, which contains 300d vectors for use. This contains a dictionary for the words in vectors. We also add an extra step for removing words which are not occurring in the text before the embedding process, to filter the input data. This is done by counting the occurrences and finding the minimum occurrence word in the dictionary.

### ***3.3 Metamorpher***

Metamorpher is used to convert the word embeddings into numerical vectors. Here, we translate the values to data that the machine can understand. The text data is not understood by the computer so we convert it into fixed size numerical vectors using stacks. This process is known as encoding. The data is stored as an encoded and padded vector. After this module, the data is ready for the CNN model. In our work, we have used a vocab2index() function to encode the embeddings with a uniform size of 40. This is the final processing done before feeding the data in the CNN model.

### ***3.4 Convolution Model***

This module is a 1D CNN model which includes 3 parameters, namely D (Dictionary size), C (Contextualizer size), and M (Max sentence length). The CNN class includes

the glove weights, embeddings, ReLU (Rectified Linear Unit) layer, pooling layer, and the convolution layer. The convolution layer is the basic building block of the CNN model. This layer has kernels. In our work, we use kernels of size 3, 4, and 5. The ReLU layer is used to eliminate negative values from the activation map by simply setting them to zero. This is an effective way of increasing the nonlinear properties of the decision function of the CNN model. In our work, we use three layers of ReLU, one for each convolution layer. The fully connected layer is a high-level reasoning layer.

### 3.5 Performance Analyser

Some of the most common factors used to analyze the performance of opinion mining models are accuracy, recall, precision, and  $f$  score or  $f1$  measure. In our work, we have used (1), (2) and (3) to compute the performance metric for our model.

$$\text{Epsilon} = 1e - 7$$

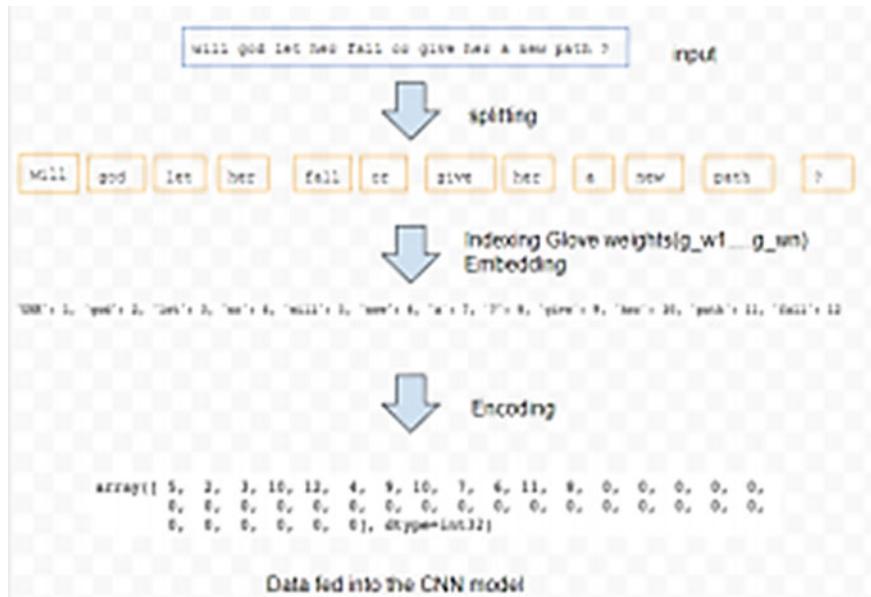
$$\text{Precision} = \text{tp} / (\text{tp} + \text{fp} + \text{epsilon}) \quad (1)$$

$$\text{Recall} = \text{tp} / (\text{tp} + \text{fn} + \text{epsilon}) \quad (2)$$

$$F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall} + \text{Epsilon}) \quad (3)$$

### 3.6 Summary

As we have discussed, the classification of the opinions in a text using CNN is a complex process. The CNN model itself includes many layers for processing the data along with the pre-processing and performance metric processes. The complete process includes reading the input text, splitting the words individually, removing the rare occurrences, using glove embeddings, encoding the data, feeding the data to the CNN model layers after training the model and finally the analysis of the results. Figure 2 shows the flow of how the sentences in the review dataset are tokenized and embedded using GloVe. The word embeddings are then encoded and fed as input to the CNN model.



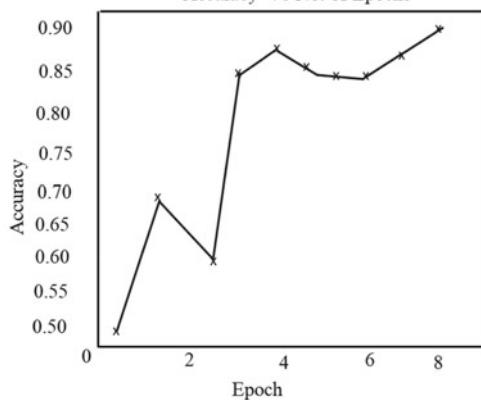
**Fig. 2** Processing of input for CNN model

## 4 Results

The experiment carried out in our work is conducted on the movie review dataset. The evaluation for the model is done with the performance metric accuracy. The loss and accuracy are calculated for both training and test datasets. We use epoch groups of 10 and train it 10 times. As we train the model, the accuracy increases and reaches a saturation point. We also plot an accuracy graph of the training process. The graph shows the accuracy for the epoch sets. The graph for first 10 epochs and last 100 epochs is shown in Fig. 3 and Fig. 4, respectively, where the saturation point is achieved. Here, we see that the maximum accuracy achieved for the training dataset is 0.921.

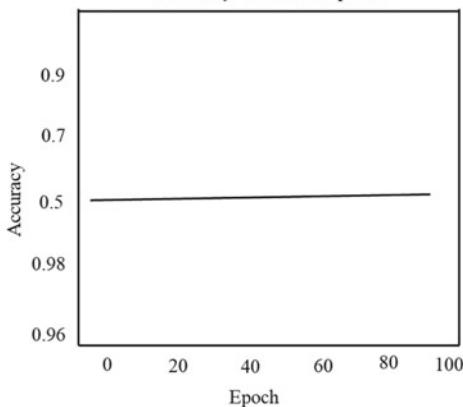
Figures. 3 and 4 show that even though the model has some deviation, and it still manages to achieve an accuracy of 0.907. Table 2 shows the results obtained by various models and our model. The formulae used to calculate the precision, recall and F1 score is stated in sub-Sect. 3.5. In Fig. 3, we have plotted the precision, recall and accuracy of all the models listed in Table 2 which are referenced from. We can see from the table that the accuracy of our model is higher than that of other models. Including multiple layers helps the CNN model to function at a more efficient level since the number of links in the model increases (Fig. 5).

Validation Set = {‘loss’, ‘0.3399101458464831’, ‘accuracy’, ‘0.8690000876429149’}



**Fig. 3** Accuracy graph for training dataset

Validation Set = {‘loss’, ‘0.32156263488532747’, ‘accuracy’, ‘0.9210336038146973’}



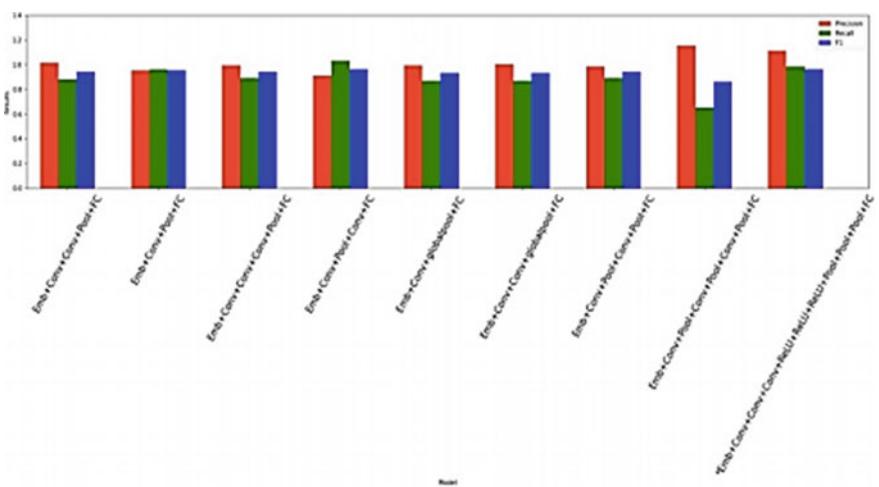
**Fig. 4** Accuracy graph for test dataset

## 5 Conclusion

In this paper, we survey the features of the classification process which is used to identify the sentiments in a text data, and the methodology used to do the same in the past. The most significant feature of 5any Natural Language processing model is that more often than not, they are specific to a language. Every language has its own vocabulary, grammar, and orientation. These NLP models require a corpus, dictionary, and other tools that are available for that language. Models that use an algorithm usually are less dynamic than the machine and deep learning techniques.

**Table 2** Comparison of performance metrics of various CNN models

Model No	Model	Accuracy	Precision	Recall	<i>F1</i>
1	Emb + Conv + Conv + Pool + FC	81.06	1.01	0.88	0.94
2	Emb + Conv + Pool + FC	79.70	0.95	0.96	0.95
3	Emb + Conv + Conv + Conv + Pool + FC	80.30	0.99	0.89	0.94
4	Emb + Conv + Pool + Conv + FC	78.17	0.91	1.03	0.96
5	Emb + Conv + globalpool + FC	77.54	0.99	0.87	0.93
6	Emb + Conv + Conv + globalpool + FC	79.06	1.00	0.87	0.93
7	Emb + Conv + Pool + Conv + Pool + FC	79.11	0.98	0.89	0.94
8	Emb + Conv + Pool + Conv + Pool + Conv + Pool + FC	74.61	1.15	0.65	0.86
9	<i>Emb + Conv + Conv + Conv + ReLU + ReLU + ReLU + Pool + Pool + Pool + FC</i>	90.7	1.11	0.98	0.96

**Fig. 5** Bar graph for Table 2

Though for training these models we require intensive pre-processing of data. From the survey, we conclude that besides the language limitations, we also require efficient processing modules. A deeply processed dataset leads to a better training model. The CNN model is much more similar to the cognitive skills of a human and should be better at classifying the subjective and objective data in the given text. The training process can take a while, unlike other algorithm based models. But, the precision of the data classified is comparatively better.

## References

1. F. Bravo-Marquez, M. Mendoza, B. Poblete, Meta-level sentiment models for big social data analysis. *Knowledge Based Syst.* **69**, 86–99 (2014)
2. C. Cardie, Sentiment analysis and opinion mining Bing Liu (University of Illinois at Chicago) Morgan & Claypool (Synthesis Lectures on Human Language Technologies, edited by Graeme Hirst, 5(1), pp. 167 (2012)
3. S. Weiss et al., Text mining: predictive methods for analyzing unstructured information (Springer, WordNet, 2011.WordNet, 2004). <http://www.cogsci.princeton.edu:80/~wn>
4. M.A.S. Cabezudo, N.L.S. Palomino, R.M. Perez, Improving subjectivity detection for spanish texts using subjectivity word sense disambiguation based on knowledge, in *Latin American Computing Conference* (2015)
5. V. N, Spam detection using sentiment analysis of text. *Int. J. Res. Appl. Sci. Eng. Technol.* **7**(3), 2252–2255 (2019). Available: <https://doi.org/10.22214/ijraset.2019.3413>
6. J. Wiebe, E. Riloff, Creating subjective and objective sentence classifiers from unannotated texts, in *Computational Linguistics and Intelligent Text Processing*, pp. 486–297 (2005)
7. J. Wiebe, R.F. Bruce, T.P. O’Hara, Development and use of a gold standard data set for subjectivity classifications, in *Proceedings of the Association for Computational Linguistics (ACL-1999)*, pp. 246–253 (1999)
8. R. Quirk, S. Greenbaum, G. Leech, J. Svartvik, A comprehensive grammar of the English language, vol. 397, Cambridge Univ Press (1985)
9. M. Saif, Challenges in sentiment analysis (2017). [https://doi.org/10.1007/978-3-319-55394-8\\_4](https://doi.org/10.1007/978-3-319-55394-8_4)
10. H. Kim, Y.-S. Jeong, Sentiment classification using convolutional neural networks. *Appl. Sci.* (2019)
11. T. Patel, M. Gupta, A.J. Agrawal, Brand analysis using named entity recognition and sentiment analysis. *Int. J. Comput. Appl.* **179**(45), 1–5 (2018). Available: <https://doi.org/10.5120/ijca2018917139>
12. A. Balahur, R. Mihalcea, A. Montoyo, Computational approaches to subjectivity and sentiment analysis: present and envisaged methods and applications, *Comput. Speech Lang.* **28**(1), pp. 1–6 (2014). Available: <https://doi.org/10.1016/j.csl.2013.09.003>
13. R. Muchhal, Investigation of ambiguity based sentiment analysis for product recommendation on E-Commerce portal. *Int. J. Res. Appl. Sci. Eng. Technol.* **6**(3), 1225–1229 (2018). Available: <https://doi.org/10.22214/ijraset.2018.3192>
14. S. Saumya, J. Singh, Detection of spam reviews: A sentiment analysis approach. *CSI Trans. ICT* **6**(2), 137–148 (2018). Available: <https://doi.org/10.1007/s40012-018-0193-0>
15. A. Luh Putu Karina Febriyanti, An analysis of the use of American slangs on Eminem’s song lyrics. *Lingua Scientia* **24**(2), 59 (2017). Available: <https://doi.org/10.23887/lsc.v24i2.18803>
16. R. Mansour, M.F.A. Hady, E. Hosam, H. Amr, A. Ashour (2015) Feature selection for twitter sentiment analysis: an experimental study, in A. Gelbukh (ed.) *Computational Linguistics and Intelligent Text Processing (CICLing, 2015)*
17. O.A. Ghaleb, A.S. Vijendran, The challenges of sentimental analysis on social web communities. *Int. J. Adv. Res. Sci. Eng.* **6**(12) (Dec 2017)
18. H. Keshavarz, M.S. Abadeh, SubLex: generating subjectivity lexicons using genetic algorithm for subjectivity classification of big social data, CSIEC (2016)
19. H. Keshavarz, M.S. Abadeh, SubLex: Generating subjectivity lexicons using genetic algorithm for subjectivity classification of big social data, in *1st Conference on Swarm Intelligence and Evolutionary Computation (CSIEC2016)*, (Higher Education Complex of Bam, Iran, 2016)
20. A. Gelbukh, in *computational linguistics and intelligent text processing* (Springer International Publishing, 2015), pp. 92–103
21. M. Asghar, Detection and scoring of internet slangs for sentiment analysis using SentiWordNet. *Life Sci. J.* (2014)
22. E. Bezerra, B. Firmino, R. Castaneda, J. Soares, E. Ogasawara, R. Goldschmidt, A subjectivity detection method for opinion mining based on lexical resources. IADIS (2011)

23. A. Go, R. Bhayani, L. Huang, Twitter sentiment classification using distant supervision. Technical report, Stanford University (2010)
24. M. Thelwall, K. Buckley, G. Paltoglou, Sentiment strength detection for the social web. *J. Am. Soc. Inform. Sci. Technol.* **63**(1), 163–173 (2012)
25. A. Esuli, F. Sebastiani, SENTIWORDNET: a publicly available lexical resource for opinion mining, in *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)*, pp. 417–422. Available at: SENTIWORDNET: A publicly available lexical resource for opinion mining (2006)
26. A. Esuli, F. Sebastiani, SentiWordNet: a publicly available lexical resource for opinion mining, in *Proceedings from International Conference on Language Resources and Evaluation (LREC)*, (Genoa, 2006)
27. A. Esuli, F. Sebastiani, Determining term subjectivity and term orientation for opinion mining, in *Proceedings of EACL-06, 11th Conference of the European Chapter of the Association for Computational Linguistics* (Trento, IT, Forthcoming, 2006)
28. B. Pang, L. Lee, A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts, in *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics ACL'04* (Barcelona, Association for Computational Linguistics, Spain, 2004)
29. S. Rustamov, E. Mustafayev, M. A. Clements, An application of hidden Markov models in subjectivity analysis, in *2013 7th International Conference on Application of Information and Communication Technologies* (Baku, 2013), pp. 1–4. <https://doi.org/10.1109/ICAICT.2013.6722756>; A. Fink, in *Markov Models for Pattern Recognition* (Springer, 2010), pp. 248
30. Y. Goldberg, Neural network methods for natural language processing. *Synth. Lect. Hum. Lang. Technol.* pp. 1–309 (2017)
31. B. Agarwal, N. Mittal, Prominent feature extraction for sentiment analysis, (Springer, 2016). <https://doi.org/10.1007/978-3-319-25343-5>
32. A. Rao, K. Shah, Model for improving relevant feature extraction for opinion summarization, in *2015 IEEE International Advance Computing Conference (IACC)*, (Banglore, 2015), pp. 1–5. <https://doi.org/10.1109/IADCC.2015.7154660>
33. Goldberg, Neural network methods for natural language processing. *Synth. Lect. Hum. Lang. Technol.*, pp 1–309 (2017)
34. A. Hassan, A. Mahmood, Deep learning approach for sentiment analysis of short texts, in *3rd International Conference on Control, Automation and Robotics (ICCAR)*, (2007), pp. 705–710
35. R. Socher, C.C. Lin, C. Manning et al (2011) Parsing natural scenes and natural language with recursive neural networks, in *Proceedings of the 28th International Conference on Machine Learning (ICML)*, (2011), pp. 129–136
36. R. Socher, A. Perelygin, J.Y. Wu et al, Recursive deep models for semantic compositionality over a sentiment treebank, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2013), pp. 1631–1642
37. T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space. [arXiv:1301.3781](https://arxiv.org/abs/1301.3781) [cs] (2013)
38. J. Pennington, R. Socher, C.D. Manning, GloVe: Global vectors for word representation, in *Empirical Methods in Natural Language Processing (EMNLP)*, (2014), pp. 1532–1543
39. Y. LeCun, L. Bottou, Y. Bengio Y, Gradient-based learning applied to document recognition. *Proc. IEEE* **86**(11), pp. 2278–2324 (1998)
40. K. Chatfield, K. Simonyan, A. Vedaldi et al, Return of the devil in the details: Delving deep into convolutional nets. [arXiv:1405.3531](https://arxiv.org/abs/1405.3531) (2014)
41. G.E. Hinton, N. Srivastava, A. Krizhevsky et al, Improving neural networks by preventing co-adaptation of feature detectors. [arXiv:1207.0580](https://arxiv.org/abs/1207.0580) (2012)
42. B. Liu, Y. Dai, X. Li et al, Building text classifiers using positive and unlabeled examples, in *IEEE 3th International Conference on Data Mining (ICDM), 2003* (IEEE, 2013), pp. 179–186
43. G. Li, S.C.H. Hoi, K. Chang et al, Micro-blogging sentiment detection by collaborative online learning, in *IEEE 10th International Conference on Data Mining (ICDM), 2010* (IEEE, 2010), pp. 893–898

44. Y. Jia, E. Shelhamer, J. Donahue et al, Caffe: convolutional architecture for fast feature embedding, in *Proceedings of the ACM International Conference on Multimedia*, (ACM, 2014), pp. 675–678
45. T. Mikolov, W. Yih, G. Zweig, Linguistic regularities in continuous space word representations, in *HLT-NAACL*, (2013), pp. 746–751
46. D. Vu, T. Nguyen, T.V. Nguyen, T.N. Nguyen, F. Massacci, P.H. Phung, A convolutional transformation network for malware classification, in *2019 6th NAFOSTED Conference on Information and Computer Science (NICS)*, (Hanoi, Vietnam, 2019), pp. 234–239. <https://doi.org/10.1109/NICS48868.2019.9023876>
47. M. Anbarasan, B. Muthu, C. Sivaparthipan, R. Sundarasekar, S. Kadry, S. Krishnamoorthy, A.A. Dasel, Detection of flood disaster system based on IoT, big data and convolutional deep neural network. *Comput. Commun.* **150**, 150–157 (2020). <https://doi.org/10.1016/j.comcom.2019.11.022>
48. X. Ouyang, P. Zhou, C. H. Li, L. Liu, Sentiment analysis using convolutional neural network, in *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing, Liverpool*, (2015), pp. 2359–2364. <https://doi.org/10.1109/CIT/IUCC/DASC/PICOM.2015.349>

# Enhanced Movie Recommender System Using Hybrid Approach



R. Lavanya, V. S. Bharat Raam, and Nikil Pillaithambi

**Abstract** Current COVID-19 pandemic scenario, people's exposure toward movies and series has increased phenomenally. At present, each family has a Netflix/Amazon Prime account. But the question is "What to watch?" Few movie recommendation sites exist, but these are predominantly based on user's watch history. Our project differs by utilizing reviews, comments, watch history and also information from similar users. Based on these, we recommend movies. We are using collaborative filtering, content filtering, and demographic filtering to recommend movies for the users. The machine learning algorithms are implementing K-means and KNN algorithm.

**Keywords** Collaborative filtering (CF) · Content-based filtering (CBF) · Recommender system · K-means clustering

## 1 Introduction

In the umbrella of artificial intelligence, machine learning lies as a part. Machine learning is a concept in which without the intervention of human, the machine is able to learn on its own. In machine learning, the learning is done by analyzing the data from the past. There are some machine learning algorithms, and we are implementing K-means clustering and K-nearest neighbors algorithm. Machine learning

---

R. Lavanya (✉)

Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai 603203, India  
e-mail: [lavanyar@srmist.edu.in](mailto:lavanyar@srmist.edu.in)

V. S. B. Raam · N. Pillaithambi

Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Chennai 603203, India  
e-mail: [vs4754@srmist.edu.in](mailto:vs4754@srmist.edu.in)

N. Pillaithambi

e-mail: [np9234@srmist.edu.in](mailto:np9234@srmist.edu.in)

finds its application in various real-time scenario like Netflix and Amazon Prime movie recommendation systems.

Big data refers to an extremely large volume of data that is constantly increasing at a higher rate. Hidden patterns and correlation between the data either in structured or unstructured format are found by examining the data, and this is termed as big data analytics. Data mining is the technique of extracting valuable information from raw data. The information is then used in taking informed business decisions. Big data is used to increase the efficiency of the recommendation system's algorithm by using it as training data. Big data is used by companies like Netflix to improve their streaming quality and decide on subscription charges based on analyzing the data from the user's viewing habits and their preferences.

## 2 Proposed Method

The recommendation in today's scenario mainly encounters two issues. One is cold start issue and another is data sparsity issue. Cold start problem is a problem faced by the recommender system when an user is new to the system. At this time, it is difficult for the recommender system to recommend movies for the user. Similarly, data sparsity is a problem faced by the recommender system when a new movie is introduced into the system and the system has no information about the movie and recommendation becomes difficult. Our idea behind the project is to handle cold start problem. Our proposed methods for overcoming this problem are as follows:

- Favorite movies and ask them to select list of adjectives for reviews.
- Sync contacts.
- Give Twitter handle and perform NLP on the tweets.
- Psychometric test.
- Use Twitter API to get list of their like pages in Twitter and analyze them to recommend movies.
- Utilize their location, age, occupation and location to provide wide genre of film according to the above criteria later if a person selects a particular genre movie in the next set of recommendation add more movies of that genre.
- Use an int value for each genre continuously increment and decrement to recommend movies of certain genres.

## 3 Related Work

This paper [1] overcomes the absence of prehistory of the users to recommend movies. One big problem in movie recommendation system is the problem of getting review and recommending new films when released. This problem makes the recommendation system face difficulty in recommending movies to users. In this method, tweets from Twitter are retrieved and sentimental analysis is performed on the tweets;

an overall conclusion of the movie's emotion is found. With this information, recommendation is done. In the current paper, CBF and CF are performed on the movie datasets from Netflix, MovieLens, and Imdb. Machine learning algorithm like K-means clustering is applied. The sentimental analysis on the tweets is done using Vader and SentimentIntensityAnalyser in Python. The drawback of this methodology is that it is difficult to retrieve and clean tweets from Twitter. Recommendation systems usually use CB to recommend films to its users. CB filtering uses the data of other users to recommend films. This method increases the time complexity drastically in real time.

This paper [2] aims to overcome the problem of time complexity by electing a virtual leader. Similar users having the same taste in movies are grouped together. A virtual leader is elected for each cluster. Now recommendation becomes easier saves lot of time and uses K-means to cluster users. Weighted slope one VU is applied on virtual leaders for recommendation. Drawback of this methodology is when electing a virtual leader we may get a general recommendation. Sparsity problem is when there isn't enough information about a movie or an user and his/her ratings. This results in poor prediction of user's ratings and movie recommendation. The methodology proposed in this paper is to combine CF and demographic filtering.

In this method [3], age, gender, occupation, and location of user are combined with their ratings to movies with this ratings and for movies with data sparsity is calculated, thus reducing data sparsity. This uses K-nearest neighbors' algorithm to find similar users. Further, random forest and neural networks are used for recommendation. The drawback of this algorithm is that similar users derived from demographic filtering need not necessarily like the same types of movies or be interested in the same regional movies. Demographic filtering might restrict the range of movie recommendations to the user.

This paper [4] aims to surge the precision of CF. This uses the k-clique method. First the user's personal information is fetched like age and location. Later using cosine similarity, the users are grouped. Similar users are given a value of 1. Now based on the characteristics, users are clustered and the cluster with maximum 1's is introduced movies based on their characteristics. This methodology uses k-clique method and cosine similarity to cluster users. Further KNN, CF is used for evaluation purpose. The drawback of this time complexity of clustering users. If users increase time, complexity may become an issue affecting the speed and efficiency of recommendation.

The current method [5] uses the methodology of CF, CBF, and natural language processing to recommend movies. This hybrid method of combining CF and CBF is reduced the computational complexity and increases the accuracy. In this method apart from extracting the ratings of movies by users, additionally the side information (genome tag) given by the users is also taken into consideration and natural language processing is applied on to it. This content can be huge, to compress this autoencoder is used. Later with this information similarity between users is found. The ratings derived using matrix factorization are combined with this to reduce computational complexity. The drawback of this model is that the side information may not be syntactically and grammatically correct to apply NLP on it.

In the current [6] methodology, a weight is assigned to each movie, and this weight is not same for everyone. This weight is generated based on the user and his preferences. In the first step, the user's favorite actor, director, genre, movie, and year is taken into consideration. Later using this weights and user information, using K-means algorithm movies are recommended. Using this method, accuracy of movie recommendation can be improved. But the drawback of this methodology is that when the number of users increases, computing weights based on user information is time-consuming and results in time and space complexities.

This idea [7] aims to reduce the error of CBF and CF. This methodology uses K-nearest neighbor and K-means to recommend movies. Since in this method, the clustering of users is reduced and two powerful algorithms are K-means and KNN. This uses python modules like numpy and pandas for data handling. Clustering of users might be too simple, and no specific constraints is imposed, and thus, a general clustering might not give a more personalized movie recommendation; this is the drawback of this methodology.

This paper [8] aims at providing scalable and robust movie recommendation and increasing the accuracy of recommendation. This is done by utilizing the user's past behavior and interaction with movies. In this method, matrix of users and movies is combined to form an utility matrix. Later K-means clustering is used to cluster similar users. Sparsity problem in the matrix is solved by alternating least squares method. Tags are assigned to each movies. With these tags, users need to select their favorite out of the list of tags. Later with this information, further recommendations are done. Thus, tags of movies act as the connection between users and system. The drawback of this methodology is the time complexity of clustering users.

This paper [9–11] focuses on improving the accuracy of clustering process to improve the overall recommendation accuracy. The user's review is available in the MovieLens, and this is used as the input for the clustering process. Kullback–Leibler Divergence-based fuzzy C-means clustering is used to improve the film accuracy. The users are clustered based on the similarity computed using Hellinger distance measure. This uses cosine similarity which results in better outcomes. The three potential steps involved in this method are KL divergence-based fuzzy C-means user clustering, closest cluster computation, and prediction of best recommended items on the basis of peer user's ratings.

The current paper [12] aims at improving the efficiency of CF method by using a similarity measure to find the similarity between users. After that, Kcliques is applied to generate clusters and suggest movies in homogeneous collections. The most optimal value for k in k-clique algorithm is found and then used to increase the accuracy of the recommender system. The suggested method in this paper is better than existing maximal clique and CF. The time taken to calculate the k-clique methods is long.

This paper [13] uses K-means algorithm and artificial bee colony (ABC) to improve the accuracy of the prediction. ABC and K-means optimization technique are used to cluster the dataset. ABC is an optimization technique that detects comb conduct of honeybee group which can be used in optimization works. Fitness function is developed for the clustered data by the ABC to improve the user's centroid

distance. The proposed system in this paper offers high performance in terms of reliability, accuracy, and personalization of movie recommendations using specific number of clusters.

This paper [14] uses an amalgamation of recommendation system that culminates both CF and CBF for reducing cold start problem and data sparsity issues in movie recommendation systems. Content-based filtering will be used on the properties of the movies interacted by the user to solve the cold start problem. Deep autoencoder network combined with CF is used to predict the ratings for movies not watched by the user. The deep autoencoder network is trained on user's ratings, interests, and personal influence. This approach can also be used in other recommendation systems. This method totally depends on the availability of the user's social profile.

This paper [15] tries to improve the film recommendation system and reduce the data sparsity problem. Heterogeneous SMR network is used that exploits a movie's multimodal content that incorporates its visual poster and text description, their public relationships, and corresponding preference for film recommendation. The proposed method tries to overcome the absence of discerning features for multimodal film that current models suffer from. In two sub-networks, multi-modal neural network is introduced to learn from both the textual and visual representation of the data. The heterogeneous SMR network is integrated with the neural network to create a social-aware movie recommendation framework.

This is [16] methodology uses the temporal user preference to recommend movies. In this process, the movies liked by users are taken into account. Later the content of the liked movies is analyzed, and based on this information, the system predicts and recommends further movies to the users. Similarity and latent factors are the two factors that are used here. The latent factor connects the users and the movies into a common space. In this model, the user's information such as location, age, time, and social network is correlated with the movie / content information such as name, genre, cast, and crew. Single value decomposition is used for obtaining rating matrix. Further here NLTK is used in the movie contents for sentimental analysis. The drawback of this methodology is the authenticity of the content and ratings. Further time complexity can be an issue in CBF.

In the current paper, CF-based recommendation system is implemented in this paper using Apache Mahout. The user rating is taken as the data to provide the recommendation. The similarity and the correlation between the items are used to develop this system, utilizing both item-based and user-based filtering techniques. User-based filtering involved finding the explicit ratings and implicit ratings, and then the nearest neighbor is found. Item-based filtering is used to detect the similarity between two products or items, and recommendation is based on that. Susceptible to cold start and or sparsity problems. The accuracy of hybrid recommendation systems is higher than CF. It also suffers from data sparsity and cold start problems. The comparison of existing methodologies is given in Table 1.

**Table 1** Comparison of existing methodologies

Paper	Objective	Abstract	Techniques	Drawbacks
<i>Movie Recommendation System Using Sentiment Analysis From Microblogging Data (centered)</i>	Recommend movies based on tweets	Using sentimental analysis tweets of users are analyzed and the user's emotion are deciphered and corresponding movies are recommended to the users. The idea of the journal is to use movie tweets to understand the current trends, public sentiment, and user response of the movie	Content-based filtering, Collaborative-based filtering, and Sentimental analysis. KNN Algorithm is used	Scraping all tweets by user and implementing sentimental analysis will be time-consuming Considering all tweets of an user might result in a lot of unrelated data
<i>Personalized real-time movie recommendation system: Practical prototype and evaluation</i>	The objective of this journal is to reduce the time complexity of collaborative filtering and implement content-based filtering along with it	This journal is about implementing a highly efficient recommender system by reducing the time complexity of CBF. This is done by grouping users in a cluster which is headed by a virtual opinion leader	This is implemented using K-means clustering algorithm. A new CF is implemented to reduce the time complexity, this is called Weighted KM-Slope-VU	Giving a virtual leader might make the movie recommendation much generic than personalized
<i>Improved Movie Recommendations Based on a Hybrid Feature Combination Method</i>	The objective of this journal is to combine collaborative filtering, content-based filtering and knowledge-based filtering	Implementing the above results in reduced errors in rating predictions based on users' past interactions, which leads to improved prediction accuracy. Demographic Filtering helps in recommending relevant movies to relevant users using their location	Uses CF, CBF, Knowledge-based filtering and Demographic Filtering	Demographic filtering might restrict the range of movie recommendations to the user

(continued)

**Table 1** (continued)

Paper	Objective	Abstract	Techniques	Drawbacks
<i>An efficient movie recommendation algorithm based on k-clique method</i>	This paper aims to improve the accuracy of collaborative filtering. In this method, k-clique method is used for evaluation purpose	This methodology uses k-clique method and cosine similarity to cluster users. Further KNN, collaborative filtering is used for evaluation purpose	K-clique, KNN collaborative filtering	If users increases time complexity may become an issue affecting the speed and efficiency of recommendation
<i>Weight-based movie recommendation system using K-means algorithm</i>	This method helps in recommending movies based on weights assigned to each movie	In this journal, the idea is to create weights to a movie based on a formula they've derived. These weights are also dependent on the user's interaction and reviews. Later the movies with similar weights are combined to clusters and later recommended to users	K-means algorithm, CF and CBF	Assigning weights to a movie is subjective and may vary from users to users. This might result in irrelevant movie recommendations sometimes
<i>Movie Recommender System Using K-Means Clustering AND K-Nearest Neighbor</i>	The objective of this journal is to make movie recommendation algorithm based on improved k-clique methods which provides accurate movie recommendations	The main purpose of this document is to achieve a more effective solution than collaborative filtering. K-means is used to group similar users. K-means creates clusters, and KNN introduces movie in similar groups	Uses CF, CBF and K-means and KNN algorithm	Clustering users might result in certain movies not being recommended to certain clusters
<i>Hybrid Recommendation System with Collaborative Filtering and Association Rule Mining using Big Data</i>	This paper aims at providing scalable and robust movie recommendation and increasing the accuracy of recommendation	In this method matrix of users and movies are combine to form an utility matrix. Later K-means clustering is used to cluster similar users	K-means clustering	Time complexity of clustering users

(continued)

**Table 1** (continued)

Paper	Objective	Abstract	Techniques	Drawbacks
<i>A Kullback–Leibler divergence-based fuzzy C-means clustering for enhancing the potential of an movie recommendation system</i>	Kullback–Leibler divergence-based fuzzy C-means clustering is proposed for enhancing the accuracy of movie recommendation system	The KL divergence-based cluster is included in clustering methods to enhance the stability and robustness in clustering process which results in improved sqrt-cosine similarity, used to find the effective nearest neighbors for an active user	KL divergence-based cluster, Fuzzy C-means clustering algorithm, KLD-FCM	It clusters an entity based on only its self-features and do not incorporate the influence of the entity's neighborhoods which makes clustering prone to additive noise
<i>An Efficient movie recommender algorithm based on improved k-clique</i>	Increase accuracy of recommender systems using improved k-clique methods	The improved k-clique method used in this paper performs better than the maximal clique method used in social network analysis, when used for movie recommendation systems	K-clique algorithm	It takes a long time to calculate the k-clique methods
<i>Movie recommender system with metaheuristic artificial bee</i>	A hybrid recommendation system with high scalability, performance and produce accurate recommendations by reducing cold start problem	A hybrid recommender system utilizing K-means clustering algorithm with bio inspired artificial bee colony (ABC) optimization technique	ABC-KM Artificial Bee Colony algorithm K-means clustering algorithm	Cross domain data can be used to improve recommendation

(continued)

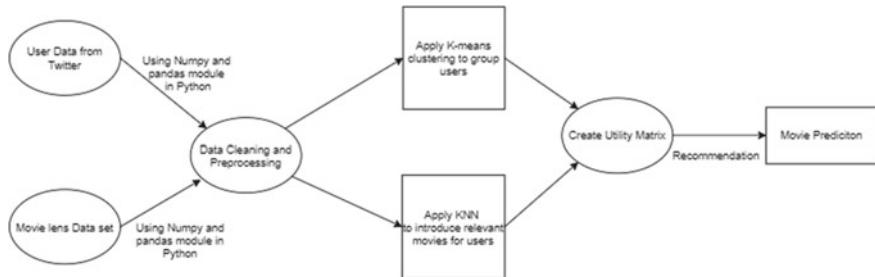
**Table 1** (continued)

Paper	Objective	Abstract	Techniques	Drawbacks
<i>Social movie recommender system based on deep autoencoder network using Twitter data</i>	This method makes use of user's social influence (social characteristics and behaviors on Twitter) to increase the accuracy and effectiveness of recommendation system	A hybrid social recommender system utilizing deep autoencoder network to recommend movies based on collaborative, content-based and social influence of the user	Deep autoencoder network (Deep learning). SRDNet	Depends on user's social data. User may not be active on Twitter
<i>Social-Aware Movie Recommendation via Multimodal Network Learning</i>	This approach is to resolve the sparsity problem inherent in SMR data	The sparsity problem is addressed by learning a multimodal network representation for ranking movie recommendations. heterogeneous SMR network exploits textual representation and movie poster image of each movie as well as user ratings for movie recommendation	Multimodal Network Representation Learning	Unrecognizable characters in movie poster reduces performance
<i>Movie Recommendation via Markovian Factorization of Matrix Processes</i>	This paper presents a new model family termed Markovian factorization of matrix process (MFMP)	MFMP although simple and primitive, already have comparable or even better performance than time SVD + + and a standard tensor factorization model	Markovian Factorization of Matrix Process (MFMP)	Inrequent visitors affect accuracy
<i>A Content-based Movie Recommender System based on Temporal User Preferences</i>	This is methodology uses the temporal user preference to recommend movies	In this process the movies like by users are taken into account. Later the content of the liked movies are analyzed, and based on this information, the system predicts and recommends further movies to the users	Temporal user preference, NLTK	Time complexity problems, authenticity of the content and ratings cannot be verified

(continued)

**Table 1** (continued)

Paper	Objective	Abstract	Techniques	Drawbacks
<b><i>Movie Recommendation System Using Collaborative Filtering</i></b>	This paper presents an approach to improve recommendation efficiency using collaborative filtering method	The user rating is taken as the data to provide the recommendation. The similarity and the correlation between the items is used to build this system, employing both user-based and item-based filtering techniques	Apache Mahout	Stuffers from data sparsity and cold start problems



**Fig. 1** Enhanced movie recommender system architecture diagram

## 4 Architecture Diagram

This movie recommendation system exploits the usage of collaborative and content-based filtering. The proposed system takes data from the Twitter handle of the user, using Twitter API. The twitter data of the user and the dataset from MovieLens is first cleaned and preprocessed. Later machine learning algorithms like K-means clustering and KNN algorithms are applied to the data. The process involved in this movie recommendation is represented in Fig. 1.

## 5 Conclusion

Using this hybrid approach of recommending movies, i.e., combining many filtering models to recommend movies, we are able to recommend much personal movies. In this paper, we have handled the cold start problem. This problem occurs when the system experiences the entry of new users. During this time, the system is completely unaware of the user's taste in movies, thus making it difficult for the system to recommend movies. We tackle this issue by conducting a psychometric and fetching twitter data of the users and later apply NLP on it to deduce information about the users to recommend movies.

## References

1. S. Kumar, P.P. Roy, K. De, Movie recommendation system using sentiment analysis from microblogging data 28 May 2020, pp. 915–923. <https://ieeexplore.ieee.org/document/9103168>
2. Personalized real-time movie recommendation system: Practical prototype and evaluation, 02 Sept 2019, pp. 180–191. <https://ieeexplore.ieee.org/document/8821512>
3. R. Ahuja, A. Solanki, A. Nayyar, Movie recommender system using K-means clustering and K-nearest neighbor, 29 July 2019, pp. 915–923. <https://ieeexplore.ieee.org/document/8776969>

4. M.T. Himel, M.N. Uddin, M.A. Hossain, Y.M. Jang, Weight based movie recommendation system using K-means algorithm, 14 Dec 2017, pp. 915–923. <https://ieeexplore.ieee.org/document/8190928>
5. G. Alshammari, S. Kapetanakis, A. Alshammari, N. Polatidis, M. Petridis, Improved movie recommendations based on a hybrid feature combination method, 13 June 2019, pp. 363–376. <https://www.worldscientific.com/doi/10.1142/S219688819500192>
6. Movie Recommender System Using K-Means Clustering AND K-Nearest Neighbour: <https://ieeexplore.ieee.org/document/8776969>
7. Hybrid Recommendation System with Collaborative Filtering and Association Rule Mining using Big Data: <https://ieeexplore.ieee.org/document/8529683>
8. A Kullback–Leibler divergence-based fuzzy C-means clustering for enhancing the potential of an movie recommendation system: <https://link.springer.com/article/10.1007/s42452-019-0708-9>
9. S. Ramesh, C. Yaashuwanth, B.A. Muthukrishnan, Machine learning approach for secure communication in wireless video sensor networks against denial-of-service attacks. *Int. J. Commun. Syst.* **33**(12) (2019). <https://doi.org/10.1002/dac.4073>
10. N.T. Le, J.-W. Wang, C.-C. Wang, T.N. Nguyen, Novel framework based on HOSVD for Ski goggles defect detection and classification. *Sensors* **19**, 5538 (2019). <https://doi.org/10.3390/s19245538>
11. An Efficient movie recommendation algorithm based on improved k-clique: <https://hcisjournal.springeropen.com/articles/10.1186/s13673-018-0161-6>
12. Movie recommender system with metaheuristic artificial bee: <https://link.springer.com/article/10.1007/s00521-017-3338-4>
13. Social movie recommender system based on deep autoencoder network using Twitter data: <https://link.springer.com/article/10.1007/s00521-020-05085-1>
14. Social-Aware Movie Recommendation via Multimodal Network Learning: <https://ieeexplore.ieee.org/document/8010448>
15. A Content-based Movie Recommender System based on Temporal User Preferences: <https://ieeexplore.ieee.org/abstract/document/8311601>
16. Movie Recommendation System Using Collaborative Filtering: <https://ieeexplore.ieee.org/document/8663822>. Utilizing an Autoencoder-Generated Item Representation in Hybrid Recommendation System. <http://ieeexplore.ieee.org/document/9075162>

# Optimization of CNN in Capsule Networks for Alzheimer's Disease Prediction Using CT Images



P. R. Ananya, Vedika Pachisia, and S. Ushasukhanya

**Abstract** Alzheimer's disease is a cumulative and irreparable disorder that affects remembrance and some other mental functionality, commonly observed in elderly and ageing population. This disease involves the gradual deterioration of brain cells, gradually shattering memory and other important brain related functions. Amnesia and uncertainty are the common indicators. No cure is known to exist for Alzheimer's disease, but medication and management action plans may provisionally act upon the symptoms giving the patient more time before severity is reached. Medications can occasionally help people having Alzheimer's disease magnify brain functions for a particular duration of time. The use of innovative computer aided detection or diagnosis (CAD) approach with Content-Based Image Retrieval (CBIR) system has drastically opened up novel possibilities in the area of image retrieval, classification and training for detection of Alzheimer's disease during the starting stages using magnetic resonance imaging (MRI) technology. In this paper, we have proposed a CBIR system using the following modes: 3D Capsule Network, 3D Convolution Neural Network and a sparse 3D autoencoder. A 3D Capsule Network (CapsNet) has a capability of learning fast, for very small datasets as well. It can conclusively control image transitions and rotations. It was found that, a collaborative model using 3D CapsNet and 3D CNN with 3D sparse autoencoder, multiplied the prediction performance when compared with the traditional deep CNN.

**Keywords** Magnetic resonance imaging · Alzheimer's disease · Content based image retrieval · Capsule network · Convolution neural networks · Computed tomography

---

P. R. Ananya (✉) · V. Pachisia · S. Ushasukhanya

Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, India

S. Ushasukhanya

e-mail: [ushasuks@srmist.edu.in](mailto:ushasuks@srmist.edu.in)

## 1 Introduction

Various revolutionary computing technologies have come up today that have begun to make medical diagnosis easier. Especially, imaging technologies like image recognition, classification etc. using machine and deep learning techniques have ploughed their way into medical science thereby greatly benefiting and encouraging physicians, researchers and scholars for further research and analysis. The importance of imaging technology in the industry is very well understood, which has led to development and technical alterations through the purpose of analysing symptoms in medicinal imageries and deliver a contiguous understanding of the symptoms. Numerous medical imaging modal techniques such as advanced X-ray technique Digital Radiography, Mammogram (MG), MRI, Ultra sonographs, CBIR, cross sectional images from Computed Tomography (CT) are used today in order to obtain medical images for diagnosis. Databases comprising of collection of brain images obtained through MRI scans are of immense significance. The benefit of scanned MRI images is that it produces high three-dimensional resolutions as well as gives a detailed view of the images for thorough diagnosis of disease [1, 2].

One of the most common neurodegenerative diseases include Alzheimer's Disease which is a health condition commonly viewed in elderly and ageing population, revealed by slow memory loss and reduced cognitive functions [3, 4]. Worldwide, AD cases are estimated to grow suggestively over the next 40 years because of the ageing population and increasing life expectancy rate posturing a massive challenge to the society [5]. Alzheimer's disease regards with brain proteins which functions abnormally, interrupt the function of nerve cells and lead to a sequence of noxious actions. The nerve cells get harmed, misplace networks among themselves and ultimately get deceased. Generally, region of the brain that pedals memory is where the damage starts. But the procedure of degeneration commences years before the initial symptoms. The brain shrinks significantly by the time the later stages of the disease is reached. Researchers and medical professionals suggest that there could be a predictable pattern in which the neuron degeneration spreads to the other parts of the brain. The brain shrinks significantly by the time the later stages of the disease is reached.

Our purpose in this paper is to study the behaviour of Capsule Networks, which is an innovative architecture model to encrypt the characteristics and three-dimensional relations of an attribute in a data image that depicts heartening outcomes of an image classification. Here, CapsNet model is compared with the traditional conventional ConvNets below constrictions of databases of biomedical images such as class imbalance and inadequate volume of labelled data. Here, we will walk through the technological supremacy which makes Capsule Neural Networks more suitable to work with the above-cited issues and demonstrate its enhanced performance experimentally. We determine that the equal variance attributes of CapsNets decreases the robust data necessities and is hence a favourable architecture in medicine and health image examination.

## 2 Literature Survey

AD is an accelerating illness that causes cells of brain to degenerate and pass out. One of the most communal cause of dementia is Alzheimer's disease which causes nonstop weakening in thought processes, behavioural and social aids which upsets a man's skill to perform individualistically. Medications currently available for AD might momentarily recover indications or decrease the degree of deterioration. These medications might occasionally benefit individuals having AD exploit function and continue individuality through certain period. Utilization of radical computer aided diagnosis methods alongside Content-Based Image Retrieval (CBIR) [6] have opened up new abilities in the field magnetic resonance imaging (MRI) in similar image revival and recognition training of Alzheimer's disease movement during the initial phases. CapsNet model appears as a favourable recent system for image classifying, for researches and experimental work that need additional vigorous computational assets and polished CapsNet model architectures may deliver even healthier results [7].

There are different techniques that currently being used today to diagnose AD at early stages. Several studies for the recovery of pictures of brain via MRI databases linked through collections of data of images of brain have been done already. Brain imaging techniques have been benefiting the diagnosis of a large number of diseases [1, 2, 8–11]. Brain images now are widely used in pinpointing evident anomalies associated with constraints apart from AD like strokes, tumours or trauma that might cause intellectual modification. New imaging implementations, presently used mainly in chief medical hubs or in medical trials might permit medics to sense precise brain variations triggered through Alzheimer's.

Brain design images contain MRI and Computerized Tomography (CT). MRI releases comprehensive images of brain by means of radio waves and a strong magnetic field. To exclude other conditions, MRI scans are used [1, 2, 8–10]. Though it might demonstrate brain reduction, evidence does not presently add any notable value in making of a diagnosis. A Computerized Tomography scan and a very specialized X-Ray technology, processes cross-sectioned pictures of the brain and is presently utilized for removing strokes, tumours, and other brain wounds. Tomography of illness procedures can be accomplished using positron emission tomography (PET). When a low-level radioactive tracer is inserted into the blood to expose a particular characteristic in the brain, it is called as a PET scan [12].

The most significant trait of AD anatomy is nerve cell drop trailed by brain atrophy developing in AD chief areas (e.g. amygdala and hippocampus) to the whole cortical area which could also be recognized through MRI scan. Those recognizable layout variations have occurred well in advance perceptible deterioration in perception and hence offers a chance to AD initial recognition through an image tool. For this purpose, CBIR, one of the most powerful image reclamation tool and it might as well be used for handling huge datasets effectively. CBIR schemes are systems which empower effective image recovery grounded in the picture information. The working of a CBIR can be defined as the discovery of key visual fundamentals within a group

of images. The visual characteristics which may refer to colour, shape or texture can form visual words which describes an area in a data image which is utilized in a pictorial record which designates the entire image [13–18].

Multiple researches have been carried out to optimize the existing methods using techniques such as SVM and multi-tier technologies that provide flexibility in terms of experimentation with classification, representation, feedback and ranking [19]. Their accuracies were found to be about 90.65% and 82.45%, respectively. Deep learning methods are proficient in knowing such depictions using data, specifically with CNN [20]. Yet additional modern work analysed use of heavy CNN to difficulties and making computer aided detection (CAD) usage within the medicinal division.

Capsule Network is powerful for retrieval of information rotations, transformation and involves lesser datasets in model training with a lesser learning curve [21, 22]. Several other techniques projected to identify AD with various unique autoencoders or Three-Dimensional CNNs [1, 8]. The paper sightsees structure recovery by intermixing dual transfer training techniques, a 3D pre-trained autoencoder with shallow neural networks, 3D CapsNet to categorize Alzheimer’s disease sufferer through steady controller built on structures of brain scanned images using MRI. significant feature of Capsule Network was labelled as “routing by agreement”, denoting the below-layer capsules can accomplish foreseeing the outcomes and results of upper-layer capsules. Hence, the initiation of capsules to upper layers match with the harmony of few such estimations. CapsNet architecture model not only learns good weights for characteristic extraction and classification of images but also learns in what way to deduce parameters like three-dimensional pose within an image. For instance, a capsule model acquires to decide not only if an airplane is in the picture, but also the orientation (to the right or to the left or whether at all it is rotated). This property is called the equivariance.

### 3 Inferences

CNN is known to be the most appropriate model for image transitions using layer-pixel pooling mode. This enables retrieval of data arranged in the form of multiple arrays. This information in relay form is obtained as input through multiple layers, known as routing. Even though, CNNs need a huge amount of data from datasets for the process of training, they are barely able to control the input data transformations orderly and miss out on important points in certain situations.

CapsNet happens to be an innovative model that beats the drawbacks of the standard CNN model. It was recently introduced as a structural design coming under the machine learning process. CapsNet proposes to completely give a new dimension and thereby transforming the dee learning process. Compared to CNN, CapsNet requires datasets much less in number. CapsNet was found to handle the MRI images from the medical scans much more effectively with strong dynamic routing, involving parameters such as size, orientation, locations, rotations etc. This particular study involved

3D CapsNet and hence was found to have an accuracy of 94.06% for Alzheimer's disease prediction. Whereas, other models were found to have lesser accuracy in comparison. Furthermore, CapsNet model also helps researchers and medical staff obtain accurate results. CapsNet also is proficient in the handling of larger datasets with smaller sample size and less training turns. This brings the model closer to a successful model.

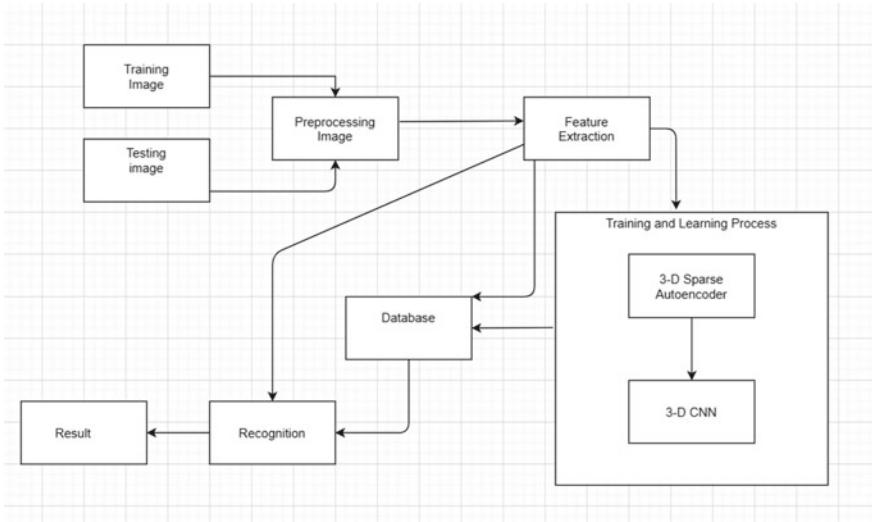
## 4 Proposed System

To curb the insufficiency and disadvantages in Convolutional Neural Networks, using recently announced machine learning operational model proposals called Capsule Network (also called CapsNet) that's appropriate for fast and profound learning of information image data. Capsule Network is powerful for information retrieval transformation and rotations, while requiring less data for model training and with a lower learning curve. One of the utmost significant feature of CapsNets is called as "routing by agreement", denoting that the junior-level capsules can predict the outcomes or results of capsules in upper-levels. Hence, the initiation of capsules within the upper layers depends on the lower layers.

In lower-level capsules, position data [23] is "position-coded" through the current active capsule because grading stays elevated, further more position based data is "rate-coded" [24, 25] within the actual data module of the capsule output vector. All those infer that as we climb up the ladder, the size of capsule sits essentially rise. Nevertheless, CapsNets gives permit to require complete benefit of characteristic three-dimensional relation [26] and imitate the flexibility of grasping data image variations to higher recapitulate which we observe. Alongside these noted features, we are employing CBIR architecture scheme for categorization for Alzheimer's disease while acting with huge datasets. With 3-D Convolutional Neural Networks and CapsNet from scrape with query-based techniques and CBIR for determining the F1 score. Ultimate outcome is equated with the introductory experimental results to AD estimations. The architecture of the proposed system is given in Fig. 1

### 4.1 Feature Extraction

Feature extraction is also known to be a procedure of dimensionality trimming through which a preliminary set of data is decreased to being more controllable and sensible clusters for process. This approach takes two stages. The initial phase is the training of a sparse 3D autoencoder used for convolutional filter learning. The next phase indicates construction of CNN, inside of what the above learned filter is employed for primary autoencoder covering.

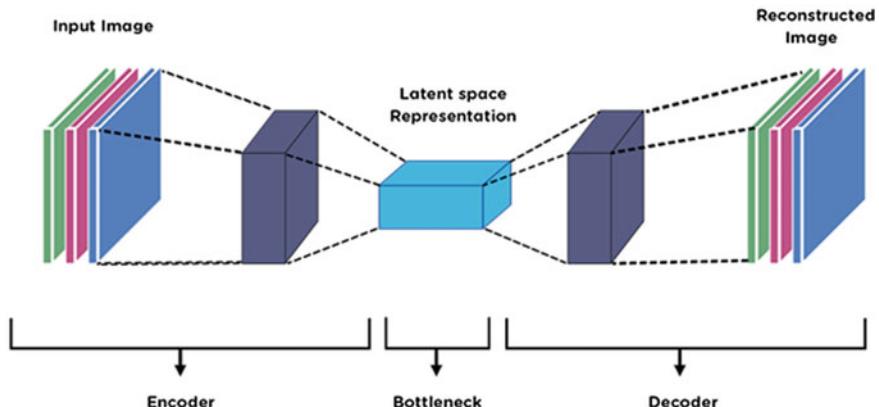


**Fig. 1** Architectural diagram of the proposed ensemble model

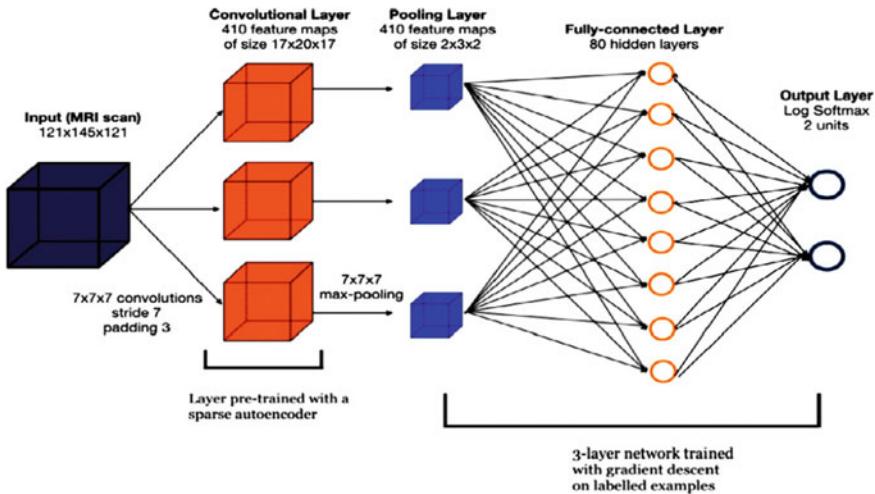
#### 4.2 Autoencoder—Sparse Autoencoder

Followed by the feature extraction, the autoencoder plays the vital role in this part for which the procedure is described in the following steps, followed by the structure of the autoencoder given in Fig. 2

- An artificial neural network of 3 layers which is used in unsupervised learning datasets and can take out information features from a dataset or a picture, by accessing computer file within the variety in “small parts or patches”, while



**Fig. 2** Structure of an autoencoder



**Fig. 3** Architectural classification of *neural network*

extracting the composition buried within the information or data is called an autoencoder.

- Three layers are—input or encoding, output or decoding, middle layers.

### 4.3 Training Process

The diagrammatic representation for the classification of the neural network is given in Fig. 3.

The second stage within the procedure contains the training part (3D-CNN). The Convolutional Neural Network model we use during this paper has 2 linear layers, 1 pooling layer, 1 convolutional layer, and eventually a SoftMax log layer. The trained model of autoencoder transports the biases and weights to 3-D layer of filter of first convolutional layer of 3D Convolutional Neural Network.

### 4.4 CapsNet Model

CapsNet comprises of a collection of neurons called capsules. They behave in a way, that the vector measurements signify the existence probability of three-dimensional features while various parameter poses are denoted as activity vectors of nerve cells. Mostly, inadequacies of CNN are due to the pooling layers present in the model. However, CapsNet follows a concept called “routing by agreement” which happens to be better suited in this case. According to this particular concept, the result generated

within the initial layer is transported to the next layer (parent-capsules). Even then, the capsule's coupling coefficient do not seem to be identical. Every capsule attempt in viewing the results of higher capsules. The linking coefficient between the capsules is improved.

## 5 Conclusion and Future Work

When CapsNet architecture was suggested, model architecture was quite simple, and several aspects required further improvement. Optimization of CapsNet model can be done in many ways like: “optimization of routing mechanism” and “increase Dropout operation” [28]. It can be concluded that for the capsule network very less dataset is needed to get trained whereas CNN will consume huge dataset to bring the results. Hence, computational cost can be minimized keeping the accuracy and prediction more accurate than CNN. CapsNet serves better computationally as well as in other aspects such as reducing time, cost, improving efficiency. Although even now presently there are no treatments that can lower down the disease progression, management and organization to the intellectual and behavioural indications of AD can remarkably advance the life of medical patients and guardians. It is known that Alzheimer's disease pathology consists of phospho-tau neurofibrillary tangles and  $\beta$ -amyloid plaques.

## References

1. M. Grossman, C. McMillan, P. Moore, L. Ding, G. Glosser, M. Work, J. Gee, What's in a name: voxel-based morphometric analyses of MRI and naming difficulty in Alzheimer's disease, frontotemporal dementia and corticobasal degeneration. *Brain* **127**(3), 628–649 (2004)
2. M. Mizotin, J. Benois-Pineau, M. Allard, G. Catheline, Feature-based brain MRI retrieval for Alzheimer disease diagnosis, in *2012 19th IEEE International Conference on Image Processing, ICIP* (2012), pp. 1241–1244
3. F. Hebert Liesi et al., Alzheimer disease in the United States (2010–2050) estimated using the 2010 census. *Neurology* **80**(19), 1778–1783 (2013). <https://doi.org/10.1212/WNL.0b013e31828726f5>
4. M. Langa Kenneth, Is the risk of Alzheimer's disease and dementia declining? *Alzheimer's Res Ther* **7**(1), 34 (2015). <https://doi.org/10.1186/s13195-015-0118-1>
5. W. Anders, J. Linus, B. John, P. Martin, W. Bengt, The worldwide economic impact of dementia 2010. *Alzheimer's Dementia* **9**(1), 1–11 (2013). <https://doi.org/10.1016/j.jalz.2012.11.006>.
6. MS-CapsNet: A Novel Multi-Scale Capsule Network Canqun Xiang; Lu Zhang; Yi Tang; Wenbin Zou; Chen Xu
7. K.R. Kruthikaa, B. Rajeswari, H.D. Maheshappab, *CBIR system using Capsule Networks and 3D CNN for Alzheimer's Disease Diagnosis*. Alzheimer's Disease Neuroimaging Initiative
8. J. Liu, J. Wang, B. Hu, F.-X. Wu, P. Yi, Alzheimer's disease classification based on individual hierarchical networks constructed with 3-D texture features. *IEEE Trans. NanoBioscience* **16**(6), 428–437 (2017). <https://doi.org/10.1109/TNBB.2017.2731849>
9. E. Worrall Daniel et al., Harmonic networks: deep translation and rotation equivariance, in *IEEE Conference on Computer Vision and Pattern Recognition IEEE Computer Society 2017*, pp. 7168–7177. <https://doi.org/10.1109/CVPR.2017.758>

8. I. Leonardo, Atrophy Measurement Biomarkers Using Structural MRI for Alzheimer's disease, in *The 15th Int. Conference on Medical Image Computing and Computer Assisted Intervention*. MICCAI (2012), p. 258
9. <https://www.mayoclinic.org/diseases-conditions/alzheimers-disease/diagnosis-treatment/drc-20350453>. PET/CT of Dementia
10. K. Zukotynski, P.H. Kuo, D. Mikulis, P. Rosa-Neto, A.P. Strafella, R.M. Subramaniam, S.E. Black, Content based image retrieval in the context of Alzheimer's disease, January 2014, in *Conference: 9th Annual South East European Doctoral Student Conference*, DSC 2014, Katarina Trojacanec, Ivan Kitanovski, Ivica Dimitrovski, Suzana Loshkovska, ADNI
11. M. Agarwal, J. Mostafa, Image retrieval for Alzheimer's disease detection, in *MICCAI International Workshop on Medical Content-Based Retrieval for Clinical Decision Support* (Springer, Berlin, 2009), pp. 49–60
12. M. Agarwal, J. Mostafa, Content-based image retrieval for Alzheimer's disease detection, in *IEEE, 9th International Workshop on Content-Based Multimedia Indexing CBMI* (2011), pp. 13–18. <https://doi.org/10.1109/CBMI.2011.5972513>
13. C.B. Akgul, D.L. Rubin, S. Napel, C.F. Beaulieu, H. Greenspan, B. Acar, Content-based image retrieval in radiology: current status and future directions. *J. Digit Imag.* **24**(2), 208–222 (2011)
14. R.A. Ben, B. Faouzi, A. Hamid, Diagnosis of Alzheimer diseases in early step using SVM (Support Vector Machine), in *IEEE 13th International Conference on, Computer Graphics, Imaging and Visualization (CGIV)* (2016), pp. 364–367. <https://doi.org/10.1109/CGIV.2016.76>
15. N.T. Le, J.-W. Wang, C.-C. Wang, T.N. Nguyen, Automatic defect inspection for coated eyeglass based on symmetrized energy analysis of color channels. *Symmetry* **11**, 1518 (2019). <https://doi.org/10.3390/sym11121518>
16. B. Muthu, C.B. Sivaparthipan, G. Manogaran, R. Sundarasekar, S. Kadry, A. Shanthini, A. Dasel, IOT based wearable sensor for diseases prediction and symptom analysis in healthcare sector. *Peer-to-Peer Netw. Appl.* **13**(6), 2123–2134 (2020). <https://doi.org/10.1007/s12083-019-00823-2>
17. F. Ammarah, A.S. Muhammad, A. Muhammad, R. Saad, A deep CNN based multi-class classification of Alzheimer's disease using MRI, in *IEEE International Conference on Imaging Systems and Techniques IST* (2017), pp. 1–6. <https://doi.org/10.1109/IST.2017.8261460>
18. M. Joani, Content-based image retrieval tutorial (2016). arXiv preprint arXiv 1608.03811. A. Rueda, J. Arevalo, A. Cruz, E. Romero, F.A. González, Bag of features for automatic classification of Alzheimer's disease in magnetic resonance images, in *Iberoamerican Congress on Pattern Recognition* (Springer, Berlin, 2012), pp. 559–566
19. A.A.M. Al-Saffar, H. Tao, M.A. Talab, Review of deep convolution neural network in image classification, in *IEEE 2017, International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications, ICRA MET* (2017), pp. 26–31
20. K. Simonyan, A. Zisserman, *Very Deep Convolutional Networks for Large-Scale Image Recognition* (2014). arXiv preprint arXiv 1409.1556 <http://arxiv.org/abs/1409.1556> (visited on 05/02/2018)
21. S. Liu, S. Liu, W. Cai, P. Sonia, K. Ron, D. Feng, Early diagnosis of Alzheimer's disease with deep learning, in *IEEE 11th International Symposium on Biomedical Imaging ISBI*, pp. 1015–1018 (2014). <https://doi.org/10.1109/ISBI.2014.6868045>
22. E. Xi, S. Bing, Y. Jin, *Capsule Network Performance on Complex Data* (2017). arXiv preprint [arXiv:1712.03480](https://arxiv.org/abs/1712.03480)
23. C.Q. Xiang et al., MS-CapsNet: a novel multi-scale capsule network. *IEEE Signal Process. Lett.* **25**(12), 1850–1854 (2018)
24. T.Y. Whye, G.E. Hinton, Rate-coded restricted Boltzmann machines for face recognition, in *Advances in Neural Information Processing Systems* (2001)
25. K.J. Eric, G.P. Abousleman, Adaptive-Rate Coded Digital Image Transmission, U.S. Patent No. 6154489, Nov. 28, 2000
26. M.D. Ward, K.S. Gleditsch, *Spatial Regression Models*, vol. 155 (Sage Publications, 2018)
27. S. Hoo-Chang, H.R. Roth, M. Gao, L. Lu, Z. Xu, I. Nogues, J. Yao, D. Mollura, R.M. Summers, Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset

- characteristics and transfer learning. *IEEE Trans. Med. Imag.* **35**(5), 1285 (2016). <https://doi.org/10.1109/TMI.2016.2528162>
28. M.T. McCann, K.H. Jin, M. Unser, Convolutional neural networks for inverse problems in imaging: a review. *IEEE Signal Process Mag.* **34**(6), 85–95 (2017). <https://doi.org/10.1109/MSP.2017.2739299>

# Red Lesion Detection in Color Fundus Images for Diabetic Retinopathy Detection



P. Saranya, K. M. Umamaheswari, Satish Chandra Patnaik,  
and Jayvardhan Singh Patyal

**Abstract** Diabetic retinopathy (DR) is a chronic disease in the eye due to blood leakages which causes vision impairment and can be identified on the surface of the retina. Diabetic patients are the most common subject to this disease, and ignorance to it can result in permanent visual damage and eventual blindness. DR falls in two categories: proliferative diabetic retinopathy (PDR) and non-proliferative diabetic retinopathy (NPDR). Non-proliferative DR is the most common form of DR, whereas proliferative DR is the severe stage of DR which causes the blood vessels to close off. DR can be detected in its early stages by the red lesions, i.e., microaneurysms and hemorrhages. In this paper, we implement a method of detecting red lesions for detection of NPDR form. The proposed methodology uses fundus images as its input and employs modified approach to extraction of retinal blood vessels and median filtering. To train the model, multiclass support vector machine classifier is implemented using the extracted features. The method is tested on 1928 fundus images from ‘KAGGLE APTOS’ and 103 images from ‘IDRiD’ dataset. The performance of the proposed methodology is as follows: sensitivity 75.6% and accuracy 94.5% on ‘KAGGLE APTOS’ and 78.5% sensitivity and 93.3% accuracy on ‘IDRiD’ dataset.

**Keywords** Red lesion · Diabetic retinopathy · Adaptive histogram equalization · SVM

## 1 Introduction

The advancement in computer technologies and specifically in artificial intelligence and machine learning has become one of the most important factors in development of medical field and its resources. Diabetic retinopathy is complication which causes damages to the blood vessels of the light sensitive tissue of the back of the eye called retina. DR has fatal consequences if it is not dealt with for a long period of time, one

---

P. Saranya (✉) · K. M. Umamaheswari · S. C. Patnaik · J. S. Patyal

Department of Computer Science and Engineering, SRM Institute of Science and Technology, Kattankulathur, Kancheepuram District, Tamilnadu 603203, India  
e-mail: [saranyap@srmist.edu.in](mailto:saranyap@srmist.edu.in)

of them is permanent blindness. One of the early symptoms of DR is red lesions, viz. microaneurysms and hemorrhages.

The paper proposes an algorithm to detect red lesions for prediction of diabetic retinopathy. The automated system for prediction of diabetic retinopathy uses color fundus images obtained by fundus camera as its input. The microaneurysms are tiny red dots that appear in the eye and due to vascular leakage are surrounded by yellow rings. Microaneurysms cause swelling of tiny retinal blood vessels and release fluid into the retina. They do not have any effect on the visual abilities of a person and have no other symptoms. The blot hemorrhages appear due to the ruptured microaneurysms in the retinal layer. Due to the blots, the blood flow to the retina reduces considerably, leading to PDR. The feature extraction process results in the extraction of a set of five features which are subsequently used in the multiclass support vector machine classifier training of the training dataset. The dataset is mixed with true and false lesion objects.

## 2 Related Work

The following literature presents various methods of detecting diabetic retinopathy using red lesions using fundus images. Harangi [1] used a combination of CNN and hand-crafted features to detect the disease. In which, he got an accuracy of 90.07%, but the major disadvantage of the model is classification of images with different positions. Kirange [2] used Gabor features and Naïve Bayes classification for the detection of DR where he got an accuracy of 74.46% and precision of 78.3%. The problem in the model was Naïve Bayes assumes that all features are not dependent and this algorithm faces the ‘zero-frequency problem’ where it assigns zero probability to a categorical variable. Alzami [3] used fractal dimension feature, which help them characterize the retinal vasculature in the DR predictions which gave him an accuracy of 91.6%, but random forest models are not all that interpretable and for very large datasets it can take lot of memory.

Gangwar [4] used Inception-ResNet-v2 and added a custom block of CNN layers on top of Inception-ResNet-v2 for building the hybrid model and presented an automated system for the detection of DR and got an accuracy of 82.18%. Shaban [5–7] used a deep convolutional neural network (CNN) with 18 convolutional layers and 3 fully connected layers for the prediction of DR, but the problem with CNN is that it doesn’t encode the position and orientation of the object and lacks of ability to be spatially invariant to the input data for which he got an accuracy of 88%. Our proposed model used SVM multiclass classifier model to the detect the diabetic retinopathy using SVM to get an accuracy of 93.3% in IDRiD and 94.5% in Kaggle Aptos dataset with 78.5% and 78.4% of sensitivity and precision in IDRiD dataset and 80.6% and 75.6% of sensitivity and precision in Kaggle Aptos dataset.

### 3 Proposed Model

In this paper, the proposed method detects presence of red lesions using the images from fundus camera as its input. Figure 1 shows the flow diagram for the proposed methodology.

#### 3.1 Preprocessing

Preprocessing is needed in order to remove any noise, get a better contrast and so that a more consistent dataset with only relevant features is obtained. In preprocessing, the first step involved is to filter and convert the fundus images from a 2D matrix to a 3D matrix in order to get 3 planes (red, blue, and green) so as to make it easier to perform morphological operations on the image. After the conversion to 3D matrix, a median filter is applied to remove additional noise or dilutions from the image. After the application of filter on the images, the green plane is extracted to clear the dark spots of red lesions, high of red lesions, and low value of other lesions. In order to spot the lesions, based on its properties of color it having low pixel in the green plane, resulting in dark spots in the green channel providing with the best contrast for the images. The green channel shows the proper numbers and dimensions including the major axis lengths where the region props are present.

The red lesions detection algorithm steps:

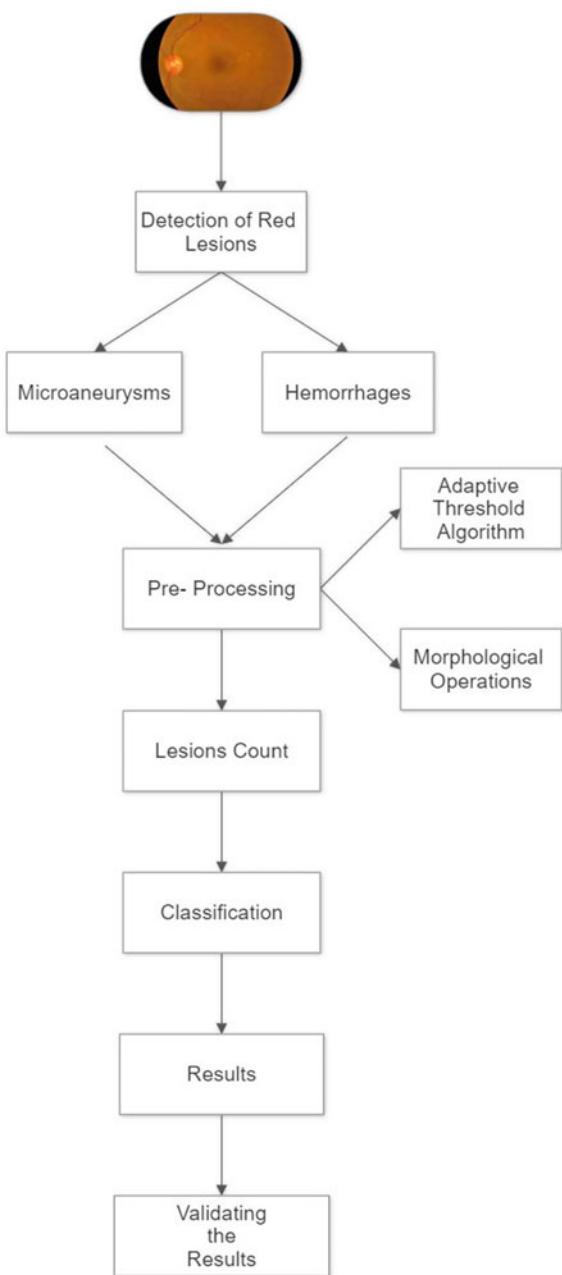
1. Adaptive threshold.
2. Filtering operations through morphological operations.

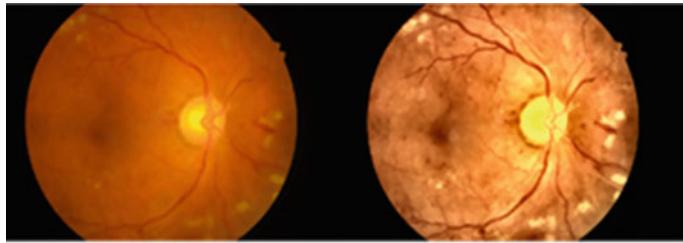
The method starts firstly by applying the adaptive threshold algorithm to the improved image with sensitivity level of 0.15, this image then to morphological operations which tries to filter any noise and perform area-based rejection to the output image as seen in Fig. 2 depicts the eye before and after the preprocessing stage. As we can see, the right picture has a better clarity than the left picture. So, in total, two filtering techniques were used.

#### 3.2 Extraction of Retinal Blood Vessels

Due to the blood vessels being the widest at the origin as well as being darker in the green channel, the process of detecting lesions after the removal of blood vessels is easier. Blood vessels are widest near the point of lesions and become narrow as we move away from that point. The gray level profile in the image is estimated using Gaussian function. But, due to the random nature of the blood vessels, performing morphological operations to extract feature is difficult, and therefore, matched filtering is used to determine the character of the feature [8–10]. The Gaussian kernel of the matched filtering is dependent on the value of variance parameter,

**Fig. 1** Flow of the proposed model





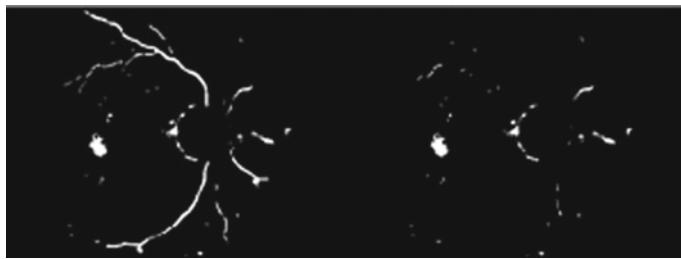
**Fig. 2** Enhanced figure of fundus image after applying filtering methods

sigma ( $\mu$ ). Results have shown that the value of this parameter for human retinal blood vessels ranges from 1.5 to 3. The classical approach uses just one value of  $\mu$ , whereas in our approach we use two values for  $\mu$ : 1.5 and 2. This is done in order to extract wider vessels as larger value extract wider vessels. Due to the random orientation of the blood vessels as mentioned before, the kernel rotates from  $0^\circ$  to  $180^\circ$  through an angle of  $15^\circ$ . The preprocessed image now has 12 kernels for the matched filtering. Each pixel of the image has 12 responses from the kernel, and the resulting image of the matched filtering is obtained by using the best value (maximum) for each pixel from the 12 responses. Similarly, we use the values 1.5 and 2 to obtain two output images. A threshold value has been set to separate the enhanced blood vessels. To search the threshold required, we use automatic thresholding technique. A co-occurrence matrix stores the derived threshold values when gray level intensity ‘ $x$ ’ has a close by value ‘ $y$ ’. The co-occurrence matrix is split into four quadrants ‘A’ to ‘D’ using the threshold value ‘ $th$ ’.

### 3.3 Candidate Lesions Detection and Classification

Candidate lesion detection has a multitude of approaches which perform morphological operations using techniques like h-maxima transformation, thresholding, and region growing. We employ the morphological operation of opening and then closing on the length filtered images and local entropy threshold of 0.15. The preprocessing operations result in a loss of a few lesions near the vessel segment, which are recovered by the sequential morphological operations of abrasion and dilation, performed on the image.

The recovered lesions are added to the images and are now detected. The resulting image was cleaned to obviate noise as seen in Fig. 3. The sensitivity level for the adaptive threshold is below 0.15, and therefore, the morphological operations are roughly an equivalent in bright regions. However, when extracting red regions, the veins were also extracted; this is often the most motivation to feature the anomaly rejection algorithm which tries to filter the veins by having a few ratio thresholds, and an expand threshold which measures how the tested object fills its corresponding bounding box. To distinguish the lesions and non-lesions, the proposed methodology



**Fig. 3** Removing noise and other disturbances and getting the final lesions

applies multiclass support vector classifier. SVM segregates the different classes using a hyper-plane in the feature space in the given image. This segregation window is maximized in order to best separate the categories using the multiclass SVM. The SVM classifier uses the extracted features to classify the images into categories of 0 for non-lesion and 1 for true lesion. For the values 1, lesions are counted in the output image.

## 4 Results and Discussions

We analyze and compare the performance of our model against the existing implementations of detection of diabetic retinopathy using red lesions on the datasets ‘KAGGLE APTOS’ and ‘IDRiD’. The output parameters for comparison are as follows: (i) accuracy, (ii) sensitivity, and (iii) precision. The model proposed by us has better output parameters as compared to the existing models. The performance metrics to analyze the model is defined in terms of true positives (TP), true negatives (TN), false positives, (FP) and false negatives (FN). These metrics are used to calculate true positive rate and false positive rate through which we are able to derive the ROC curves for the proposed model. The ROC curve represents the proposed model on the two datasets and tests the multiclass SVM classifier used in the model.

$$TP = (TP/TP + FN)$$

$$FP = (FP/FP + TN)$$

The steepness of the curve is ideal to maximize the true positive rate and minimizes the false positive rate. Quantitative results on the IDRiD and Kaggle Aptos datasets. The reported results compare the parameters accuracy, sensitivity, and precision between the existing models and proposed model, respectively.

Table 1 shows the comparison between the accuracy, sensitivity and precision in two different datasets. It clearly shows that our proposed model has better results than the earlier existing models. The highest accuracy achieved in IDRiD dataset is

**Table 1** Quantitative results on the IDRiD and Kaggle Aptos datasets

IDRID DATASET			
	Accuracy (%)	Sensitivity	Precision
Eman AbdelMaksoud [7] (2020)	90.2	—	—
Zhan Wu [8] (2020)	56.19	64.21%	—
Balazs Harangi (2019)	90.07	—	—
Dr. D. K. Kirange (2020)	74.46	—	78.3%
Farrikh Alzami (2019)	91.6	—	—
Proposed Model	93.3	78.5%	78.4%

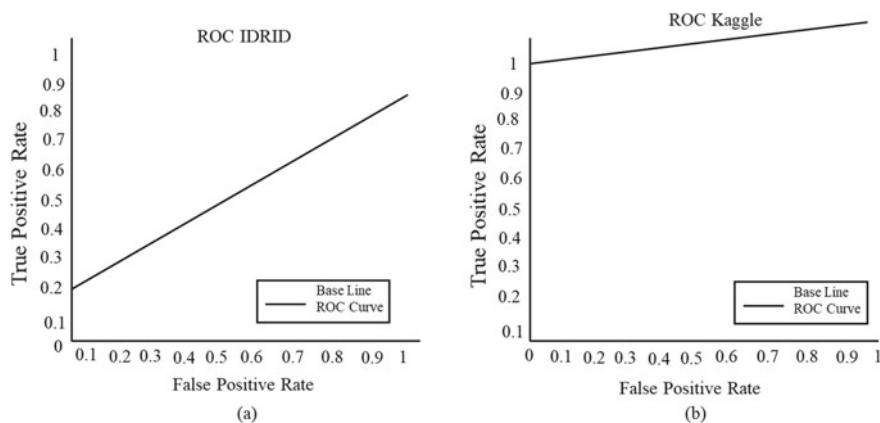
  

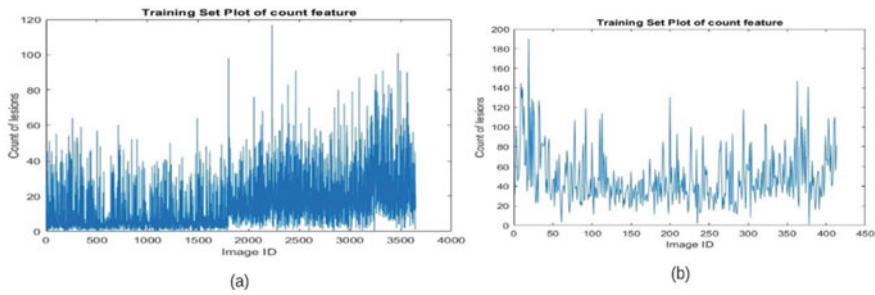
KAGGLE DATASET			
	Accuracy (%)	Sensitivity	Precision
Sara Hosseinzadeh Kassani (2019)	83.09	—	—
Akhilesh Kumar Gangwar (2020)	82.18	—	—
Mohamed Shaban (2020)	88	—	—
Proposed Model	94.5	80.6%	75.6%

The reported results compare the parameters accuracy, sensitivity, and precision between the existing models and proposed model, respectively

93.3% and in KAGGLE dataset is 94.5%. This shows our model can perform and show better results.

Figure 4a, b depict the ROC curves for datasets IDRiD and KAGGLE APTOS. ROC curve for IDRiD dataset for the predictive probability using the TP, TN, FP, and FN. ROC curve for KAGGLE APTOS dataset for the predictive probability using the TP, TN, FP, and FN. ROC graph is used for distinguishing the given classes, in terms of the predicted probability of the multiclass SVM classifier applied. Since multiclass SVM is a discrete classifier, the output graph is a straight line instead

**Fig. 4** **a** ROC Curve for IDRiD dataset, **b** ROC Curve for KAGGLE APTOS dataset



**Fig. 5** **a** Training set plot of count feature for IDRiD, **b** Training set plot of count feature for KAGGLE APTOS

of a curve. The parameters of TP, TN, FP, and FN were derived by measuring the accuracy of test data.

A graph to represent the count for the features per image for training set of both datasets is plotted as shown in Fig. 5a, b for the datasets IDRiD and KAGGLE APTOS. An automated algorithm using for loop is applied to the number of lesions extracted for each image in Sect. 3.2 and plotted against the image ID (the image for which the lesion was extracted), represented in as a graph as Fig. 5a, b for IDRiD and KAGGLE APTOS datasets, respectively.

## 5 Conclusion

This paper has proposed an easy and effective method for the detection of diabetic retinopathy using red lesions in fundus camera pictures. In the proposed method, we apply preprocessing technique of adaptive histogram equalization and morphological operations to improve the picture contrast using the green channel of the image. Gaussian kernel of the matched filtering is applied for the extraction of retinal blood vessels and an adaptive threshold of 0.15 to filter the veins of the eye. The features extracted by the candidate lesion are used to train the multiclass SVM classifier, and the performance on the testing images is better than the existing methods of red lesion detection. We intend to extend our work on detection of diabetic retinopathy to severity grading of DR and add more output parameters. We also anticipate to work on detecting DR using bright lesion method of detection.

## References

1. B. Harangi, J. Toth, A. Baran, A. Hajdu, Automatic screening of fundus images using a combination of convolutional neural network and hand-crafted features, in *41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)* (2019)

2. D.K. Kirange, J.P. Chaudhari, K. P. Rane, K.S. Bhagat, N. Chaudhri, Diabetic retinopathy detection and grading using machine learning. *Int. J. Adv. Trends Comput. Eng.* **8**(6), 3570–3576 (2019)
3. F. Alzami, R.A. Megantara, A.Z. Fanani, Abdussalam: diabetic retinopathy grade classification based on fractal analysis and random forest, in *International Seminar on Application for Technology of Information and Communication (iSemantic)* (2019)
4. A.K. Gangwar, V. Ravi, Diabetic retinopathy detection using Transfer Learning and Deep Learning, in *Evolution in Computational Intelligence*, pp. 679–689 (2020)
5. N.T. Le, J.-W. Wang, C.-C. Wang, T.N. Nguyen, Automatic defect inspection for coated eyeglass based on symmetrized energy analysis of color channels. *Symmetry* **11**, 1518 (2019). <https://doi.org/10.3390/sym11121518>
6. B. Muthu, C.B. Sivaparthipan, G. Manogaran, R. Sundarasekar, S. Kadry, A. Shanthini, A. Dasel, IOT based wearable sensor for diseases prediction and symptom analysis in healthcare sector. *Peer-to-Peer Netw. Appl.* **13**(6), 2123–2134 (2020). <https://doi.org/10.1007/s12083-019-00823-2>
7. M. Shaban, Z. Ogur, A. Mahmoud, A. Switala, A. Shalaby, H.A. Khalifeh, M. Ghazal, L. Fraiwan, G. Giridharan, H. Sandhu, A.S. El-Baz, A convolutional neural network for the screening and staging of diabetic retinopathy. *Public Library of Science ONE* (2020)
8. V.M. Mane, R.B. Kawadiwale, D.V. Jadhav, Detection of red lesions in diabetic retinopathy affected fundus images, in *IEEE International Advance Computing Conference (IACC)* (2015)
9. E.A. Maksoud, S. Barakat, M. Elmogy, A comprehensive diagnosis system for early signs and different diabetic retinopathy grades using fundus retinal images based on pathological changes detection. *Comput. Biol. Med.* **126** (2020)
10. Z. Wu, G. Shi, Y. Chen, F. Shi, X. Chen, G. Coatrieux, J. Yang, L. Luo, S. Li, Coarse-to-fine classification for diabetic retinopathy grading using convolutional neural network. *Artif. Intell. Med.* 108 (2020)

# Video Based Human Gait Activity Recognition Using Fusion of Deep Learning Architectures



P. Nithyakani and M. Ferni Ukrit

**Abstract** Human activity recognition (HAR) plays a vital role in the fields like security, health analysis, gaming, video streaming, surveillances, etc. Many applications were developed based on HAR using video due to its user-friendly nature and affordable cost. We propose the fusion of convolutional layer and Long-Short-Term Memory deep architecture for HAR. In this model, three convolution layer, two BiLSTM layer is used. Convolution layer extracts local features of the frame data. Extracted features were flattened and feed into BiLSTM layers to process the frame sequence and handles the time dependencies. The performance of the model is evaluated using two public datasets namely UCF101 dataset and HMDB-51 dataset and obtained the average accuracy of 96% and 95%, respectively. Influence of hyperparameter was analysed and tuned parameter is used for implementation. The proposed model provides the better accuracy when compared with other deep learning architectures used for HAR using video.

## 1 Introduction

Human activity recognition (HAR) recognize the activities of daily living by processing the signals acquired through wearable sensors, smartphones or cameras. Research direction towards home and abroad has become prevalent due to the development of human computer interface. Individuals could automatically obtain the various facts about their activities by extracting features of ADL (Activities of Daily Living) and classifying them, in succession provides a premise to smart applications. Furthermore, this technology has been widespread in the fields of healthcare, robotics, surveillance, remote monitoring, gesture recognition, gait analysis and gaming.

---

P. Nithyakani (✉) · M. Ferni Ukrit

School of Computing, Department of Information Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai 603203, India  
e-mail: [nithyakp@srmist.edu.in](mailto:nithyakp@srmist.edu.in)

M. Ferni Ukrit

e-mail: [ferniumk@srmist.edu.in](mailto:ferniumk@srmist.edu.in)

Monitoring human activities have become widespread due to the advancement of sensor technology and less cost of devices. In order to maintain the healthy lifestyle, people are interested to log their activities like sleeping, walking, meals, heart beat rate, blood pressure, physical exercises, energy consumption etc. Thereby sensor-based HAR is gaining popularity and it is extensively used with security. Researchers have analysed the various techniques of human activity recognition with video to increase the accuracy. Generally human activity recognition is divided into sensor-based and vision based approaches. Wearable sensor is widely used comparatively to other sensors due to its low cost, low power consumption, high capacity, easy to deploy and imprecise to environmental change. Human activity recognition based on wearable sensors due to its portable nature and favourable reception in everyday life. In accordance with, a huge number of researches have been conducted to examine the capability of wearable sensor in recognizing human activities. Wu et al. [7–9] collected angle and angular velocity data by attaching wearable sensor in waist and thigh. In their study, they have used Fisher discriminant analysis as a hierarchical classifier for pre-fall detection system to classify into 3 types of fall such as non-fall, backward fall and forward fall. Zhang et al. [10] attached wearable sensors in shoulder, elbow, wrist, knee and ankle joints to monitor the human sport activity of soccer and basketball. Gani et al. [11] has used smartphone (inbuilt sensor) to classify the human activities into 11 category using maximum likelihood classifier and Gaussian mixture model.

Vision based approach has proven to provide enhanced recognition accuracy. Moreover, the factor like illumination, occlusion and background challenges the accuracy of vision based human activity recognition.

The paper is organized as follows: Sect. 2 presents the recent video based human activity recognition using machine learning and deep learning architectures. Section 3 provides details about the dataset and pre-processing techniques used for proposed network. Section 4 deals with the proposed architecture of Convolutional LSTM. Section 5 presents the experimental and result analysis. In addition, the impact of hyper parameters is discussed. Section 6 discusses the conclusion and future scope of this paper.

## 2 Related Work

Over the past few years, more research works has been explored various sensing technology for wearable sensor-based human activity recognition [1]. Earlier, researchers have classified the human activities using the traditional machine learning algorithms like support vector machine (SVM), naïve Bayes (NB), K- nearest neighbour (KNN), etc. Jain et al. [14] proposed HAR system which extract the features using histogram of gradient and centroid signature. SVM and KNN classifier is used to classify the features. Hsu et al. [16] proposed nonparametric weighted feature extraction for recognizing human activities and sports activities. In their study, they have used

principal component analysis (PCA) and support vector machine (SVM) for classification. Tian et al. [18] proposed kernel Fisher discriminant analysis to improve the discrimination of feature vector in various human activities using inertial signals. Ronao et al. [19] proposed convent for extracting the features and classifying the raw time series signal received from the smartphone device. In their study, they have reduced the feature complexity and increase the accuracy on motion activities.

Traditional machine learning methods are highly depend on manual feature extraction in recognizing human activities in daily life. This method is usually bounded by human domain knowledge. To resolve these issues, researchers have relied on deep learning approaches to extract adequate features automatically from wearable sensor data during the training stage. Abstract sequence is done in high level whereas original temporal features are extracted in low level. In consideration of flourishing application of deep learning architectures in the field of image classification, speech recognition, music recognition, natural language processing and other fields, researchers are interested to apply deep learning methods for human activity recognition. Jiang et al. [13] proposed activity recognition using DCNN. In their study, they converted the raw signals of accelerometer and gyroscope into activity image and used DCNN for classification. Mario [16] 2018 proposed CNN for human activity recognition. In this study, he trained the CNN with the square acceleration image obtained from the signals of single tri-axial accelerometer.

Shi et al. [20] proposed CNN-RNN network for human action recognition in video stream. In their study, they have used sequential Deep Trajectory Descriptor for extracting Texture images from the video captured. They have implemented CNN-LSTM to learn the long term memory representations. Zhao et al. [21] proposed 1D CNN and 2 CNN with LSTM to recognize speech emotion. In their study, they have used local feature learning blocks and one LSTM Layer for learning long term dependencies. Bai et al. [20] categories the scene image with the coordinated architecture of CNN and LSTM. In their study, CNN is used to extract features from the scene image captured in multi-view and multi-level of abstraction. LSTM is used for classifying the image features in the form of sequence. Wang et al. [23] developed a radar based stacked recurrent neural network (RNN) along with LSTM for recognizing six different human motion. In their study, they have used a Stacked RNN and two 36-cell LSTM layer for classifying the time varying Doppler and micro-Doppler signatures.

LSTM model depends on the previous data for predicting the current status. LSTM runs in only one direction and it is less reliable. Bi-directional LSTM overcomes this problem by referring to the previous input and subsequent information to predict the current status. Hence, we choose BiLSTM network which stores the time series sequence and tracks the dependencies between the current and stored inputs. BiLSTM can learn the time varying signatures which can increase the performance of classification of sequential human activities. Combinations of CNN and BiLSTM architecture has increase the recognition accuracy. In this paper, we propose the state of the art fusion architecture of convolutional operation with BiLSTM to automatically extract features from the frames of video dataset and classify the human activities with few

parameters. The experimental results high accuracy with reduced parameters and also increased the generalization ability and the processing speed.

### Contribution of the paper.

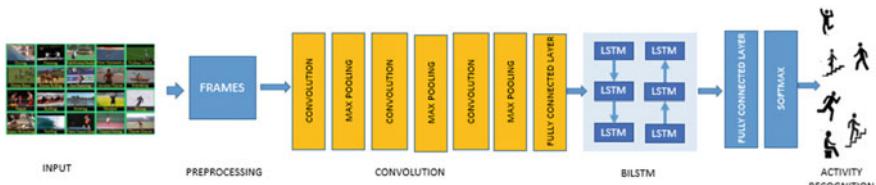
- (1) Fusion of flatten convolution with BiLSTM is proposed for human activity recognition using video.
- (2) The proposed approach reduced the dependencies of handcrafted features by extracting the features automatically.
- (3) The proposed fusion architecture yield better accuracy than the methods used recently for human activity recognition.

## 3 Proposed Architecture

We proposed the fusion of CNN and BiLSTM network to recognize the human activities. In general, fusion of deep learning architectures is used to extract the spatial features and the temporal features. In this work, CNN extracts the spatial features and BiLSTM extract the temporal features. Video frames are fed into three-dimensional convolutional neural network for feature extraction. The extracted feature vector is fed into BiLSTM to further extraction of temporal feature. The obtained spatial-temporal feature is flattened into 1D feature vector and Softmax classifier gives the probability of activities. The overall structure of the proposed model is shown in the Fig. 1. In this section, the details of convolutional neural network and Bi-directional LSTM is discussed.

### Convolutional Neural Network:

Convolutional neural network (CNN or ConvNet) is a deep learning architecture which extracts local features and associate them to develop the complex features to provide exorbitant classification. CNN has more popularity in image classification. In CNN architecture, convolution layer plays an important role in convolving the inputs using convolution kernel whereas pooling layers reduce the dimensionality of the feature map obtained from the previous convolution layer. Equation (1) shows how forward propagation takes place in CNN with  $\omega$  as filter ( $n \times n$  dimension) and  $\psi$  as non-linearity weight matrix. Equation (2) illustrates the gradient of each weight. Equation (3) represents the weight suitable for constructing a convolution layer.



**Fig. 1** The overall structure of Convolution BiLSTM network for HAR using video

$$x_{ij}^l = \psi \left( \sum_{a=0}^{n-1} \sum_{b=0}^{n-1} \omega_{ab} y_{(i+a)(j+b)}^{l-1} \right) \quad (1)$$

$$\frac{\partial E}{\partial x_{ij}^l} = \frac{\partial E}{\partial y_{ij}^l} \frac{\partial y_{ij}^l}{\partial x_{ij}^l} = \frac{\partial E}{\partial y_{ij}^l} \frac{\partial}{\partial x_{ij}^l} (\psi(x_{ij}^l)) \quad (2)$$

$$\omega_{ab} = J \left( \sum_{a=0}^{n-1} \sum_{b=0}^{n-1} \frac{\partial E}{\partial x_{(i-a)(j-b)}^l} \frac{\partial x_{(i-a)(j-b)}^l}{\partial y_{ij}^{l-a}} \right) \quad (3)$$

$$J = \left( \sum_{a=0}^{n-1} \sum_{b=0}^{n-1} \frac{\partial E}{\partial x_{(i-a)(j-b)}^l} \right)^{-1} \quad (4)$$

To obtain the spatial feature, weights are given according to the layers in the model. Each layer of the model is fully connected. In order to feed the feature vector resulted with CNN into BiLSTM layer, vector is converted 3-dimensional shape.

In this study, video frame is given as the input to the Convolution BiLSTM model. Temporal features are extracted using convolution operation consisting of three convolution layer and its max pooling layer. The filters used for the convolutional layer is 128, 256 and 512. Each convolution layer has Rectified Linear Unit (ReLU) activation function Max pooling layer is used after each convolution to reduce its feature dimension. Fully connected layer process the flattened input into 3 dimension and feeds it to BiLSTM layer. BiLSTM solves time dependence issue in time series data. BiLSTM learns the temporal features and its output to feed to fully connected layer (FC Layer). Softmax is used as a classifier to predict the probability of the human activity.

## 4 Dataset

The description of two public dataset for video based HAR is given in Table 1.

**Table 1** Dataset description

DATASET	HMDB 51 dataset	UCF 101
Video clips	6766	13,320
No. of activities	51	101
Source	YouTube	Movie Clips, YouTube and Google video
Training set	70 video clips	14,900
Test set	30 video clips	6360

## 5 Experimental Results

For this study, UCF101 dataset and HMDB dataset were used to evaluate the performance and accuracy of the proposed Convolution BiLSTM model. Both the dataset values were recorded with the same methodology in continuous manner. The video is converted into frames. In order to reduce the overfitting, data augmentation is used. 70% of the dataset is allotted for training the model whereas remaining 30% of dataset is used for testing the model. Categorical cross entropy is used to evaluate the loss in predicting the human activity. Adam optimizer which utilized the Adam stochastic optimization algorithm to compute the learning rates from the first order gradient (Beta 1-0.9) and second order gradient (Beta 2-0.99) for various parameters. Batch size is set to 128 and number of epochs is 120. The value of dropout is 0.02. The confusion matrix and performance measures like precision, recall and F1 score is in given Tables 1 and 2 for UCF101 and HMDB dataset, respectively, with selected activities. The proposed model Convolution BiLSTM have predicted accurately with 96% of accuracy for UCF101 dataset. In HMDB dataset, proposed model have acquired 95% accuracy. Performance of the proposed model Convolution BiLSTM over UCF101 and HMDB dataset is shown in the Figs. 2 and 3, respectively, (Table 3).

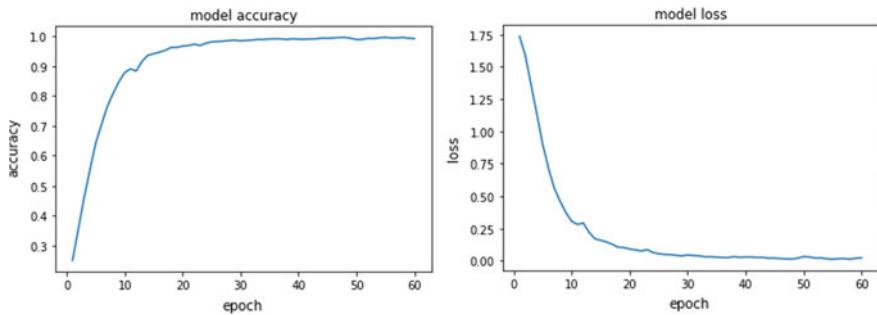
Hyperparameters greatly influence the performance of the classification model. In this study, the parameters such as learning rate, kernel size, dropout rate and batch normalization are evaluated with variation over UCF-101 dataset. Kernel size with variations of [3, 5] and dropout rate ranging [0: 0.2]. With and without Batch normalization, the model is executed and analysed. The best accuracy is achieved with kernel size 3, with batch normalization and dropout rate 0.02. Learning rate with 0.0014 provides the highest accuracy and the variations of learning rate with accuracy is given in the Table 4.

## 6 Conclusion

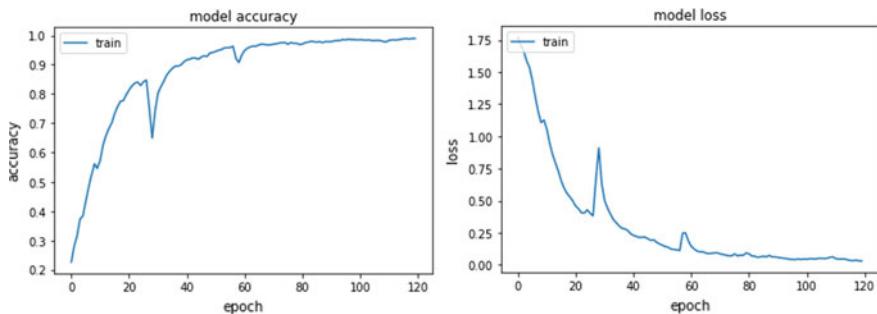
Fusion of convolutional neural network and BiLSTM have achieved better performance in image classification, speech recognition and object detection. Though, various deep architectures algorithms were used for HAR with video, fusion of deep architectures have not been implemented. The proposed method Convolution BiLSTM combines the convolution operation (convolution + pooling) and BiLSTM to recognize the human activity using captured video. Spatial features of video frame were extracted using the convolution operation whereas temporal features were extracted by BiLSTM. Time dependencies problem is solved by BiLSTM layer. The model was experiment with UCF-101 dataset and HMDB 51 dataset. The experimental results proves the state of the art Convolution BiLSTM gives better accuracy when compared with convolutional neural network. The performance of the model is evaluated with precision, recall and F1 score. The performance measure provides

**Table 2** Confusion Matrix and performance measure (precision, Recall, F1 score) of Convolution LSTM for UCF101 dataset

UCF	Walk front	Golf Swing	Kicking front	Lifting	Riding Horse	Run Side	Precision	Recall	F1 score
Walk front	96	0	4	0	0	0	0.98	0.96	0.97
Golf Swing	2	98	0	5	1	0	0.97	0.98	0.98
Kicking front	0	12	88	0	23	0	0.95	0.88	0.91
Lifting	1	0	1	97	1	0	0.93	0.97	0.95
Riding Horse	3	1	0	5	91	0	0.96	0.91	0.94
Run Side	0	0	1	0	99	0.98	0.99	0.98	



**Fig. 2** Model accuracy and loss of UCF 101 HAR dataset using Convolution BiLSTM



**Fig. 3** Model accuracy and loss of HMDB-51 dataset using Convolution BiLSTM

**Table 3** Confusion Matrix and performance measure (precision, Recall, F1 score) of Convolution LSTM for HMDB-51 dataset

HMDB-51	Jump	Climb Stairs	Kick	Tennis Swing	Jump Rope	Diving	Precision	Recall	f1 score
Jump	94	0	3	1	2	0	0.96	0.94	0.95
Climb Stairs	4	93	0	2	1	0	0.95	0.93	0.94
Kick	0	2	95	3	0	0	0.96	0.95	0.95
Tennis Swing	0	3	1	96	0	0	0.93	0.96	0.94
Jump Rope	4	1	0	2	93	0	0.96	0.93	0.94
Diving	0	0	0	0	1	99	0.96	0.99	0.97

the average of 96% and 95% recognition rate for UCF-101 dataset and HMDB 51 dataset, respectively. Furthermore, the impact of hyper parameters learning rate, kernel size, and dropout and batch normalization is explored and the optimum values

**Table 4** Effects of Learning rate on the performance of Convolution LSTM in UCF 101 dataset

Learning rate	Accuracy (%)
0.0017	93.6
0.0016	94.2
0.0015	94.9
0.0014	96.7
0.0013	95.4

were implemented in the proposed model. Moreover, we target to utilize the deep learning architectures on larger dataset for HAR.

## References

- Y. Wu, Y. Su, R. Feng, N. Yu, X. Zang, Wearable-sensor-based pre-impact fall detection system with a hierarchical classifier. *Measurement* **140**, 283–292 (2019)
- J. Zhang, Y. Cao, M. Qiao, L. Ai, K. Sun, Q. Mi, Q. Wang et al., Human motion monitoring in sports using wearable graphene-coated fiber sensors. *Sens. Actuators A* **274**, 132–140 (2018)
- M.O. Gani, T. Fayezeen, R.J. Povinelli, R.O. Smith, M. Arif, A.J. Kattan, S.I. Ahamed, A light weight smartphone based human activity recognition system with high accuracy. *J. Netw. Comput. Appl.* (2019)
- W. Jiang, Z. Yin, Human activity recognition using wearable sensors by deep convolutional neural networks, in *Proceedings of the 23rd ACM International Conference on Multimedia - MM '15* (2015)
- A. Jain, V. Kanhangad, Human activity classification in smartphones using accelerometer and gyroscope sensors. *IEEE Sens. J.* **18**(3), 1169–1177 (2018)
- Y.-L. Hsu, S.-C. Yang, H.-C. Chang, H.-C. Lai, Human daily and sport activity recognition using a wearable inertial sensor network. *IEEE Access* **6**, 31715–31728 (2018)
- S.-I. Chu, B.-H. Liu, N.-T. Nguyen, Secure AF relaying with efficient partial relay selection scheme. *Int J Commun Syst.* **32**, e4105 (2019). <https://doi.org/10.1002/dac.4105>
- M. Balaanand, N. Karthikeyan, S. Karthik, R. Varatharajan, G. Manogaran, C.B. Sivaparthipan, An enhanced graph-based semi-supervised learning algorithm to detect fake users on Twitter. *J. Supercomput.* **75**(9), 6085–6105 (2019). <https://doi.org/10.1007/s11227-019-02948-w>
- Y. Tian, X. Wang, L. Chen, Z. Liu, Wearable sensor-based human activity recognition via two-layer diversity-enhanced multiclassifier recognition method. *Sensors* **19**(9), 2039 (2019)
- C.A. Ronao, S.-B. Cho, Human activity recognition with smartphone sensors using deep learning neural networks. *Expert Syst. Appl.* **59**, 235–244 (2016)
- M.-O. Mario, Human activity recognition based on single sensor square hv acceleration images and convolutional neural networks. *IEEE Sen. J.* **1**–1 (2018)
- Y. Shi, Y. Tian, Y. Wang, T. Huang, Sequential deep trajectory descriptor for action recognition with three-stream CNN. *IEEE Trans. Multimedia* **19**(7), 1510–1520 (2017)
- J. Zhao, X. Mao, L. Chen, Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomed. Signal Process. Control* **47**, 312–323 (2019)
- S. Bai, H. Tang, S. An, *Coordinate CNNs and LSTMs to Categorize Scene Images with Multi-views and Multi-levels of Abstraction. Expert Systems with Applications* (2018)
- M. Wang, Y.D. Zhang, G. Cui, Human motion recognition exploiting radar with stacked recurrent neural network. *Digital Signal Processing* (2019)

# A Deep Learning Based Palmar Vein Recognition: Transfer Learning and Feature Learning Approaches



M. Rajalakshmi and K. Annapurani

**Abstract** A method of palm dorsal vein recognition using transfer learning and feature learning approaches has been proposed in this paper. This deep learning framework does the learning process automatically so that the features are extracted from the original image without undergoing any preprocessing mechanisms. The system proposed is dealing with two methods of learning: The first method employs the pre-trained models of CNN (AlexNet, VGG19, ResNet50 and ResNet101) for feature extraction from the deeper fully connected (fc6, fc7 and fc8) layers. K-Nearest Neighbour (KNN) and Support Vector Machine (SVM) algorithms are used for classification with combination of Error-Correcting Output Codes (ECOC). The later method employs the transfer learning approach for the extraction and classification of both features with CNN (AlexNet, ResNet50, ResNet101 and VGG19) models. The experiments have been done with Dr. Badawi's dataset of hand veins containing 500 images. In the first method, all the models are given the recognition accuracy that gives promising results when features are extracted from layers of 'fc6'. Using ECOC with SVM for classification shows greater accuracy rate than the models using ECOC with KNN. Also in models using ECOC with SVM, the ResNet101 model achieves improved performance. The recognition accuracy for all models provides the best result in the experimentation of the second approach when the epoch number is 25 where ResNet101 achieves a 100% recognition rate.

**Keywords** Authentication · Patterns recognition · Palm dorsal vein · KNN · SVM · CNN · Transfer learning

---

M. Rajalakshmi (✉) · K. Annapurani

School of Computing, SRM Institute of Science and Technology, Kattankulathur, India  
e-mail: [rajalakm2@srmist.edu.in](mailto:rajalakm2@srmist.edu.in)

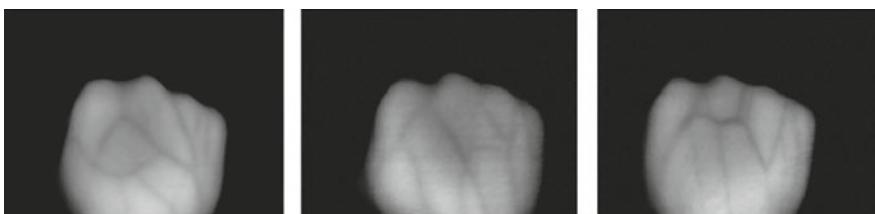
K. Annapurani  
e-mail: [annapook@srmist.edu.in](mailto:annapook@srmist.edu.in)

## 1 Introduction

Biometric systems are the invention of technology that can be used uniquely and effectively to recognize individuals in automated systems. Traditional authentication systems like swiping cards, passwords, keywords, keys and smart cards offer only less security and believed to be unreliable. The cards or keys may be lost or may be stolen by unknown persons. Passwords might be revealed to unauthorized persons. As an better and effective alternative to overcome the security threats which is expected to be in an authentication or recognition systems, biometrics field has been extensively explored to improve the reliability of personal authentication techniques. The biometric traits falls into two broad classes: Physiological characteristics includes iris, palm print, retina, fingerprint, face, ear, hand geometry, finger vein, palm and its dorsal vein etc., and behavioural characteristics includes signature, typing speed, gesture characteristics, key clicks, gait and lips, etc. [1]. Biometric modalities always possess any one of the following properties as per the literature studies:

- Distinct: A trait should be unique from one person to other person.
- Universality: Every individual should possess that trait.
- Permanence: The trait should remain unchanged.
- Collectability: The trait should be measurable in terms of quantity.
- Acceptability: It specifies a level of acceptability by the general public in their routine life. It should be user-friendly.
- Performance: The speed and accuracy as well as an environmental effect on its performance to gain better recognition.
- Circumvention: The versatility of the system should not get corrupt by outsider attack.

The vein pattern of the hand dorsal surface is presented in Fig. 1. Below the skin surface, blood vessels form a network which is considered as a vein pattern. According to literature [1, 2], vein or vascular patterns are sufficiently distinguished by individuals despite ageing and also observed that there are no major changes in adults throughout his/her lifetime. The vein structures are unique even between two identical persons (twins). Due to the recent advancements of deep learning technology, which has the special capability of extracting the features automatically, it suits several applications like natural language processing, recognition systems



**Fig. 1** Hand vein images [3]

and identification systems. Convolution Neural Network (CNN) was discovered in 1990 by LeCun et al. [4], which well suits all image-based identification and recognition systems. In this paper, a palmar vein recognition system-based on dorsal venal network of one's left or right hand using transfer learning and feature learning approaches has been proposed.

The remaining paper is organized as follows: Section 2 explains the different methods of dorsal hand vein recognition as mentioned in the literature. Section 3 provides the framework of the proposed CNN method. Section 4 details about the experimentation results of the proposed recognition system. Finally, Sect. 5 discusses the conclusion.

## 2 Survey of Literature Works

Due to the inherent stability feature of hand vein patterns, the vascular structure remains constant throughout one's lifetime, vein based biometrics prove to be promising and accurate compared to other biometric traits. Few literatures involving the hand vein pattern as its traits are discussed in this survey.

Huang et al. [5] proposed a method which integrates universal and native analysis then joint hierarchically. The results obtained are promising to increase the efficiency of the system. Lee et al. [6] proposed a directional filter approach which encodes features of one's hand vein into binary code by adopting a method called minimum directional filtering response (MDFR). Later, the classification is carried out using Hamming Distance (HD). In this paper, the major advantage is that non-venal regions are identified by calculating the minimum filtering modification which achieves high accuracy. Trabelsi et al. [7] suggested a new descriptor Circular Difference and Statistical Directional Patterns (CDSDP) which extracts the venal map structure and classification being carried out through ANN and FMNN.

Yun et al. [8] proposed a new hand vascular system based on feature points from which the reference points are extracted. The extracted features are then identified by the distance measures between the two feature points and their angles. Finally, the feature points along with image translation and rotation are combined which shows improved accuracy. Chuang et al. [9] suggested a process depending on minutiae features extracted from vein networks. The minutiae points comprise bifurcation and end points. Also a dynamic pattern tree (DPT) has been suggested which speeds up the matching process. Zhu et al. [10] uses the texture along with the features of hand geometry in their proposed method. It identifies the region of interest where the main vein area lies, then computes its vein map by the thinning process. Then texture and geometry clues are combined lastly for final decision.

Wan et al. [11] proposed a framework based on deep learning for the identification of vein patterns at the back of hand, which relies on CNN. The input image is pre-processed to segment its region of interest (ROI) and then applied with noise filtering techniques like histogram equalisation and Gaussian smoothing filter. CNN models like Reference-CaffeNet, AlexNet and VGG are used for Feature Extraction

and logistic regression for classification. The proposed recognition method automatically learns to extract features from the original input image by not applying any preprocessing techniques using deep learning framework, as opposed to literature.

### 3 Proposed Framework for Palm Dorsal Vein Recognition System

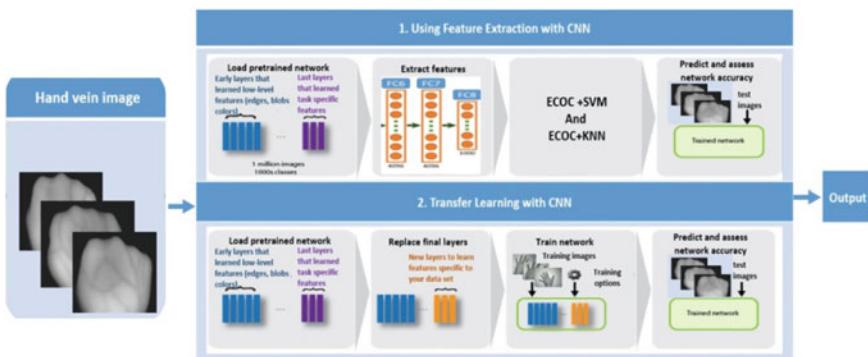
The proposed vein recognition system uses CNN models like (AlexNet, ResNet50, ResNet101 and VGG19) models. This system explains two methods as shown in Fig. 2:

- Utilizing the Pre-trained CNN models: For extraction of features KNN and SVM algorithm together with Error-Correcting Output Codes (ECOC) is utilized for classification.
- Transfer learning with CNN: The pre-trained CNN models such as AlexNet, ResNet50, VGG19 and ResNet101 are used for extracting features and also for doing classification.

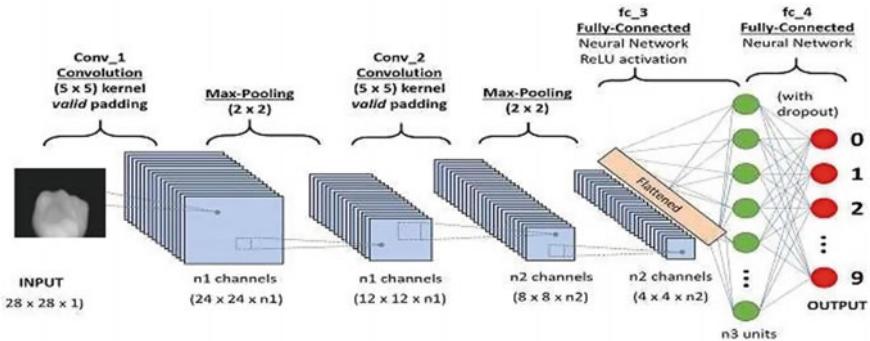
#### 3.1 Feature Learning with CNN Approach

In this paper, the deep CNN approach as suggested earlier in 1988 by Fukushima has been used. CNN is mainly employed in the processing of the images. The general structure of a CNN has two key sections: feature extractors (Convolution and Pooling layers) and a classifier, as shown in Fig. 3.

Both layers in the network get the input from their immediate previous layer during feature extraction and send their output to the following layer as the input



**Fig. 2** The proposed framework of palm dorsal vascular patterns recognition system

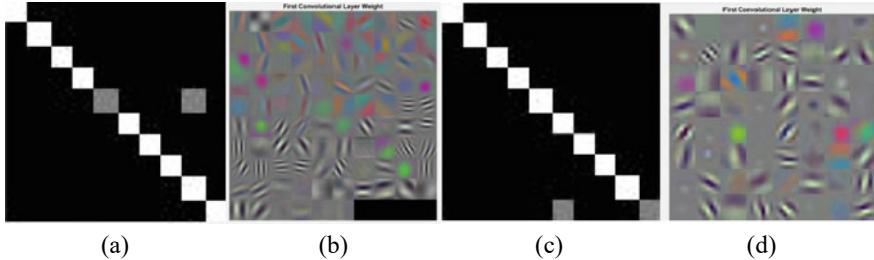


**Fig. 3** Generalized representation of a CNN Network

[12]. The highest classification layer determines individual class weight, which determines results based on the best weight for the corresponding classes [13]. The CNN Network's convolutional layer is at the heart of the network's name. Convolution operates on input data and 2D weight data in the form of arrays to perform linear multiplication. The Convolutional Layer, Pooling Layer and Fully Connected Layer are the layers that make up the CNN architecture [14].

- **Convolutional Layer:** The convolutional layer takes the input image of a hand vein and applies filters to extract feature maps, resulting in many feature maps. The change in features depends on the filter that has been added. Low-level features such as thresholds and angles are obtained from the primary convolution layer. The last layers have higher level characteristics.
- **Pooling Layer:** This layer decreases the image size and its solution where depth of the image remains unchanged. It also does features' creation which are robust to noise. There are two ways in which the pooling can be done: max pooling and average pooling. Usually max pooling is preferably used.
- **Activation functions:** The input is mapped to a set of values in which the output of the layer that comes before it falls within that set of values. Sigmoid, ReLU and Tanh are the commonly used activation functions. Since ReLU trains the network faster, it is applied here.
- **Fully Connected Layer:** Features of previous layer is identified in this last layer of a CNN.

Pre-trained CNN models such as AlexNet, ResNet50, ResNet101 and VGG19 are used [15–17] in this work instead of building a CNN model from scratch. Fully connected layers extracts the features required for classification. Figure 4 illustrates the confusion matrix and Convolutional layer weight obtained from AlexNet and ResNet101 models, respectively.



**Fig. 4** **a, b** Confusion matrix and CL weight of AlexNet, **c, d** Confusion matrix and CL weight of ResNet101

### 3.2 ECOC Framework

In this model, the weights of the final output are classified using ECOC with K-Nearest Neighbour (KNN) and support vector machine (SVM), as described in Sect. 3. ECOC is a most widely used classification framework for which there is no necessity to be associated with any specific method of classification. In this work, KNN and SVM classification methods are used with ECOC. ECOC comprises the following two main phases: encoding and decoding. The encoding method entails the development of a Coding Matrix (CM). Binary classifiers and code words for classes are used to represent CM in columns and rows, respectively. Binary or ternary coding may be used to execute certain interface tasks [5]. In former method, the coding matrix elements which are arranged in rows and columns either belong to the set  $\{1, -1\}$  or to  $\{0, 1\}$ , respectively. Here,  $M$  denotes the number of classes and  $N_c$  is the number of binary classifiers [18]. The one-vs-all (OVA) strategy has been chosen for this work to construct the coding matrix [18] which uses binary coding.

This approach is also used with transfer learning. The transfer learning involves the usage of a pre-trained model which is already trained on some other input images. Such a model is trained to classify a different set of images in this work which is expected to provide a good prediction. The pre-trained CNN (AlexNet, ResNet101, ResNet50 and VGG19) acts as an automatic feature extractor from the input image. The matching score is generated by the feature vector and fed into the softmax layer.

A Softmax layer uses a Softmax function [11], Eq. (5), to the input:

$$P(y = j|z^{(i)}) = e^{z^i} / \sum_{j=0}^k e^{z_k^{(i)}}$$

where  $z$  is

$$z = \sum_{j=1}^m w_j y_j + b$$

where  $w$  represents weight vector,  $y$  represents feature vector and  $b$  represents bias unit [19].

## 4 Experimental Results and Discussion

### 4.1 Results After Using Feature Extraction with CNN: First Method

A desktop with (Intel(R) Core (MT) i7 CPU, RAM 16 GB, NVIDIA graphics card specification and win10 64-bit operating system is used for this work. Deep learning toolbox of MATLAB R2020a is used for implementation.

After extracting the higher and lower level features from the deeper convolutional layers and from the first few layers, respectively, a hierarchical representation of input hand vein images is derived from CNN. The features were extracted from both the training and testing images from the deeper fc6, fc7 and fc8 layers in this paper using the activation functions. Because of the size difference in the input image, it is resized before being fed as an input. The resized images are initially stored in augmented image data stores, which are then used as input arguments to activation functions. Features derived from the ‘FC6’, ‘FC7’ and ‘FC8’ layers of all models are used to train the ECOC with SVM and KNN classifiers. The used hand vein database is divided into three categories: 60–40%, 70–30% and 80–20%. Tables 1 and 2 display the classification accuracy of ECOC with SVM and ECOC with KNN, respectively (Table 3).

The KNN with AlexNet model has the highest accuracy, ranging from 88.5% to 98.5%, according to Table 1. The 80–20 per cent split provides the most consistency. The model offers fair accuracy when features are extracted from the ‘fc6’ layer, varying from 97.3% to 98.5%. The KNN with AlexNet model has the highest accuracy, ranging from 88.5 per cent to 98.5%, according to Table 1. The 80–20%

**Table 1** Recognition accuracy of ResNet50, ResNet101, VGG19 and AlexNet models under different split ratio and ECOC with KNN classification

Name of the CNN Model	Feature Extractors								
	Fc6			Fc7			Fc8		
Train versus Test Ratio used									
	80–20	70–30	60–40	80–20	70–30	60–40	80–20	70–30	60–40
ResNet50	97.4	97.1	97	96	96	93.50	91.1	91	88.3
ResNet101	95	95	94.50	92	92	88.500	87	87	76
VGG19	96	96	91.50	93	93	93.500	87	87	86
AlexNet	98.5	98	97.3	97	97	94.50	91	91	88.50

**Table 2** Classification accuracy using ResNet50, ResNet101, VGG19 and AlexNet models under different split ratio and ECOC with SVM classification

Name of the CNN Model	Feature Extractors								
	Fc6			Fc7			Fc8		
	Train versus Test Ratio used								
	80–20	70–30	60–40	80–20	70–30	60–40	80–20	70–30	60–40
ResNet50	93.40	91.50	90.7	93.6	88.7	89.3	87	82.6	81.9
ResNet101	99.8	99.8	99.1	98	98	97.50	96	96	95
VGG19	94	94	93.5	92	92	91.8	89	89	89.5
AlexNet	99.6	99	99	97	97	97	97	97	96

**Table 3** Transfer learning with ResNet50, ResNet101, AlexNet and VGG19 models improves recognition performance

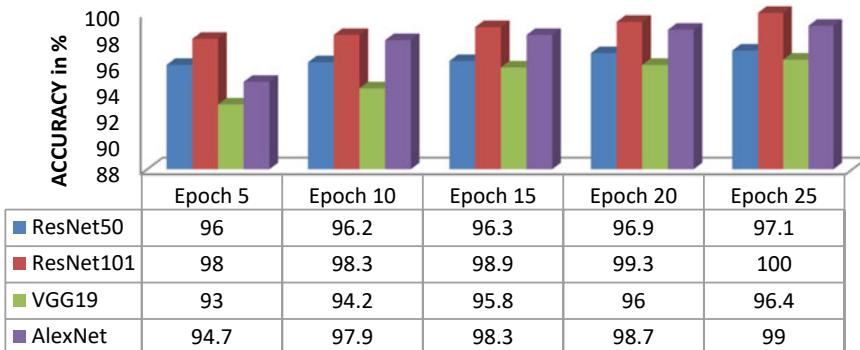
Model Name	Epoch No 5	Epoch No 10	Epoch No 15	Epoch No 20	Epoch No 25
ResNet50	95	96	96.3	96.9	97.1
ResNet101	95.8	98	98.3	98.9	100
VGG19	93	94.2	95.8	96	96.4
AlexNet	94.7	97.9	98.3	98.7	99

split provides the most consistency. The model offers fair accuracy when features are extracted from the ‘fc6’ layer, varying from 97.3% to 98.5%.

Finally, when features from ‘fc6’ are extracted, the SVM and KNN classifiers with ResNet50, ResNet101, AlexNet and VGG19 models achieve the best performance, after which the recognition accuracy decreases. Furthermore, it is clear that ECOC with SVM classifies better than ECOC with KNN, and recognition accuracy of ECOC with SVM classifier gives good results with the ResNet101 model when the split is 80–20%.

#### 4.2 *Results After Using Transfer Learning with CNN: Second Method*

The dataset images are fed into pre-trained CNN (AlexNet, ResNet50, ResNet101, and VGG19) models, which extract the features and perform classification automatically. These models are used to recognize 50 classes in the Badawi dataset in this analysis, where it is already trained to classify 1000 classes. This number is determined by the number of dataset classes, and it ensures that the pre-trained network’s performance fits the new classification assignment. This method involves calculating an epoch number and mini-batch size [23].



**Fig. 5** Recognition accuracy of all models using transfer learning approach from Epochs 5 to 25

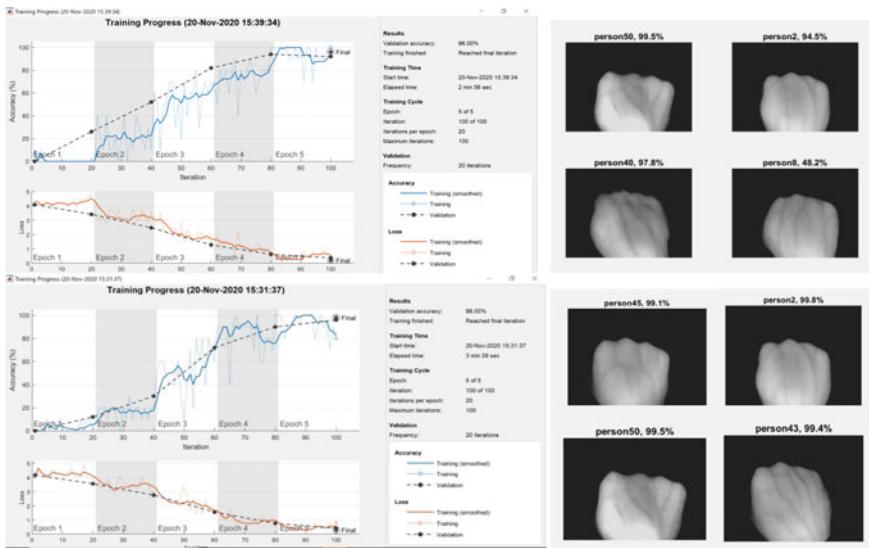
The new dataset was used in this paper with epoch numbers of 5, 10, 15, 20, and 25. The mini-batch size is set to ten, with an initial learning rate of 0.0003. For the dataset used, transfer learning is performed using AlexNet, ResNet50, ResNet101 and VGG19 models after all the parameters have been set. Fig. 5 shows the recognition accuracy achieved by transfer learning. The recognition accuracy ranged from 93% to 100% and ResNet101 and AlexNet gives good accuracy reaches to 100% and 99% recognition rate in 25 epochs.

After evaluating the findings from Tables 1, 2, and Fig. 5, it is clear that for the given dataset, ECOC with SVM has a higher accuracy of 99.8% than ECOC with KNN, which has a 98.5 per cent accuracy, and transfer learning with CNN, where the classification using SVM approach produces 100% accuracy in the ResNet101 Model. The obtained results are depicted in Fig. 6.

## 5 Conclusion

In recent years, biometric systems find its advancement sufficiently well with improved recognition rates, but the robustness, reliability and user friendly requirements still offer so much space for improvement [1]. Authentication systems based on venal patterns appear as a well advanced approach where their features in the back of hand are affirmed to be unique and stable (last for long term) even for the genetically identical twins [2].

This paper suggested a two-method vein recognition scheme for the back of the hand. The first approach uses pre-trained models such as AlexNet, ResNet50, ResNet101 and VGG19 to extract features, which are then classified using ECOC with KNN and SVM. The findings revealed that features derived from the fully linked layer 'fc6' improve all models' recognition accuracy. It also concludes that models using ECOC with SVM have a higher accuracy rate than models using ECOC with KNN. In ECOC with SVM models, the ResNet101 model outperformed the others. In



**Fig. 6** Graph of accuracy and loss function for transfer learning using ResNet101 when epoch is 5

ECOC with SVM models, the ResNet101 model outperformed the others. When the epoch number is 25, the second method blends transfer learning with CNN models for feature extraction and classification, resulting in the highest recognition accuracy for both models, with ResNet101 and AlexNet achieving 100% and 95% recognition rates, respectively.

**Acknowledgements** The Authors would like to thank Prof. Ahmed M. Badawi, Professor and Head of Biomedical Engineering, Cairo University, Egypt for providing us with the palm- Dorsal hand vein database which consists of 50 subjects (250 images in total).

## References

1. G. Neha, S. Tuly, Palmprint recognition: a selected review. *Int. J. Eng. Sci.* **6**(6), 6776–6780 (2016)
2. B. Sontakke, V. Humbe, P. Yannawar, Dorsal hand vein authentication system: a review. *Int. J. Sci. Res. Eng. Technol.* **6**(5), 511–514 (2017)
3. M. Shahin, A. Badawi, M. Rasmy, Multimodal biometric system based on near-infra-red dorsal hand geometry and fingerprints for single and whole hands. *World Acad. Sci. Eng. Technol.* **4**(4), 268–283 (2010)
4. Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, in *Proc. IEEE*, vol. 86, No. 11 (1998), pp. 2278–2324
5. D. Huang, X. Zhu, Y. Wang, D. Zhang, Dorsal hand vein recognition via hierarchical combination of texture and shape clues. *Neurocomputing* **214**, 815–828 (2016)
6. J. Lee, T. Lo, C. Chang, Dorsal hand vein recognition based on directional filter bank. *Signal Image Video Process.* **10**(1), 145–152 (2016)

7. R. Trabelsi, A. Masmoudi, D. Masmoudi, Hand vein recognition system with circular difference and statistical directional patterns based on an artificial neural network. Springer Multimedia Tools Appl. **75**(2), 687–707 (2014)
8. Y. Hu, Z. Wang, X. Yang, Y. Xue, Hand vein recognition based on the connection lines of reference point and feature point. Infrared Phys. Technol. **62**(1), 110–114 (2013)
9. S. Chuang, Vein recognition based on minutiae features in the dorsal venous network of the hand. Signal Image Video Process. **12**(3), 1–9 (2018)
10. X. Zhu, D. Huang, Hand dorsal vein recognition based on hierarchically structured texture and geometry features, in *Proceedings of Chin. Conference Biometric Recognition* (2012), pp. 157–164
11. H. Wan, L. Chen, H. Song, J. Yang, Dorsal hand vein recognition based on convolutional neural networks, in *Proceedings of IEEE International Conference. Bioinformatics Biomedicine* (2017), pp. 1215–1221
12. M. Alom, T. Taha, C. Yakopcic, S. Westberg, M. Hasan, B. Van Esen, A. Awwal, V. Asari, *The History Began From Alexnet: A Comprehensive Survey on Deep Learning Approaches*. arXiv preprint [arXiv:1803.01164](https://arxiv.org/abs/1803.01164) (2018)
13. M. Alaslani, L. Elrefaei, Convolutional neural network based feature extraction for iris recognition. Int. J. Comput. Sci. Information Technol. **10**(2) (2018)
14. S. Hijazi, R. Kumar, C. Rowen, *Using Convolutional Neural Networks for Image Recognition* (Cadence Design Systems Inc, San Jose, CA, USA, 2015)
15. A. Krizhevsky, I. Sutskever, G. Hinton, Imagenet classification with deep convolutional neural networks. Proc. Adv. Neural Inf. Process. Syst. 1097–1105 (2012)
16. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. Computer Science (2015)
17. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, S. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 1–9 (2015)
18. H. Joutsijoki, M. Henry, J. Rasku, K. Aalto- Setälä, M. Juhola, Error-correcting output codes in classification of human induced pluripotent stem cell colony images, in *BioMed Research International* (2016), pp. 1–13
19. M. Nasrabadi, Pattern recognition and machine learning. J. Electron. Imaging **16**(4), 140–155 (2006)
20. The SUAS Database. <http://SUAS.ee.boun.edu.tr/>. Accessed on 1/8/2018
21. A. Badawi, Hand vein biometric verification prototype: A testing performance and patterns similarity, in *Proc. Int. Conf. Image Process. Comput. Vis. Pattern Recognit.* (2006), pp. 3–9
22. M. Shahin, A. Badawi, M. Kamel, Biometric authentication using fast correlation of near infrared hand vein patterns. Int. J. Biomed. Sci. **2** (2007)
23. W. Kim, S. Jong, P. Kang, Multimodal biometric recognition based on convolutional neural network by the fusion of finger-vein and finger shape using Near-Infrared (NIR) camera sensor. Sensors **18**(7), 1–34 (2018)

# Survey of Popular Linear Dimensionality Reduction Techniques



Anne Lourdu Grace and M. Thenmozhi

**Abstract** Big Data analytics related solutions are one of the prime industrial focuses across all domains. In this digital era, the volumes of the data being generated by both machine and man are humongous. The key challenges are to store the data which requires more space and also to retrieve the data in an optimized way that saves time and money. The number of features for any given dataset is another problem statement. As in most real-life datasets, the number of features present is more and we are not sure, which features or dimensions are good enough to be considered. Visualization of the high dimensional data is also one of the most complex issues that impair meaningful insights since the visuals are not precise anymore due to feature explosion. In this paper, we have done a comprehensive survey on some of the most popular dimensionality reduction approaches broadly categorized under linear dimensionality reduction techniques. We have also done a critical comparative analysis focused on the utility of these algorithms. Finally concluded the survey with few notes on future work.

**Keywords** Dimensionality reduction · Feature extraction · Feature reduction · Factor analysis · Principal Component Analysis · High dimensional data

## 1 Introduction

Large volumes of data being generated in both batch and streaming form are truly relentless across every digital medium and especially evident in the domains such as Banking, Insurance, Healthcare, Telecom and Manufacturing [1, 2]. These voluminous real-world data can essentially be categorized into three classes as structured (Database), semi-structured (XML, Jason) and unstructured (Image, Video

---

A. L. Grace (✉)

Department of Computer Science, SRM University, Chennai, India

e-mail: [ag3512@srmist.edu.in](mailto:ag3512@srmist.edu.in)

M. Thenmozhi

Department of Information Technology, SRM University, Chennai, India

e-mail: [thenmozm@srmist.edu.in](mailto:thenmozm@srmist.edu.in)

and Audio) data [2]. As a result, the volume, velocity and variety of data keep growing exponentially which further increases the complexity of storage, retrieval and security [1].

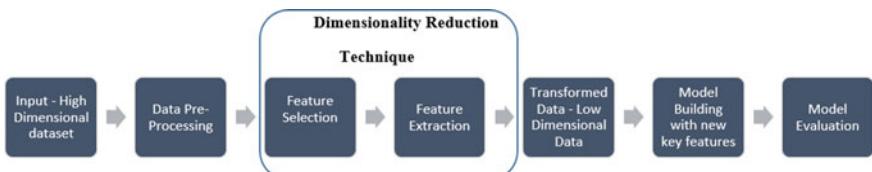
Most of the real-world analytics use cases tend to have a very large number of features which essentially does not mean that every feature is important, thus creates a definite opportunity to reduce these features or dimensions [2, 3]. The key challenge in reducing dimension is to accurately identify essential features amidst the higher possibilities of having a low variance, outliers, missing values and redundant data. Moreover, any treatment performed to handle high dimensional data might also increase the complexity in terms of computational time, storage capacity, maintenance and cost [2].

Dimensionality reduction is a technique that helps to transform high dimensional data with several features into a more meaningful dataset by reducing unwanted dimensions or features [3]. As the number of features or dimension increases the complexity to handle data, it also increases the difficulty of building any analytical solutions which ultimately makes the data irrelevant for many applications [4]. Dimensionality reduction facilitates many data-driven tasks like classification, visualization, and compression of high dimensional data [3]. Dimensionality Reduction Technique (DRT) helps to identify and derive the most relevant features from given high dimensional data efficiently before applying it to the machine learning models [5].

## 2 Dimensionality Reduction Technique

To drastically improve computation time and efficacy of pattern recognition, the High Dimensional Data (HDD) needs to be reduced into lower dimensions and the transformation techniques employed to achieve it is commonly known as Dimensionality Reduction (DR) [3]. Dimensionality reduction technique plays a pivotal role in the overall data preprocessing phase through effective feature selection and extraction process (Fig. 1).

Dimensionality Reduction Technique (DRT) is mainly two steps process focusing on Feature Selection and Feature Extraction, though both of these terms may appear very similar, there are quite a lot of differences in the way they work [6]. The existing features in HDD can be combined for the creation of a new reduced feature using



**Fig. 1** Dimensionality reduction flow diagram

DRT popularly known as Feature Extraction (FE) [2]. The newly reduced features will have the maximum amount of information retained in it without much information loss. Feature selection is the process of identifying the most significant feature for the given high dimensional data which can be applied for various applications [2, 5, 7]. Selecting the right features not only guarantees a significant decrease in the effort but also results in quicker convergence of models [5]. While on the other hand extracting a feature vector from an original object is the key objective of Feature Extraction which does not try to reduce already extracted features from a high dimensional feature vector [7].

Analysis or visualization of high dimensional data is very complex and computationally expensive. The real challenge in implementing any feature reduction technique is to establish the right approach that can identify irrelevant and unwanted features called noise [8, 9]. At this stage, noise removal needs to be performed in a complete loss-less manner. The meaningful, relevant and important features/information is called Signal and rest all is noise [10].

$$\mathbf{Xi} = \mathbf{Si} + \varepsilon_i$$

where  $\mathbf{Si}$  denotes the signal and  $\varepsilon_i$  the independent noise.

The low dimensional data which has been reduced from high dimensional data using DRT helps to overcome some of the major issues caused by the curse of dimensionality [11]. Reduced features help in better analysis, visualization, and eventually the optimization of storage cost. Dimensionality reduction techniques help to transform the high dimensional data (mathematically represented below) as

$$\begin{aligned}\mathbf{Y} &= [y_1, y_2, \dots, y_m] \in \mathbb{R}^{k \times l} \\ \mathbf{Z} &= [z_1, z_2, \dots, z_m] \in \mathbb{R}^{k \times n}\end{aligned}$$

where  $\mathbf{Y}$  is HDD transformed to  $\mathbf{Z}$  lower dimension with  $k$  observations. Ideally, the dimensions  $l \ll n$

Visualization is one of the key challenges of the HDD, where the visuals become difficult for more than 3 dimensions, whereas DRT helps to reduce nD dimension into 2D or 3D which helps for better visualization and the storage space is optimized as well [12, 13]. Based on the category of transformation technique applied to achieve dimensionality reduction, Linear Dimensionality Reduction Technique (LDRT) uses a simple linear function, whereas the non-Linear Dimensionality Reduction Technique (nLDRT) is designed to handle the non-linear data [14, 15] (Table 1).

**Table 1** List of Key terminologies in Dimensionality Reduction Technique

Terminology	Brief description
Curse of Dimensionality	Adversely impacts the predictive power of ML models along with extremely increased training time [11]
Linear Techniques	Applies the linear transformation approaches for dimensionality reduction [16, 17]
Non-Linear Techniques	Applies the non-linear transformation approaches for dimensionality reduction [18, 19]
Feature Extraction Techniques	Focus is purely on reducing the dimensional vector space [3, 2]
Feature Selection Techniques	Identification and selection of the relevant, right feature subset [3]

### 3 Linear Dimensionality Reduction Technique

Linear dimensionality reduction technique is the foundation to visualize and analyze high dimensional data, by reducing the features to lower dimension, the computation speed increases phenomenally along with better visualization [16]. The different methods of Linear DR produce a low dimensional linear mapping of the original HDD that preserves some key interesting features in actual data [16, 19, 20].

#### 3.1 Factor Analysis

Factor Analysis (FA) is a technique used to reduce high dimensional data into a reduced number of dimensions called factors, for this reason, it is also one of the “dimension reduction techniques” [21, 22]. It is an interdependent technique in which both the dependent and independent variables cannot be separated or identified. It helps to find a correlation between the observed and unobserved (latent) variable. This technique can also extract all the common variance data and map it to a variable called Factor [23]. Factor Analysis is not only used to reduce the number of variables but also uncovers the clusters of responses. The key notion behind FA is that multiple observed variables have similar patterns of responses due to their associative relationships within an underlying set of latent variables, these factors or unobserved variables cannot be measured directly [24]. As an example, Job satisfaction of a person cannot be measured directly, it might have many features like pay scale, travel time, last promotion, co-worker support, supervisor support, education level, designation etc.

It is very evident in the above diagram that all the variables used are expressed in a linear combination of the factors. One of the key points is how Factor Analysis is different from Principal component analysis, FA considers both the common variance and the unique variance but PCA considers equal variance for all variables [25, 26].

In terms of mathematical steps in Factor Analysis, it starts with setting up the factor score coefficients in which the first set of weights are chosen. This activity ensures that the first factor should be able to explain the largest portion of the total variance [21, 22]. In the next step, the second set of weights are selected so that the second factor explains most of the residual variance, subject to being uncorrelated with the first factor.

One of the most critical assumptions to perform FA, the number of factors used from the input dataset is always equal to the number of variables [27]. So that each identified factor can contribute some amount of variance in the observed variables [28, 29]. Once the variance is calculated for all the factors, they are arranged in descending order of explained variance [21, 23].

#### ***Steps in Factor Analysis:***

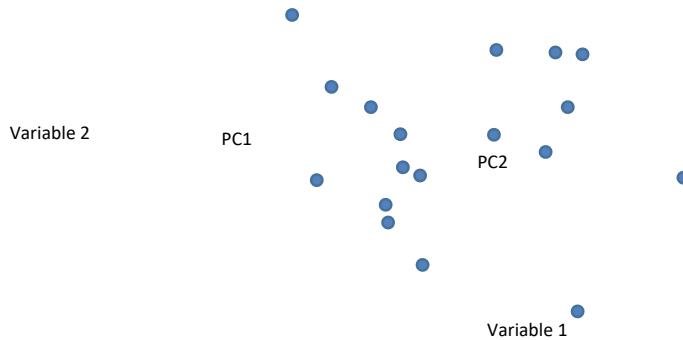
1. Understand the business problem and formulate the problem statement
2. Collect the data and do the analysis, test if the data is a good fit for FA
3. Correlation matrix needs to be constructed
4. Factor Analysis method needs to be determined
5. Appropriate factors need to be identified that explains the variance
6. Identify the rotation technique like Oblique or Orthogonal
7. Label the factor and interpret the data
8. Calculate Factor Scores
9. Determine Model Fitness

### ***3.2 Principal Component Analysis***

The widespread applications such as compressing data for optimal storage, analyzing complex genetic data, and facial recognition are all powered by a very well-known dimensionality reduction technique called Principal Component Analysis (PCA). The DRT literature claims that hoteling [30] immensely contributed to the indigenous development of PCA, whereas the original idea was introduced by Pearson [10].

It involves identifying the set of features or dimensions into a reduced set of uncorrelated dimensions. Though it helps to reduce the number of dimensions, it retains the needed information for the given input data without much information loss during the dimensionality reduction process [31, 32]. PCA is used to reduce the number of features as well as to summarize the variance of the covariance structure through a linear combination using the input variables.

PCA has a long successful history with much wider acceptance due to its simplicity and reliability [33]. The effective linear projection with a simple objective function of mean square error (MSE) helps transform the features from HDD into lower dimension using the efficient eigenvector concepts [31]. Though PCA has many variants, the most popular one is Moniker PCA. It is also widely used in communications, text mining etc. [30].



**Fig. 2** Principal component analysis and its principal components

PCA helps to simplify the given complex data. It expresses all the variables in a linear transformation which identify new coordinates or axes for the given dataset [34–36]. Once the data is collected, data is centered and then the covariance matrix is formed. Eigenvalue and Eigenvectors are derived from the covariance matrix [32]. The combination of eigen vectors with large eigenvalues corresponding to the dimensions needed to be checked which has the strongest correlation (Fig. 2).

The first principal component represented by PC1 is the first axis projecting highest variance of the original data. The next highest variance is just perpendicular to the PC1 is called the second principal component represented by PC2 [31, 32]. Thus, maximum variation can be explained by the first few principal components. Once all the principal components are formed, later using the scree plot we can identify the number of components and the rest of the low variance principal component can be reduced [37, 38].

#### *Steps for PCA:*

1. Understand and formulate the problem statement
2. Collect the input dataset
3. Calculate the mean for the variable
4. Center the data, by subtracting each variable with its respective mean
5. Calculate the covariance matrix
6. Calculate eigenvalues and eigenvectors
7. Identify and select the principal components
8. Create the projected data or the feature vector
9. Forming principal components

Besides the usual benefits of PCA in reducing high dimensional data and visualization, it also helps to find patterns in the high dimensional data [21].

**Table 2** List of key linear dimensionality reduction techniques based on ICA

Techniques	Brief description
Fast ICA	Effectively extracts features from non-Gaussian blended sources [15]
Mixed ICA	Different sources can be ranked effectively using its estimator [42]
RAICAR	Efficiently solves spatial ICA challenges [43]
IICA	Automatically eliminates EEG artifacts to derive the right features [44]
CICA	Widely used in classification use cases [42]
ICAclust	Does not require defining the number of clusters in advance [45]

### 3.3 Independent Component Analysis

An unsupervised, multivariate and linear DRT to extract independent components from a linear mixture of non-Gaussian multi-channel independent sources is called as Independent Component Analysis (ICA) [23, 39, 40]. The resulting uncorrelated and independent components can be achieved effectively by maximizing the likelihood and minimizing mutual between them. ICA has a wide array of applications such as Bayesian Detection, Compression, Data Analysis, Source Localization, and Source Separation and Identification [41].

#### Steps for ICA:

1. Understand and formulate the problem statement
2. Collect the input dataset
3. Assume the data to be linear and a non-Gaussian mixture
4. Perform ICA preprocessing through Centering and Whitening
5. Measure the non-Gaussianity through Kurtosis and Negentropy
6. Minimize the mutual information
7. Maximize the likelihood estimation
8. Perform ICA and Projection Pursuit
9. Forming the independent components

There are many flavors of ICA based on their objective functions to estimate the independent components, e.g. Fast fixed point ICA (FastICA), Mixed ICA, Ranking and Averaging ICA by Reproducibility (RAICAR), Capola ICA (CICA), ICAclust, Probabilistic ICA (PICA), and Sparse Gaussian ICA (SGICA), etc (Table 2).

## 4 Comparative Analysis

The Linear Dimensionality Reduction Techniques (LDRTs) have many variants that offer different core functionalities. Some of the most popular LDRT variants

**Table 3** Comparing popular linear dimensionality reduction techniques

Techniques	Analysis
FA	To uncover the clusters of responses. Gaussian distribution; Parametric; Deterministic
	Popularly used for latent variable identification
	Applied in marketing, microarray analysis, and biological science
PCA	To summarize & preserve variance of the covariance. Gaussian distribution; Parametric; Non-Deterministic
	Popularly used for the reduction of computation time, sparsity & noise
	Applied in Time Series, Image and Face recognition data
ICA	To explore multi-channel independent data sources. Non-Gaussian distribution; Non-Parametric; Deterministic
	Popularly used for Blend Source Separation (BSS)
	Applied in Signal Processing and Gene Expression Data

mentioned under Sect. 2 can also be analyzed based on key comparative properties such as their positive and negative aspects, issues and challenges, and different application areas [46] (Table 3).

Most of the ML algorithms are prone to over-fit the redundant, irrelevant high dimensional data, drastically reducing efficiency and increasing the response time of the ML model [47]. The HDD may have additional issues such as sparsity and noise which dramatically spikes up the difficulty level in identification and management. Thus increases the chances of failure of many DRTs. Performance of the ML model is also affected by other HDD issues such as the curse of dimensionality [48], the small sample size problem, and the absence of class labels. Such challenges can be solved by selecting an appropriate DRT.

## 5 Conclusion

We conclude that there is no doubt that High Dimensional Data (HDD) can heavily undermine the Machine Learning (ML) capabilities but selecting the right Dimensionality Reduction Technique (DRT) focused on reducing computation time and cost can ensure very high predictive power [49, 50]. Principal Component Analysis (PCA) outperforms many other DRTs as their linear transformation requires less computation power. PCA and its variants have been widely implemented across several domains and application areas such as biomedical, video and audio. Some of its variants also work very effectively on non-linear data.

The need to improve and innovate DRTs will keep evolving with the growth of HDD. To achieve effective human–machine collaboration, the most suitable linear DRT should be chosen based on clear business requirements [6, 51, 52]. In future research work, we will perform a detailed review of Non-Linear Dimensionality

Reduction Techniques (NLDRTs). We will also try to solve some of the major pain points by proposing a strong new hybrid DRT.

## References

1. L. Gao, J. Song, X. Liu, J. Shao, J. Liu, J. Shao, Learning in high-dimensional multimedia data: the state of the art. *Multimedia Syst.* **23**(3), 303–313 (2017)
2. L.V.D. Maaten, E. Postma, J.V. Herik, Dimensionality reduction: a comparative. *J. Mach. Learn. Res.* **10**, 66–71 (2009). G. Chandrashekhar, F. Sahin, A survey
3. A.N. Escalante-B, L. Wiskott, How to solve classification and regression problems on high dimensional data with a supervised extension of slow feature analysis. *JMLR* **14**, 3683–3719 (2013)
4. D. Amarasinghe, J. Cabrera, High-dimensional data, *J. Natl. Sci. Found.* **44**
5. S. Ayesha, M.K. Hanif, R. Talib, Overview and comparative study of dimensionality reduction techniques for high dimensional data. *Information Fusion* **59**, 44–58
6. P. Comon, Independent component analysis, a new concept? *Signal Process.* **36**(3), 287–314 (1994)
7. B. Tang, M. Shepherd, E. Milioto, M.I. Heywood, Comparing and combining dimension reduction techniques for efficient text clustering, in *International Workshop on Feature Selection for Data Mining*, vol. 39 (2005), pp. 81–88
8. M. Holmes, A. Gray, C. Isbell, Fast SVD for large-scale matrices. In *Workshop on Efficient Machine Learning at NIPS* **58**, 249–252 (2007)
9. Z. Zhang, F. Yang, K. Xia, R. Yang, A supervised lpp algorithm and its application to face recognition [j]. *J. Electron. Inf. Technol.* **3**, 8 (2008)
10. S. Buchala, N. Davey, T.M. Gale, R.J. Frank, Analysis of linear and nonlinear dimensionality reduction methods for gender classification of face images. *Int. J. Syst. Sci.* **36**(14), 931–942 (2005)
11. X.L. Zhang, Nonlinear dimensionality reduction of data by deep distributed random samplings, in *Asian Conference on Machine Learning*, vol. 2015, pp. 221–233
12. F. Namugera, *Dimensionality Reduction of High-Dimensional Noisy Data*. African Institute of Mathematical Sciences (AIMS), Senegal (2017)
13. T. Lesort, N. Díaz-Rodríguez, J.-F. Goudou, D. Filliat, State representation learning for control: an overview. *Neural Netw.* **108**, 379–392 (2018)
14. E. Shchurenkova, Dimension Reduction Using Independent Component Analysis with an Application in Business Psychology, University of British Columbia, 2017 PhD. thesis
15. J. Rahmanishamsi, A. Donati, M.R. Aghabozorgi, A copula-based ica algorithm and its application to time series clustering. *J. Classif.* **35**(2), 230–249 (2018)
16. Y. Xin, Q. Wu, Q. Zhao, Q. Wu, Semi-supervised regularized discriminant analysis for Eeg-based Bci system, in *International Conference on Intelligent Data Engineering and Automated Learning* (Springer, 2017), pp. 516–523
17. M. Verleysen, D. François, The curse of dimensionality in data mining and time series prediction, in *International Work-Conference on Artificial Neural Networks* (Springer, 2005), pp. 758–770
18. J.P. Cunningham, Z. Ghahramani, Linear dimensionality reduction: survey, in-sights, and generalizations. *J. Mach. Learn. Res.* **16**(1), 2859–2900 (2015)
19. A. Gisbrecht, B. Hammer, Data visualization by nonlinear dimensionality reduction. *Wiley Interdiscip. Rev.: Data Min. Knowl. Discov.* **5**(2), 51–73 (2015)
20. B. Kuster, A.M. Gholami, A.C. Culhane, Dimension reduction techniques for the integrative analysis of multi-omics data. *Brief. Bioinform.* **17**(4), 628–641 (2016)
21. C.K. Chandrasekhar, H. Bagyalakshmi, M.R. Srinivasan, M. Gallo, Partial ridge regression under multicollinearity. *J. Appl. Statistics* **43** (2016)

22. H.H. Haeman, *Modern Factor Analysis*, 3rd Revision Published 1976 by The University of Chicago
23. N.B. Erichson, P. Zheng, K. Manohar, S.L. Brunton, J.N. Kutz, A.Y. Aravkin, *Sparse Principal Component Analysis Via Variable Projection*. arXiv preprint [arXiv:1804.00341](https://arxiv.org/abs/1804.00341)
24. R.P. McDonald, *Factor Analysis and Related Methods*. Lawrence Erlbaum Associates
25. H. Hotelling, Analysis of a complex of statistical variables into principal components. *J. Educ. Psychol.* **24**(6), 417 (1933)
26. S. Deegalla, H. Boström, K. Walgama, Choice of dimensionality reduction methods for feature and classifier fusion with nearest neighbor classifiers, in *15th International Conference on Information Fusion (FUSION)* (IEEE, 2012), pp. 875–881
27. I.T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments. *Philos. Trans. R. Soc. A: Math. Phys. Eng. Sci.* **374**(2065), 20150202 (2016)
28. T. Radüntz, J. Scouten, O. Hochmuth, B. Meffert, Automated EEG artifact elimination by applying machine learning algorithms to ica-based features. *J. Neural Eng.* **14**(4), 46004 (2017)
29. M.F. Glasser, T.S. Coalson, J.D. Bijsterbosch, S.J. Harrison, M.P. Harms, A. Anticevic, D.C.V. Essen, S.M. Smith, Using temporal ica to selectively remove global noise while preserving global signal in functional MRI data. *Neuroimage* **181**, 692–717 (2018)
30. K. Pearson, Liii. on lines and planes of closest fit to systems of points in space, London Edinburgh Dublin Philos. Mag. J. Sci. **2**(11), 559–572 (1901)
31. J.C. Loehlin, *Latent Variable Models: An Introduction to Factor, Path, and Structural Analysis* (American Psychological Association Press, Washington, 1998)
32. C. Ding, H. Xiaofeng, K-means clustering via principal component analysis, in *ICML '04 Proceedings of the Twenty-First International Conference on Machine learning* (2004), p. 29
33. T. Bruce, *Exploratory and Confirmatory Factor Analysis: Understanding Concepts and Applications* (American Psychological Association Press, Washington, 2004)
34. M. Nascimento, F.F.e. Silva, T. Sáfadi, A.C.C. Nascimento, T.E.M. Ferreira, L.M.A. Barroso, C.F. Azevedo, S.E.F. Guimarães, N.V.L. Serão, Independent component analysis (ica) based-clustering of temporal rna-seq data. *PloS one* **12**(7), e0181195 (2017)
35. C.F. Beckmann, S.M. Smith, Probabilistic independent component analysis for functional magnetic resonance imaging. *IEEE Trans. Med. Imaging* **23**(2), 137–152 (2004)
36. P. Ablin, J.-F. Cardoso, A. Gramfort, *Faster ICA Under Orthogonal Constraint*. arXiv, preprint [arXiv:1711.10873](https://arxiv.org/abs/1711.10873)
37. N. Abrahamsen, P. Rigollet, *Sparse Gaussian ICA*, arXiv preprint [arXiv:1804.00408](https://arxiv.org/abs/1804.00408)
38. S.J. Press, S. Wilson, Choosing between logistic regression and discriminant analysis. *J. Am. Stat. Assoc.* **73**(364), 699–705 (1978)
39. C. Ecse, *Dimensionality Reduction*. pca. kernel pca, Lecture slides: COMP-652 and ECSE-608
40. L. Wiskott, *Lecture Notes on Principal Component Analysis* (2013)
41. M. Wan, G. Yang, C. Sun, M. Liu, Sparse two-dimensional discriminant locality-preserving projection (s2ddlpp) for feature extraction. *Soft. Comput.* 1–8 (2018)
42. D.T. Pham, P. Garat, Blind separation of the mixture of independent sources through a quasi-maximum likelihood approach. *IEEE Trans. Signal Process.* **45**(7), 1712–1725 (1997)
43. X.-s. He, F. He, A.I. He, Super-gaussian bss using fast-ica with chebyshev–pade approximant. *Circuits Syst. Signal Process.* **37**(1), 305–341 (2018)
44. Z. Yang, S. La Conte, X. Weng, X. Hu, Ranking and averaging independent component analysis by reproducibility (raicar). *Hum. Brain Mapp.* **29**(6), 711–725 (2008)
45. H. Ince, T.B. Trafalis, A hybrid forecasting model for stock market prediction. *Econ. Comput. Econ. Cybern. Stud. Res.* **51**(3), 263–280 (2017)
46. N. Kambhatla, T.K. Leen, *Dimension Reduction by Local Principal Component*
47. B. Wang, Y. Hu, J. Gao, Y. Sun, H. Chen, B. Yin, *Locality Preserving Projections for Grassmann Manifold*, arXiv preprint [arXiv:1704.08458](https://arxiv.org/abs/1704.08458)
48. S. Ahmadkhani, P. Adibi, Face recognition using supervised probabilistic principal component analysis mixture model in dimensionality reduction without loss framework. *IET Comput. Vision* **10**(3), 193–201 (2016)

49. H. Zhao, S. Sun, Z. Jing, Local-information-based uncorrelated feature extraction. *Opt. Eng.* **45**(2), 20505 (2006)
50. S. Chen, H. Zhao, M. Kong, B. Luo, 2D-lpp: a two-dimensional extension of locality preserving projections. *Neurocomputing* **70**(4–6), 912–921 (2007)
51. M. Balaanand, N. Karthikeyan, S. Karthik, R. Varatharajan, G. Manogaran, C.B. Sivaparthipan, An enhanced graph-based semi-supervised learning algorithm to detect fake users on Twitter. *J. Supercomput.* **75**(9), 6085–6105 (2019). <https://doi.org/10.1007/s11227-019-02948-w>
52. T.N. Nguyen, B. Liu, S. Chu, D. Do, T.D. Nguyen, WRSNs: toward an efficient scheduling for mobile chargers. *IEEE Sensors J.* **20**(12), 6753–6761, 15 June 15, 2020. <https://doi.org/10.1109/JSEN.2020.2974255>

# Healthcare Monitoring System Using Medical Smart Card



G. Sujatha, D. Hemavathi, K. Sornalakshmi, and S. Sindhu

**Abstract** Advancement in the field of information technology and communication has given huge growth to IOT in all the fields specifically in medical science and it is considered among the fastest to accept IOT, many of the medical successes are achieved due to IOT devices. It has brought huge convenience to both doctors and patients in saving their precious time and money. Keeping the convenience of doctors and patients in mind, we are creating a system for patients through which they can check all their basic healthcare conditions and store it in smart card as this card. This system has features like BP monitoring, Heart Beat monitoring, Temperature measurement and EEG monitoring in one system. It has a feature of real time monitoring whose data will be stored in the cloud and doctors can easily access it without wasting their time on these tests and can save some lives.

## 1 Introduction

Technology has become one of the integral parts of our lives. It is evolving day by day, which is hugely responsible for our change in lifestyle. Technological solutions have simplified the work of every individual and brought out efficiency. IOT is presumed to be a game changer and creating quite a buzz in almost every industry and healthcare [1, 2] is among the fastest in adopting IOT in their field and to embrace the opportunity of improving the quality and effectiveness in medical services [2–5]. IOT has grown

---

G. Sujatha (✉) · D. Hemavathi · K. Sornalakshmi · S. Sindhu  
SRM Institute of Science and Technology, Chennai, India  
e-mail: [sujathag@srmist.edu.in](mailto:sujathag@srmist.edu.in)

D. Hemavathi  
e-mail: [hemavatd@srmist.edu.in](mailto:hemavatd@srmist.edu.in)

K. Sornalakshmi  
e-mail: [sornalak@srmist.edu.in](mailto:sornalak@srmist.edu.in)

S. Sindhu  
e-mail: [sindhus2@srmist.edu.in](mailto:sindhus2@srmist.edu.in)

immensely in the field of medical science, many of the medical successes is achieved due to the research in Internet of Things (IOT) devices.

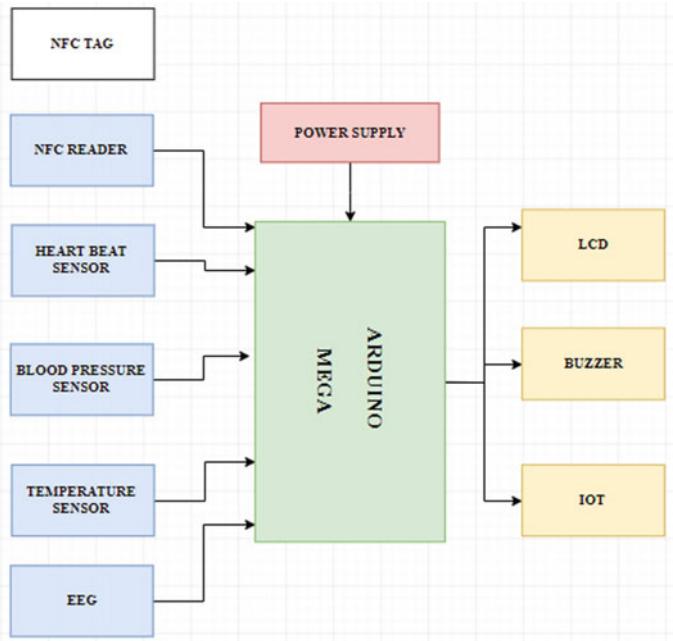
With so much vast opportunity in modernization of medical science with the help of IOT, we decided to develop a system or platform that can be considered useful in healthcare industry. Then, we came across the issues with basic medical healthcare check-ups. The existing methodology requires continuous visits to the hospital for regular check-ups, which is very time consuming and costly. There is no real time monitoring of patients unless he/she gets admitted in the hospital, which is again time consuming, and can cost a lot, it is not suitable for patients proper and continuous monitoring. It also requires a lot of manual work, which again increases work force. Their database systems for storing record about patient detail is not safe and can easily be misused [6]. That's how we got motivated by the idea of solving an issue, which can further contribute to the healthcare industry for the good of humankind.

After understanding the importance of Internet of Things and its role in the field of medical science, we decided to create a platform that will monitor the health of the patients "Smart Card Healthcare Monitoring System" using sensors [7]. This system consists of four different medical sensors, which will monitor patient's basic health condition at real time. It consists of Temperature sensor, Heart rate sensor, Blood Pressure sensor, EEG (Electroencephalogram) sensor all in one system; all these sensors will monitor patient's continuous health behaviour and will store all the readings in the cloud for easy access of records from anywhere. In case of any alarming change in patient's behaviour, it will immediately intimate the concerned person by buzzing. This system also has a smart NFC card, which can be used as a digital record of patient for doctors to assess. This system is developed to be time saving, cost effective and very efficient. It will be extremely helpful to all the people especially elderly people, as they do not have to go to hospital or clinics for these basic check-ups. It will also save doctors time [8] while assessing the patient, as the tests have already been conducted at home.

## 2 Proposed System

### 2.1 Working Principle

The proposed system here has a patient module, a security module and a monitoring section where a patient's dynamically changing conditions in the body can be monitored throughout. The patient health status can be observed from anywhere. The data gathered in the process is stored in NFC card which is very efficient and assures that the data is kept private. The monitoring of each health parameter is done by a single device. The device is easily portable and saves time as individual monitoring of each parameter is not required. As there is a continuous monitoring of the patient immediate action can be taken in case of an emergency.



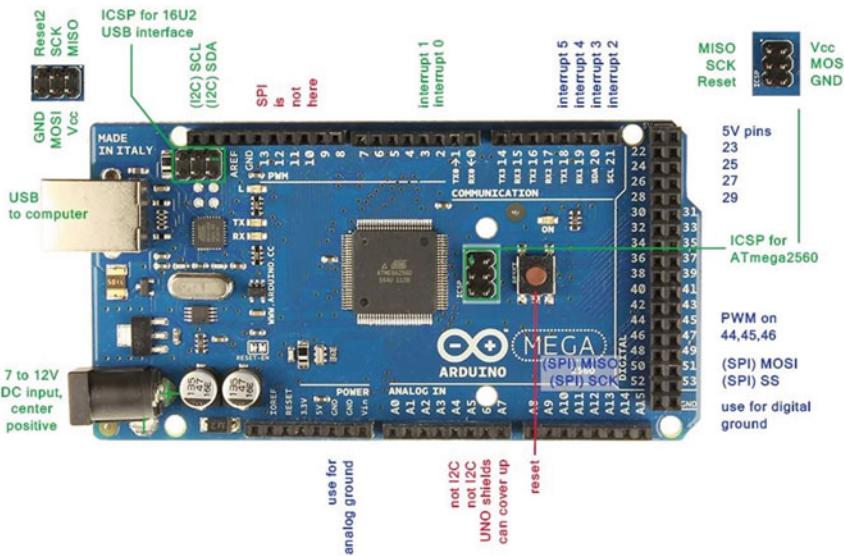
**Fig. 1** Patient module

Figure 1 represents the working principle of patient module. This NFC card gathers the information collected through sensors and uploads it to the website where the doctor and the patient's relatives can know how the health condition of the patient is. In actual application, the sensors will be present in the patient's bed and the NFC card will help to control the bed system, which has a moving mechanism, which can be used to adjust the bed by monitoring patient's condition. The buzzer will be used if there is assistance required by the patient or in a case of emergency if the patient is under any critical situation due to sudden change in any of the health parameters (Fig. 2).

## 2.2 Components

### 2.2.1 NFC (Near Field Communication)

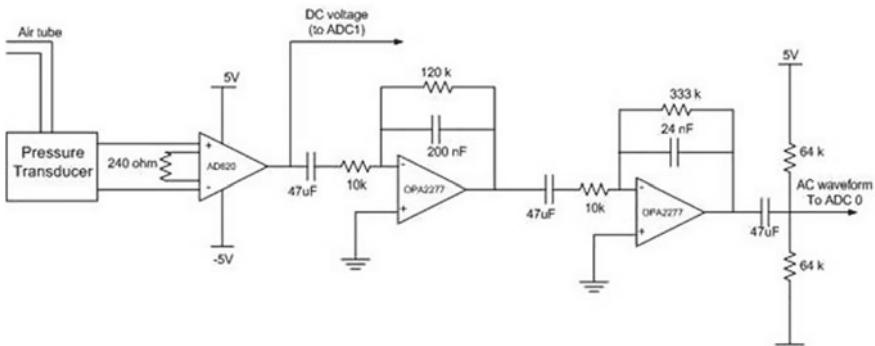
It is a wireless short-ranged communication between compatible devices using magnetic field induction. There is only a need of transmitter and a receiver. It is based on radio-frequency identification (RFID), i.e. it communicates using radio waves.



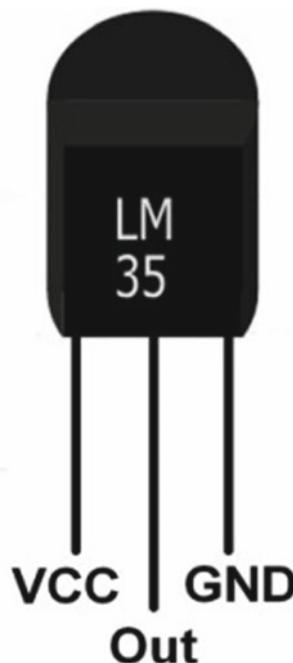
**Fig. 2** Arduino mega architecture diagram

### 2.2.2 Blood Pressure Sensor

Figure 3 shows the architecture diagram of Blood Pressure Sensor. The Blood Pressure Sensor is used for measuring human's blood pressure by using the oscillometric technique to measure the Systolic, Diastolic and Mean Arterial pressure. The measurement is mostly done by Sphygmomanometer, which uses a mercury column in which the height of mercury and fall in the column show the pressure difference, but the electronic devices do not use mercury column.



**Fig. 3** BP architecture diagram

**Fig. 4** Temperature sensor

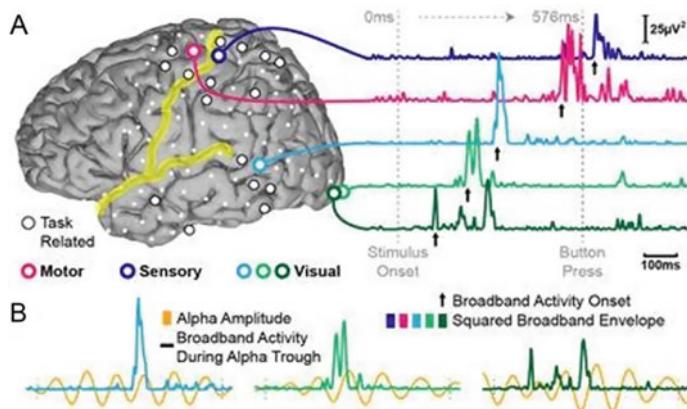
### 2.2.3 Temperature Sensor

LM35 is a sensor shown in Fig. 4, is used to measure temperature and display the value in °C. The sensor's circuit needs to be contained to prevent oxidation. The sensor is more accurate as compared to the thermistor. It has less self-heating (less than 0.1 °C) and the optimum temperature range is from –55 to 150 °C. The scale factor is 0.01 V/°C.

It does not need to be calibrated and the accuracy is maintained at  $\pm 0.4$  °C (room temperature) and  $\pm 0.8$  °C (b/w 0 °C to 100 °C).

### 2.2.4 ElectroEncephaloGram (EEG)

Electroencephalogram (EEG) is used in the testing of electrical activities of the waves of the brain. It is shown in Fig. 5. Brain cells, which are used in communicating with each other through the use of impulses of an electrical nature. The EEG can be used for detecting problems potentially associated with any activities in the brain. The test is allowing the brain wave patterns to be calculated. Small, aligned sticky discs and electrodes are pasted to the scalp and the hand. These analyses the impulses in the brain and send signals to a system, where the results are calculated and stored. It is painless and safe. This setup will record our brain's electrical activity just like the variations called traces. Each trace result to a different region of the brain. EEGs



**Fig. 5** EEG function measurement

were recorded every time on paper as they provided a more accurate value, but these days the closer value is computerized, paperless

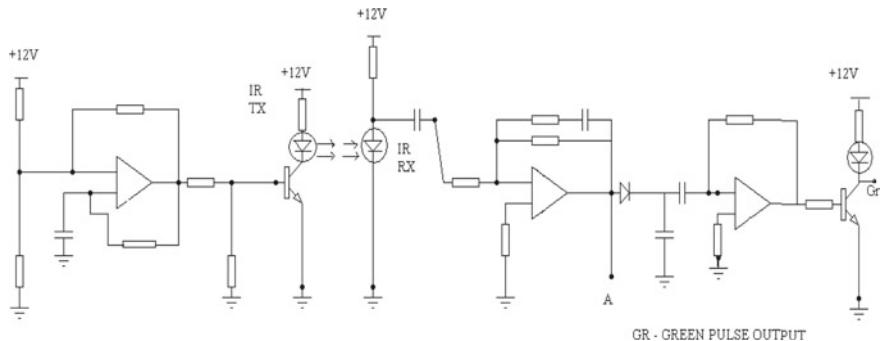
EGs are used more or less often. Electrodes are required to be put on both sides of the head and one side of the hand which are then connected by wires to the electrical box. The wires can only suggest calculated values of the brain activity, they do not release any electrical current to your head. Then it is connected to the machine.

## 2.2.5 Heart Rate Sensor

This heart beat checker allows us to provide digital output the placement of the finger is between the LED and the receptor. When the sensor is detecting the changes in the blood flow, then the LED flashes in a rhythmic synchronized manner with each heartbeat. The digital output is connected to microcontroller to measure the beat per minutes (BPM) rate. It works by the principle that light is modulated when the blood flow through the finger changes after each pulse. Figure 6 represents the architecture diagram of heart rate sensor.

## 2.2.6 Liquid Crystal Display

The liquid crystal display is device that allows one to electronically display the various data that is passed it in a digital format. A display such as these have its use in a wide array of devices on a day-to-day basis. A  $16 \times 2$  is the most basic module and is very much helpful in showing information in an easy to understand format. These modules are made over seven segments and other several multi segment LEDs.



**Fig. 6** Architecture diagram of heart rate sensor

### 2.2.7 BF—Busy Flag

Busy flag is used to show the status of the LCD, which allows the information to be passed to it in a continuous manner whenever the status is set to ready. At the time when the command is sent or data is sent to the LCD for processing instructions, at that time this flag is set and when the instruction is executed this flag is cleared. This is helpful and exact amount of delay for the processing of the LCD. To check the busy flag, the condition Read S = 0 and Read/Write = 1 must be met and The MSB of the LCD data bus (D7) act as busy flag. While BF = 1 represents that the LCD is busy and cannot accept any further commands or data and BF = 0 signals it is ready.

### 2.2.8 Buzzer

Buzzer is an audio signal device, which can be mechanical, piezoelectric or electromechanical types. The uses of buzzer as alarm devices, timer. In this work, it is used as an alarm if a patient is in critical situation or in emergency.

### 2.2.9 Internet of Things (IOT)

IOT is way of extending the connection between different devices through the Internet and to control each device separately. It involves machine learning, embedded systems, analytics and sensors. It is mostly used in home automation where all the home devices and appliances are connected through Wi-Fi and all of these devices are controlled wirelessly through voice or gestures. All can be controlled through android and IOS devices. These may include thermostat altering, altering light brightness or switching the fans, lights and air conditioners on and off.

The sensors connected to the Arduino mega are monitored via Wi-Fi where the collected data of the patients changing conditions in the body are noted and uploaded to the Internet through Wi-Fi using NFC card and NFC reader. The buzzer will

be engaged manually through Wi-Fi if the doctor or the patient's family members monitoring the patient's data find any irregularity in the patient's body condition to prevent any harm that could be caused to the patient.

### 3 Results and Discussion

Smart Card Health Care Monitoring System uses various sensors that help in calculating the values of various vital signs in the body of a person and thereby, help in getting most accurate and appropriate readings based on various parameters of health care check-up.

The following sessions were performed to check the working of the device upon different patients, where values were recorded for Heartbeat, Body Temperature, Blood Pressure and EEG. If the values exceed the normal range the alarm is turned on to indicate the critical condition. In the table, 0 represents the reading in which values went beyond the normal range causing ringing of alarm and 1 denotes the values within the normal range.

Operational constituents of the device are—Arduino Uno, heartbeat sensor, Phototransistor and IR LED. Total three sessions were performed on three separate individuals with 10 trials, to determine the accuracy of the device and proper functioning of all the sensors. With every abnormal value, there is sounding of the alarm denoting that the values are beyond the range as per normal human standards of health. Below are the normal range of various health parameters:

#### Normal Range:

Heart Beat: 60 to 100 beats per min

Body Temperature: 97 F–99 F/36.1 °C–37.2°C

Blood Pressure: 120/80–140/90

EEG: 7–12 Hz

#### 3.1 Session 1

In the first session, the patient was requested to insert the finger in heartbeat sensor for which the readings were recorded, and then upon holding the temperature sensor with fingers the readings of body temperature were noted. In the same way, readings for blood pressure and EEG were also recorded using the respective sensors. Upon performing the test, the following readings were observed as mentioned in the Table 1.

**Table 1** Session 1 trial

Session 1	Heart beat	Body temperature (F)	Blood pressure	EEG
Trial 1	40	96	100/70	6
Trial 2	72	97.1	122/82	7.2
Trial 3	68	97.3	130/82	7.8
Trial 4	58	95	110/78	8.2
Trial 5	74	98	128/85	8.4
Trial 6	73	98.4	130/85	8
Trial 7	76	99.1	126/84	9
Trial 8	59	95	115/75	6.5
Trial 9	72	97.2	128/84	7.8
Trial 10	73	97.5	130/84	8

### 3.1.1 Result

Based on the above readings the following Table 2 was plotted as per the range of the above parameters, and the result calculated as per average of all the readings comes out to be 7, which marks the system to be slightly effective.

$$\text{Session 1 : } \sum \text{Trial 1} + \text{Trial 2} \dots + \text{Trial 10} = 7 \text{ (Slightly Effective)}$$

## 3.2 Session 2

Like the first session, the second session was performed on a different individual, and the readings for Heart Beat, Blood Pressure, Body Temperature and EEG were noted down.

Table 2 mentions the readings taken during the second session of health care check-up for the second patient (Table 3).

### 3.2.1 Result

Table 4 is plotted on the basis of above-mentioned readings, and the average for which comes out to be 9, which denotes that the device is very effective in its functioning.

$$\text{Session2: } \sum \text{Trial 1} + \text{Trial 2} \dots + \text{Trial 10} = 9 \text{ (Very Effective)}$$

**Table 2** Average of all trial results

Session	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6	Trial 7	Trial 8	Trial 9	Trial 10	Results	Remarks
Session 1	0	1	1	0	1	1	1	0	1	1	7	Slightly effective

**Table 3** Session 2 trial

Session 1	Heart beat	Body Temperature (F)	Blood pressure	EEG
Trial 1	73	97.3	128/84	7.2
Trial 2	78	97.1	130/88	9
Trial 3	74	98	128/86	7.8
Trial 4	76	97.8	124/82	8
Trial 5	74	98	128/85	8.4
Trial 6	58	98.4	106/70	6.6
Trial 7	76	99.1	126/84	9
Trial 8	88	98	115/75	7
Trial 9	72	99	128/86	7.5
Trial 10	75	97.5	128/85	8.2

### 3.3 Session 3

In the third session, again the readings were observed upon a third individual with 10 trials and below are the details of the health parameters of the patient. Table 5 represents those details.

#### 3.3.1 Result

From the above readings the following result was plotted in Table 6, from which result calculated as average of all the readings comes out to be 8, which again proves the system to be very effective.

$$\text{Session 3 : } \sum \text{ Trial 1 + Trial 2 . . . + Trial 10 = 8 (Very Effective)}$$

### 3.4 Final Results

Range of result above 7 makes it very effective while between 6 and 7 makes it slightly effective and range below 6 makes it not effective. Therefore, Table 7 shows the readings which were observed and shows the average of the trail results.

**Table 4** Average of all trial results

Session	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6	Trial 7	Trial 8	Trial 9	Trial 10	Results	Remarks
Session 1	1	1	1	1	1	0	1	1	1	1	9	Very effective
Session 2	1	1	1	1	1	0	1	1	1	1	9	

**Table 5** Session 3 trial

Session 3	Heart beat	Body temperature (F)	Blood pressure	EEG
Trial 1	74	98.2	130/88	7.5
Trial 2	80	97.3	122/84	8
Trial 3	52	94	135/89	8.4
Trial 4	78	98.6	126/88	8.6
Trial 5	79	98	130/85	9.2
Trial 6	82	98	126/84	7
Trial 7	84	97.8	130/88	8.8
Trial 8	56	93	102/74	6.8
Trial 9	86	96.2	126/88	8.8
Trial 10	78	97.8	126/84	8.4

## 4 Conclusion

Advancement in the field of information technology and communication has given huge growth to the Internet of Things (IOT), to emerge up quite fast especially in case of healthcare environment, it has been a great source of help because it bringing convenience to patients and physicians as it has to be applied in fields like patient information management, constant real-time monitoring, blood test information management, medical emergency management etc. Internet connectivity into everyday objects and physical devices is the extension of the IOT (Internet of Things). In the present moment, there is lot of noise about IOT and its unsurmountable impact on everything, beginning from our daily travels, to shopping, to keeping the track of inventory and so on. So, how could the field of medical science be deprived of this expanding technology? Therefore, this work is a kind of link between the services provided in medical field and modern technology, to enhance the healthcare facilities by the means of IOT. Costly healthcare services, with the ageing global population, and rising chronic diseases have resulted into daily medical check-ups of Blood Pressure, heartbeat, etc. have become a common scene in hospitals and clinics. This work brings out a solution to this uprising problem of modern days, by helping to provide all the facilities of medical healthcare check-ups at your very doorstep and helping the community by re-emphasising the principle of: Health is Wealth!

This Smart Card Healthcare System could prove to be a boon in the field of medical science, as in near future, many more features could be incorporated into it such as Continuous Glucose Monitoring System which could help in monitoring the glucose level of the patient. Real time monitoring through connected device is a medium for life saver in the situation of emergencies related to health. In the modern fast-paced age with the rapid advancement in the field of IOT, there is a dire need of such kind of system, which can be cost effective and less time consuming for the benefit of the patient. Therefore, this work furthers automation for faster medical personnel responses. The work has enabled us to understand the quick and accurate

**Table 6** Average of all trial results

Session	Trial 1	Trial 2	Trial 3	Trial 4	Trial 5	Trial 6	Trial 7	Trial 8	Trial 9	Trial 10	Results	Remarks
Session 3	1	1	0	1	1	1	1	0	1	1	8	Very effective

**Table 7** Final outcome

Session	Trail 1	Trail 2	Trail 3	Trail 4	Trail 5	Trail 6	Trail 7	Trail 8	Trail 9	Trail 10	Results	Remarks
Session 1	0	1	1	0	1	1	1	0	1	1	7	Slightly effective
Session 2	1	1	1	1	1	0	1	1	1	1	9	Very effective
Session 3	1	1	0	1	1	1	1	0	1	1	8	Very effective

action in the remote areas, where approaching for help is a major concern, and this work is a great benefactor for the same. Therefore, this work can make the patient caring workflow automated using IOT and health care mobility solution.

## References

1. S.M. Riazul Islam, D. Kwak, H. Kabir, M. Hossain, Kyung-Sup Kwak, *The Internet of Things for Health Care: A Comprehensive Survey*, IEEE (2015)
2. S. Sharma, K. Chen, A. Sheth, *Towards Practical Privacy-Preserving Analytics for IoT and Cloud Based Healthcare Systems*, IEEE (2018)
3. Z. Zhang, B. Muthu, C.B. Sivaparthipan, The necessary of constructing preventive health intervention policy under the trend of deep aging in China. *J. Ambient. Intell. Humaniz. Comput.* (2020). <https://doi.org/10.1007/s12652-020-02594-8>
4. N.T. Le, J.-W. Wang, C.-C. Wang, T.N. Nguyen, Automatic defect inspection for coated eyeglass based on symmetrized energy analysis of color channels. *Symmetry* **11**, 1518 (2019). <https://doi.org/10.3390/sym11121518>
5. K.-H. Yeh, A secure IOT-based healthcare system with body sensor networks. *IEEE Access* (2016)
6. J.J.P.C. Rodrigues Dante Borges De Rezende Segundo, H.A. Junqueira, M.H. Sabino, R.M. Prince, J. Al-Muhtadi, V.H.C. De Albuquerque, *Enabling Technologies for the Internet of Health Things*, IEEE Access, March 2018
7. H. Al-Hamadi, I.-R. Chen, Trust-based decision making for health IOT systems. *IEEE Internet Things J.* (2017)
8. P. Gope, T. Hwang, BSN-care: a secure IoT-based modern healthcare system using body sensor network. *IEEE Sensors J.* **16**(5), 1368–1376 (2016)

# Enhanced Energy Distributed Unequal Clustering Protocol for Wireless Ad Hoc Sensor Networks



G. Parimala, A. Razia Sulthana, and S. Nithiya

**Abstract** Wireless Sensor Network is a rising innovation that comprises of a few quantities of sensor hubs to detect different parameters for various applications. Non-replaceable batteries are the greater part in this case which are enabled by using those batteries. Henceforth, amid the structure of such systems it is very essential that the sensor hubs with very less energy consumed is expected under many circumstances. We presented here a novel model namely enhanced energy distributed unequal clustering which is mainly utilized for tackling energy consumption issue in multi-hop remote sensor systems. In the proposed method with area of BS and energy gives significance as cluster metric. In view of the metrics, diverse nodes are allocated. In this, another method has been proposed to enhance EDUC, by choosing cluster head (CH) thinking about number of hubs. CHs competition radius calculations depend on the location of nodes from BS and it causes the growth of clusters that are smaller in size, if the nodes are near to BS. Residual energy is considered for selection of CH and the single-metric-based CH selection leads to minimize of network lifespan. The outcomes demonstrate that the proposed plan beats the current conventions regarding system lifetime and performances in all the scenario.

## 1 Introduction

Remote Sensor Networks will be systems that comprises of minor detecting gadgets (down to the span of a grain), called as sensor hubs. These sensor hubs are utilized to screen parameters like the temperature of a specific district and afterward transmits

---

G. Parimala (✉) · S. Nithiya  
SRM Institute of Science and Technology, Kattankulathur, Chennai, India  
e-mail: [parimalg@srmist.edu.in](mailto:parimalg@srmist.edu.in)

S. Nithiya  
e-mail: [nithiyas@srmist.edu.in](mailto:nithiyas@srmist.edu.in)

A. Razia Sulthana  
Birla Institute of Science and Technology-Pilani, Dubai Campus, United Arab Emirates  
e-mail: [razia@dubai.bits-pilani.ac.in](mailto:razia@dubai.bits-pilani.ac.in)

these detected data to another sensor hub or some other gadget in the system [1]. Subsequently, so as to transmit these detected data, the hub requires some measure of vitality [3–5]. This vitality is provided to the hubs by methods for batteries which perhaps replaceable or non-replaceable or now and again these batteries may likewise be battery-powered, contingent upon sun-oriented power for reviving [2]. Thusly, the lifetime of the hubs will be expanded which thus will likewise build the lifetime of the system as the lifetime of a system exclusively relies upon the lifetime of the hubs present in the system.

For decrease of vitality utilization in sensor hubs, additionally grouping encourages information conglomeration at bunch head by diminishing the quantity of transmitted information bundles. Depending on the received information, cluster heads will be elected by base station and will arrange the remaining nodes into clusters [6]. Single-hop transmission mode is applied to accomplish the communication within and outside of the cluster. Energy utilization for data transmission is reduced in the network because the base station has knowledge about the positions and energy of sensor nodes [7]. In such cases, multi-bounce correspondence is effective in conquering signal engendering troubles. In any case, in light of the fact that the radio disperses vitality in transmission as well as in gathering, coordinate transmission is likewise valuable [8]. Be that as it may, there is an impediment if there should be an occurrence of direct transmission moreover. If there should be an occurrence of correspondence from group make a beeline for the BS. The grouping conventions built up that utilization multi-jump correspondence for accomplishing more vitality productive between bunch correspondences. Multi-bounce LEACH, EADC, EDUC, and so on are some such conventions [9].

In this paper, an endeavor has been made to enhance arrange life expectancy of an EDUC convention utilized in ceaseless observing applications. The EDUC utilizes non-uniform bunching calculation to alleviate the vitality gap issue. Another key thought utilized in our enhanced EDUC convention is amid determination strategy of activity transferring. The cost engaged with handing-off, regarding vitality, is joined as the measurements for choosing one of the doable hubs as a transfer hub rather than just the separation data utilized in EDUC. Consequently, the principle point of this undertaking is to lessen the vitality devoured by the system amid the transmission of information and in the meantime give a vitality productive and dependable information transmission.

## 2 Literature Review

The main convention accessible in this class is low-vitality versatile grouping progressive system (LEACH) convention [10–12]. It chips away at the guideline of single-bounce correspondence between base station and the hubs. This makes it temperamental for extensive scale systems. There are different other changed LEACH conventions which are enhancements over the LEACH convention. Cluster based design is the widely accepted method in WSNs to conserve energy. Most of

the clustering algorithms do not consider the nodes distances to BS for forming the clusters in the network, and it leads to faster energy drainage for the nodes that are located near to BS. To resolve these issues and to improve network lifetime, many unequal clustering algorithms have been introduced in literature [13, 14]. Clusters that are generated closer to BS have smaller in size and larger sized clusters will be generated if the nodes are located farther from BS. In unequal clustering, cluster size is directly proportional to the distance of nodes from BS. Smaller sized clusters highlight that they have smaller number of cluster members and very less traffic within the cluster.

In this strategy, the CH determination is probabilistic and accordingly singular hubs can be delivered hybrid vitality proficient dispersed (HEED) bunching calculation, where the CHs are chosen relying on the lingering energies of the hubs and the expense of correspondence for intra-groups. For correspondence in between bunch, utilization of multi-bounce correspondence comes viable. It is viable in drawing out the lifetime of the hub. It needs in adjusting the heap as hubs near BS bite the dust rapidly. Grouping is one of the answers for better vitality preservation during the time spent correspondence in WSN [15].

The system has implemented a distributed algorithm namely Scalable Energy Efficient Clustering Hierarchy. In SEECH, CH and relay nodes are selected based upon the eligibility criteria of nodes. High degree nodes are assigned as CHs and relay nodes, respectively. Choice of relay node is from the node that is lying closer to source node and by considering maximum residual energy. CH can collect and forward the data to BS. Only relay nodes are permitted to communicate to base station either by direct mode or by multi-hop communication. In any case, it has the restriction that it is helpful for single-bounce arranges as it were [16].

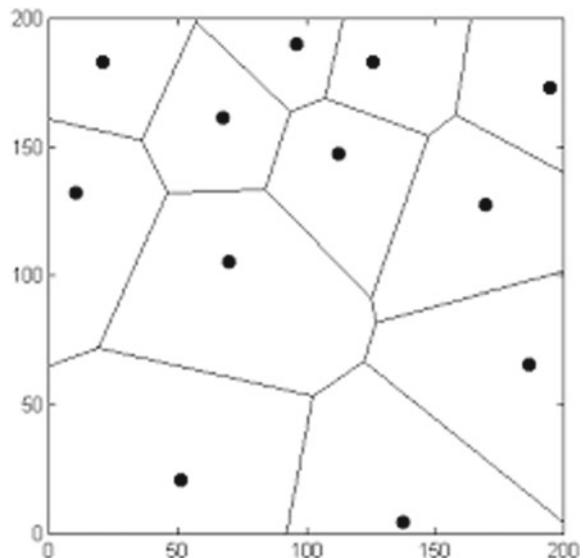
The system has proposed an Energy Driven Unequal Clustering protocol (EDUC) for generating unequal sized clusters in wireless sensor network. CH competition radius calculations depend on the area of nodes from BS, and it causes the generation of clusters that are smaller in size, if the nodes are lying closer to BS. The nodes that are residing near to BS utilize large amount of energy compared to the nodes that are located away from BS. The sensor nodes are randomly dispersed in a sensing field. Sensor nodes energy consumption and energy distribution of nodes within cluster are calculated [17].

### 3 Proposed Method

#### Improved Unequal clustering algorithm for WASN

In this system, the base station communicates a “welcome” message to all hubs at a specific power level. By thusly, every hub can figure the surmised separation to the base station dependent on the got flag quality. It not just encourages hubs to choose the correct power level to speak with the base station, yet additionally causes us to deliver bunches of unequal size. Point-by-point depictions of the unequal grouping

**Fig. 1** Clusters formed as a cell around heads



calculation is introduced in the accompanying segment. Bunching a remote sensor arrange implies apportioning its hubs into groups, each with a CH and the conventional hubs as its individuals. The undertaking of being a CH is pivoted among sensors in every datum assembling round to disperse the vitality utilization over the system. EEUC is an appropriated CH focused calculation, where CH choice is principally founded lingering vitality of every hub. The pseudo code for a discretionary hub is given underneath (Fig. 1).

### 3.1 Inter-Cluster Multi-Hop Routing

At the point when group heads convey their information to the base station, each bunch head first totals the information from its group individuals, and afterward sends the parcel to the base station by means of multi-hop correspondence. In some proposed calculations like PEGASIS, transfer hubs can total the approaching bundles from different groups together with its very own parcels. This supposition is eccentric in light of the fact that the level of detected information connection between various bunches is similarly low. The steering issue here varies significantly from that of conventional specially appointed remote systems due to the many-to-one movement design. Then again, neither inquiry driven nor occasion driven directing conventions for remote sensor systems can be connected to the group heads overlay. Hence, we plan an upgraded multi-bounce steering convention for the between bunch correspondence.

### 3.2 Improved Protocol Mechanism

The bunching technique utilized is comparable in activity to typical EDUC convention. The convention works in rounds. After nodes are sent, every center regardless figures its partition from base station. For this, BS imparts a banner, which is heard by utilizing all centers. In view of the got banner first-rate, every center approximates its partition to BS. Each circular contains pack set-up degree and determined country degree in which data transmission happens. The essential sub-degree is the neighbor center realities hoarding toward the start of information amassing sub-stage, every center point imparts a Node\_Msg, which joins its waiting force close by its recognizable proof. Every last one of the center points, with the range assortment, get the message absolutely one of its companions. The average residual energy of the cluster for each according to the below equation,

$$e_{\text{average value}} = \frac{\sum_{i=1}^n R_i, e_s}{m_c} \quad (1)$$

The upgraded EAUC conspire depends on the EADUC convention; in any case, as opposed to the EAUC, it utilizes an alternate rivalry range rule for delivering unequal groups. In the first EAUC convention, in the articulation for rivalry span, just the separation between the hubs and the BS, and the remaining vitality of the hubs is considered. So as to represent the cost engaged with accumulation, the proposed plan additionally thinks about the quantity of neighbors, notwithstanding the over two elements, while choosing the opposition radii. The gathering heads send the data package to the BS either explicitly, or through moving. In case the partition from specific pack scramble toward BS is more noticeable than limit eliminate (dist\_th), between bunch correspondences is finished; by and large, arrange data transit. For between bundle correspondence, the assurance of gathering head as next bounce center point (move center) is then drawn. In the principal EAUC show, hand-off center assurance is done as one of the neighbor nodes from the contender sending set by a boundary Erelay.

This *Erelay* parameter is computed as given below,

$$e_{\text{realy}} = D^2(R_j, R_i) + D^2(R_j, BS) \quad (2)$$

In the proposed plan, for between bunch correspondence measures, each bundle head initially conveys a message including mote id, extra essentialness, and the amount of gathering. To furthermore upgrade the execution, the gathering head turn isn't done in each round or maybe once the pack set-up is gained, and it is held two or three routine. In first routine of show task, the message transit stage and the relentless state stage incorporates different genuine spaces and for each huge opening further contains number of more modest than regular openings. In each less space, the whole method of data transmission dispose of is passed on. In the midst of the last more modest than ordinary space, the part centers send their excess essentialness close by

the data. After one imperative opening is done, the CH turn inside as far as possible is finished. The old CH gets displaced by another bundle head in a comparable cluster dependent upon the extra imperativeness of the centers and the partition from the current CH.

## 4 Result and Discussion

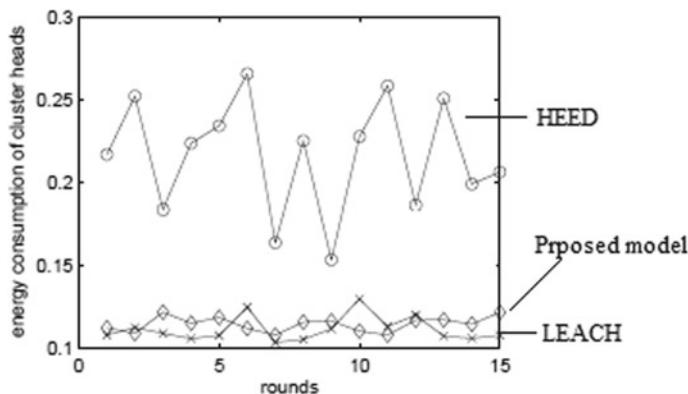
In this segment, we assess the execution of the enhanced EEUUC system by means of simulation. We contrast improved EEUUC. In multi-hop steering is utilized amid bunch heads conveying the information to the BS as indicated by a few references. We likewise run broad tests to decide the ideal number of bunches to use in LEACH, and the ideal group span to use in HEED (Fig. 2).

The fundamental point of this work is to decrease the vitality expended amid the transmission of information from source hub to goal hub. Our work is to break down the execution of the system and the vitality utilization, by fluctuating at least one element. Contingent upon how these elements influence the transmission vitality we discover the ideal incentive for these elements with the end goal that the transmission of information is done in a vitality proficient and in a solid way.

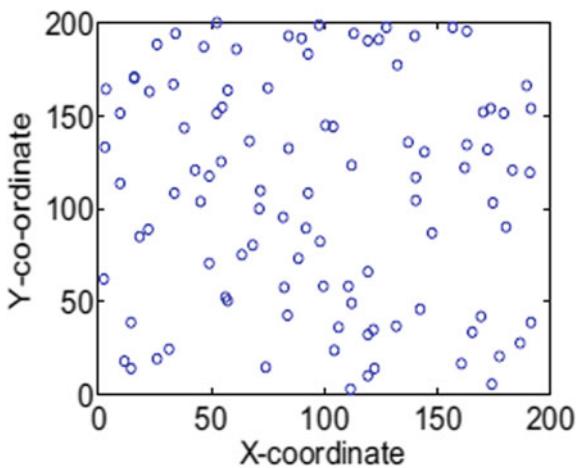
### *Simulation environment*

For over an area of  $200 \text{ m}^2 \times 200 \text{ m}^2$  120 nodes are uniformly deployed in Fig. 3.

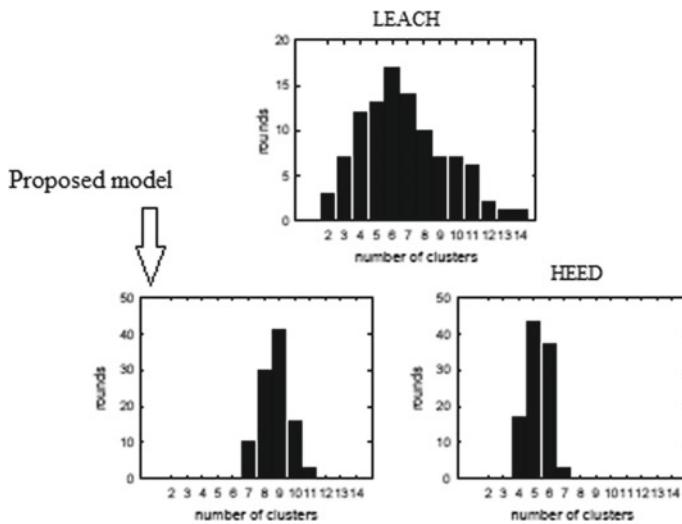
So as to watch the impact of the proposed strategy for registering rivalry span and transferring and the method of division of information transmission stage into major and smaller than expected spaces independently, the consequences of the proposed convention, the improved EDUC, are appeared in two stages in the consequent segments. In the primary execution, to be specific enhanced EDUC the proposed convention utilizes the technique for grouping and transferring just without fusing



**Fig. 2** Amount of energy spent by clusters

**Fig. 3** Network topology

the division of information transmission stage. In the second execution, to be specific enhanced EDUC, the strategy for grouping and handing-off alongside the procedure of division of information transmission stage is joined. In any case, the aggregate outstanding vitality of the system is less if there should be an occurrence of situation 3 when contrasted with situations 1 and 2 and also the stability of our clustering algorithm shows the distribution of the number of clusters Fig. 4.

**Fig. 4** Number of clusters in each round

## 5 Conclusion

In this paper, we have presented an enhanced energy distributed unequal clustering protocol (Enhanced EDUC) to improve the lifespan of WSN. The problem areas issue shows up, while utilizing the multi-hop steering in a bunching approach. We contend that both the revolution of group heads and the measurement of remaining vitality are not adequate to adjust the vitality utilization over the system. To address the issue, we initially acquaint an unequal grouping component with equalization the vitality utilization among bunch heads. Subsequently, the vitality utilization among the group head hubs is all the more adequately adjusted. The result of this investigation will be valuable for taking care of the vitality gap issue in information gathering systems.

## References

1. F. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci, A survey on sensor networks. *IEEE Commun. Mag.* **40**(8), 102–114 (2002)
2. B. Krishnamachari, D. Estrin, S. Wicker, The impact of data aggregation in wireless sensor networks, in *Proceedings of IEEE International Conference on Distributed Computing Systems Workshops (ICDCSW)* (2002), pp. 575–578
3. V. Mhatre, C. Rosenberg, Design guidelines for wireless sensor networks: communication, clustering and aggregation. *Ad Hoc Netw.* **2**(1), 45–63 (2004)
4. W. Heinzelman, A. Chandrakasan, H. Balakrishnan, An application-specific protocol architecture for wireless microsensor networks. *IEEE Trans. Wirel. Commun.* **1**(4), 660–670 (2002)
5. O. Younis, S. Fahmy, HEED: a hybrid, energy-efficient, distributed clustering approach for ad hoc sensor networks. *IEEE Trans. Mob. Comput.* **3**(4), 660–669 (2004)
6. S. Lindsey, C. Raghavendra, K.M. Sivalingam, Data gathering algorithms in sensor networks using energy metrics. *IEEE Trans. Parallel Distrib. Syst.* **13**(9), 924–935 (2002)
7. W. Choi, P. Shah, S.K. Das, A framework for energy-saving data gathering using two-phase clustering in wireless sensor networks, in *Proceedings of Int'l Conference on Mobile and Ubiquitous Systems: Networking and Services (MOBIQUITOUS)* (2004), pp. 203–212
8. S. Soro, W. Heinzelman, Prolonging the lifetime of wireless sensor networks via unequal clustering, in *Proceedings of the 19th IEEE International Parallel and Distributed Processing Symposium (IPDPS)* (2005)
9. M. Ye, C. F. Li, G.H. Chen, J. Wu, EECS: an energy efficient clustering scheme in wireless sensor networks, in *Proceedings of IEEE International Performance Computing and Communications Conference (IPCCC)* (2005), pp. 535–540
10. B. Muthu, C.B. Sivaparthipan, G. Manogaran, R. Sundarasekar, S. Kadry, A. Shanthini, A. Dasel, IOT based wearable sensor for diseases prediction and symptom analysis in healthcare sector. *Peer-to-Peer Netw. Appl.* **13**(6), 2123–2134 (2020). <https://doi.org/10.1007/s12083-019-00823-2>
11. B. Liu, V. Pham, N. Nguyen, A virtual backbone construction heuristic for maximizing the lifetime of dual-radio wireless sensor networks, in *2015 International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP)*, Adelaide, SA, Australia (2015), pp. 64–67. <https://doi.org/10.1109/IIH-MSP.2015.20>
12. C.Y. Chang, H.R. Chang, Energy aware node placement, topology control and MAC scheduling for wireless sensor networks. *Comp. Netw.* **52**, 2189–2204 (2008)
13. X. Gu, J. Yu, D. Yu, G. Wang, Y. Lv, ECDC: an energy and coverage-aware distributed clustering protocol for wireless sensor networks. *Comp. Electr. Eng.* **40**, 384–398 (2014)

14. J. Mao, Z. Wu, X. Wu, A TDMA scheduling scheme for many-to-one communications in wireless sensor networks. *Comp. Commun.* **30**, 863–872 (2007)
15. P. Ayona, A. Rajesh, Investigation of energy efficient sensor node placement in railway systems. *Eng. Sci. Technol. Int. J.* (2015). <https://doi.org/10.1016/jestch.2015.10009>
16. V. Kaundal, A.K. Mondal, P. Sharma, K. Bansal, Tracing of shading effect on underachieving SPV cell of an SPV grid using wireless sensor network. *Eng. Sci. Technol. Int. J.* **18**, 475–484 (2015)
17. S. Bandyopadhyay, E.J. Coyle, An energy efficient hierarchical clustering algorithm for wireless sensor networks, in *IEEE INFOCOM* (2003), pp. 1713–1723

# Classification and Prediction of Leaf Diseases in Grape Using Optimized CNN



J. Sujithra and M. Ferni Ukrat

**Abstract** In agriculture, some nature of leaf categorizes the degree of greatness or a region of being free from deficiencies, and significant varieties. Likewise, the ailments in leaves have hazards to the financial, and production growth in the farming sector all over the world. The spotting of affected areas in leaves is identified using digital image processing, diminishes the reliance on the farmers for the security of farming outcomes. Thus, leaf bug disclosure and analysis is the urge of the advanced work. Toward this paper, the optimized CNN technique is proposed to identify and classify grape diseases. This model achieves the correct classification rate of 92% for 2482 images from 4 classes such as Anthracnose, Powdery Mildew, black Rot, and healthy.

**Keywords** CNN · Stochastic gradient descent (SGD) · Cross-entropy · Learning rate · Grape diseases

## 1 Introduction

The characterization of plants into fitting logical scientific categorizations is of uncommon excitement for specialists in different fields like agronomists, ecological defenders, foresters, land chiefs, and beginner cultivators. Normally, plant classification attempt made by researchers relies upon visible appraisal of plant natural items, herbal portion, and leaflets. For foods grown from the ground appear on plants for a little timeframe, people can't be practiced only as a rational segment to observe plant varieties. The leaflets-based exposure is dynamically strong toward visible brief based plant distribution and detection. Most of the image processing

---

J. Sujithra (✉) · M. Ferni Ukrat

School of Computing, SRM Institute of Science and Technology, Chennai, India

e-mail: [sj1090@srmist.edu.in](mailto:sj1090@srmist.edu.in)

M. Ferni Ukrat

e-mail: [ferniumkm@srmist.edu.in](mailto:ferniumkm@srmist.edu.in)

techniques provide effective results in disease detection [1]. Regardless, this proximity of the huge amount of plant varieties gives it a very effective undertaking to distinguish and perceive all plant species over a hand-operated appraisal. The classification of the various kinds of plants can be refined by their unique leaf surface, pattern, vein structure, shading, and so on. The tone of the leaf may change in specific circumstances with a correction in the atmosphere yet their key shape, surface, and vein configuration remain about identical. Those extraordinary leaves have formed excitement for the pros to make machine-based computerized plant species ID framework. The current record describes several sorts like highlights that will be used for disease identification. Certain characteristics fuse circularity, rectangularity, unexpectedness, completeness, and a point of view extent-based features expelled of the leaf.

## 2 Literature Review

In [2], authors have explained traits that depend on text pattern, color, Fourier lemma, wavelets, centroid, contour block, vein composition, and sawtooth designs and utilized twelve traits excluded of the Flavia dataset to analyze the crop varieties. The most skilled method to deliver leaf-based characteristics to frame the last set is the extraordinary variety in the literature and that could be done concerning supplementary analysis between various crop varieties. In [3], an appended review conducted with Fourier descriptors, syntactical characteristics, including state and dimension characterizing features for the consideration of plant identification. The geometric syntactical traits adjacent source characteristics obtained related in plant analysis [4] to recognize crop varieties amidst vital precision.

The features are applied to analyze leaves images on various plant varieties. Multi-spectral image processing techniques were also used to enhance the quality of identification by setting threshold values [5]. Distinctive AI classifiers were evaluated with the help of algorithms like decision tree, Naive-Bayes, Multi-SVM, and k-nearest Neighbor. The additional 5-advanced algorithm is exhibited for the association of plant varieties in leaves images. In [2] utilized 1600 leaf pictures from the “Flavia” dataset and the addition of 625 pictures from a customized dataset to verify the speculation of the proposed approach. K means algorithm and canny detectors were used for extracting useful feature vectors from leaves [6] to obtain the higher accuracy. New spectral indices were employed in [7] to monitor different diseases on crops.

**Fig. 1** Anthracnose

### 3 Implementation

#### 3.1 Training and Testing

A wide database containing 2482 images of leaflets from normal and diseased leaves held practiced to this training and testing of these Convolution models. ImageNet has variable-resolution images [8–10], it was down sampled to obtain fixed resolution for entire dataset. An introductory transcription about the database, including a less amount of pictures, is illustrated in [11]. The entire database held at first separated into three datasets, the training set, validation set, and the testing set, by arbitrarily dividing this 2482 pictures so that 70% of them performed the training set, 20% of them made the validation including 10% made in the testing. Consequently, during the training in convolution models, 1530 images were obtained, while some residue of 956 images were made for testing the achievement of the models in incorporating unique, old “unseen” images. The benefit of practicing greyscale variants of the images for training implied unrecognized, as past works [12] become revealed that methodology scales significantly better in the training of convolutional neural networks across multiple GPUs. Figures 1, 2, and 3 present about the disease presentation on grape leaves such as Anthracnose, Powdery Mildew, and Black Rot.

### 4 Results and Discussion

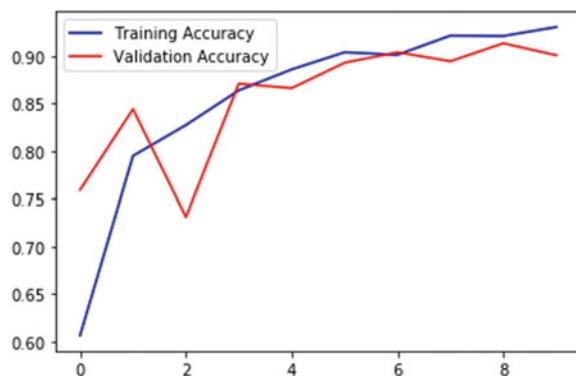
At the point of the inspection, we examined outlined CNN [13] specifically designed to deal with the variability of 2D shapes. CNN organize frameworks creates on a database containing not actually or equal the initial investment with 2482 pictures of grape unhealthy and healthy samples. The prominent performance on crop disease

**Fig. 2** Powdery Mildew**Fig. 3** Black rot

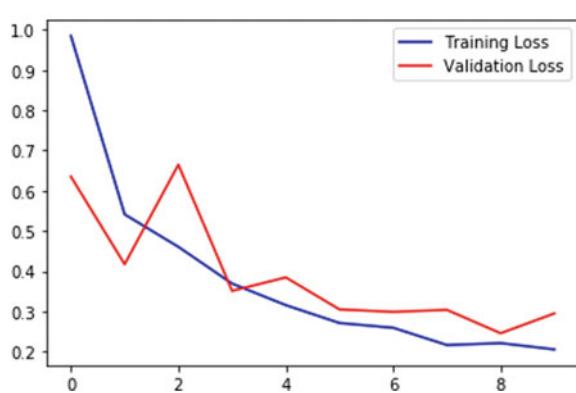
detection was achieved by Convolutional Neural network than other techniques [14]. These photos are gotten with various estimations, positions, positions, backcloths, and illuminating. The trained parameters like coefficients and bias [15] control the effect of squashing the non-linearity. The basic qualities of the difficulties transpired by the convolutional layers solely by the part extraction method and redesigned sometime later. To obtain the quality in identification and classification, the depth and width of the network should be increased as stated in [16]. The vibrant layers regarding layers abandon features of an increasingly inflated proportion of decision. Back propagation was carried out for adjusting weights in the network [17]. Some measures of images toward happening as intended zones are according to the going with  $224 \times 224$  (fascinating picture),  $224 \times 224$ , and finally  $64 \times 64$ . That exists undeniably not obscured that the tremendous framework sees whereby to designate leaves, debasements, distorts in images (low-level features). In Figs. 4 and 5, the precision and loss of both approval and testing can be shown utilizing the graph.

By taking some small dataset like unhealthy and healthy examples, the overfitting happens. I utilized SGD for enhancement with the learning rate of 0.01, decay

**Fig. 4** Efficiency of training and validation



**Fig. 5** Lack of training and validation



is  $1e-6$ , and momentum is 0.9. The loss function has been utilized to ascertain the misfortune utilizing cross-entropy.

## 5 Conclusion

Presently, horticulture is affected by some plant deficiencies, which decreases both the quantity and quality. Toward the end, we have an intense learning way had been proposed to deal with grape leaves disease characterization and analysis. Far from a conventional approach, the instruments for diagnostics are lacking in some backward countries. For feature extraction, DCNNs utilize trainable convolution layers as an extractor. The proposed model can help producers to distinguish the diseases in the grape plant as a choice. The image of leaves is required to determine the sort of disease. With the help of a smartphone, it can be done. The fundamental goal is to recognize three famous varieties of grape diseases which are Anthracnose, Powdery Mildew and Black rot by applying deep neural networks. For an ongoing situation,

the experimental outcomes have explained the fitting methodology. Future work may test with a large number of specimens to increase the accuracy of the proposed model.

## References

1. S.K. Tichkule, Prof. D.H. Gawali, Plant diseases detection using image processing techniques, in *2016 Online International Conference on Green Engineering and Technologies (IC-GET)*
2. S.G. Wu, F.S. Bao, E.Y. Xu, Y.X. Wang, Y.F. Chang, Q.L. Xiang, A leaf recognition algorithm for plant classification using PNN (2007). Retrieved from <http://flavia.sourceforge.net/>
3. A. Aakif, M.F. Khan, Automatic classification of plants based on their leaves. Biosyst. Eng. **139**, 66–75 (2015)
4. V. Satti, A. Satya, S. Sharma, An automatic leaf recognition system for plant identification using machine vision technology. Int. J. Eng. Sci. Technol. **5**(4), 874 (2013)
5. D. Cui, Q. Zhang, M. Li, G.L. Hartman, Y. Zhao, Image processing methods for quantitatively detecting soybean rust from multispectral images. Elsevier **22**(4), 186193 (2010)
6. R. Raichaudhuri, R. Sharma, On analysis of wheat leaf infection by using image processing, in *Proceedings of the International Conference on Data Engineering and Communication Technology, Advances in Intelligent Systems and Computing*, vol. 1 (2016), pp. 978–981
7. W. Huang, Q. Guan, J. Luo, J. Zhang, J. Zhao, D. Liang, L. Huang, D. Zhang, New optimized spectral indices for identifying and monitoring winter wheat diseases. IEEE J. Sel. Topics Appl. Earth Observ. Rem. Sens. **7**(6) (2014)
8. M. Anbarasan, B. Muthu, C. Sivaparthipan, R. Sundarasekar, S. Kadry, S. Krishnamoorthy, A.A. Dasel, et al., Detection of flood disaster system based on IoT, big data and convolutional deep neural network. Comput. Commun. **150**, 150–157 (2020). <https://doi.org/10.1016/j.comcom.2019.11.022>
9. D. Vu, T. Nguyen, T.V. Nguyen, T.N. Nguyen, F. Massacci, P.H. Phung, A convolutional transformation network for malware classification, in *2019 6th NAFOSTED Conference on Information and Computer Science (NICS)*, Hanoi, Vietnam (2019), pp. 234–239. <https://doi.org/10.1109/NICS48868.2019.9023876>
10. A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in *Advances in Neural Information Processing Systems* (2012)
11. D.P. Hughes, M. Salathé, An open access repository of images on plant health to enable the development of mobile disease diagnostics (2015). [arXiv:1511.08060](https://arxiv.org/abs/1511.08060)
12. A. Krizhevsky, One weird trick for parallelizing convolutional neural networks (2014). [arXiv: 1404.5997](https://arxiv.org/abs/1404.5997)
13. Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
14. J. Sujithra, M.F. Ukrat, A review on crop disease identification and classification through leaf images. Eur. J. Molecular Clin. Med. **7**(09), 2020
15. Y. LeCun, Y. Bengio, Convolutional networks for images, speech, and time series, in *The Handbook of Brain Theory And Neural Networks*, vol. 3361(10) (1995)
16. C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, et al., Going deeper with convolutions, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015)
17. M. Jhuria, A. Kumar, R. Borse, Image processing for smart farming: detection of disease and fruit grading, in *Proceedings of the 2013 IEEE Second International Conference on Image Information Processing (ICIIP-2013)*

# Diabetic Retinopathy Image Segmentation Using Region-Based Convolutional Neural Network



D. Vanusha and B. Amutha

**Abstract** Diabetic retinopathy is a retinal disease that causes permanent blindness. Without earlier detection, the patients suffer from physical conditions associated with long-term diabetes, and it also affect retina of the eye part. Deep learning can help ophthalmologist in the field of artificial intelligence and also the alternate option regarding the classification of the DR by using an automatic classifier. The classification of DR can be trained accurately in the deep learning process and it needs excessive number of images which is essential in DR limitations. Deep learning technique achieves efficient performance output on the Image Net high-scale visual recognition competition to challenge image classification. Inception-V3 model is an example for the challenge. Transfer learning is a traditional approach that helps in overcoming issue of working from scratch and helps in improving the accuracy based on the use case using standard architectures, which has achieved state-of-the-art results. The main idea to suit the DR dataset is transfer-learning, in which the deep learning architecture mainly aims on DR separation by using deep learning techniques to address the problem. To find the DR classification, this output can help the researchers to detect the transfer learning techniques and show their differences by Convolution Neutral Networks and KNN which derives the data from the training model and execute it in later stage as an output. In this deep learning technique, CNN and KNN are the existing methods, whereas RCNN is the proposed method.

**Keywords** CNN · KNN · Transfer learning · RCNN

---

D. Vanusha (✉) · B. Amutha

Department of CSE, SRM Institute of Science and Technology, Chennai, India

e-mail: [vanushad@srmist.edu.in](mailto:vanushad@srmist.edu.in)

B. Amutha

e-mail: [amuthab@srmist.edu.in](mailto:amuthab@srmist.edu.in)

## 1 Introduction

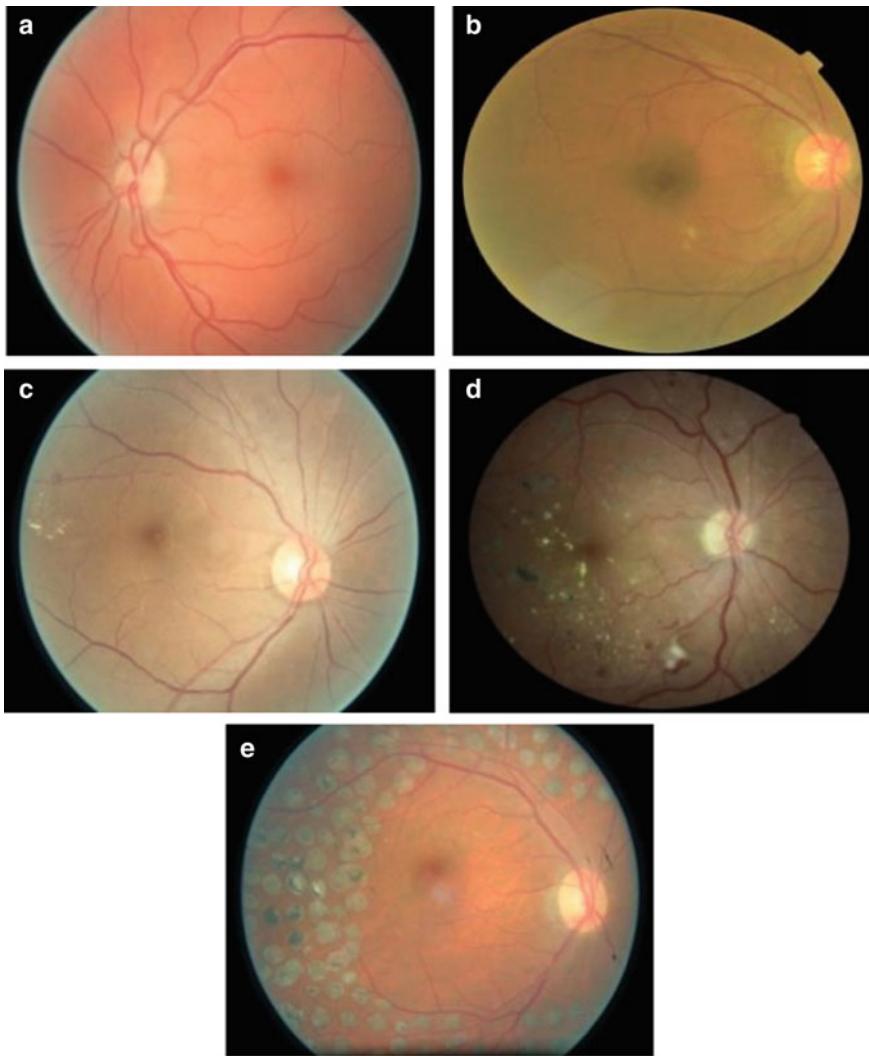
Diabetic retinopathy is the micro vascular complications that are caused by diabetic mellitus, which is dangerous, metabolic, clinically heterogeneous disorder increasing every day all over the world. The factors of DM follow as frequently hyperglycemia, which may be due to impaired insulin secretion. Due to diabetic mellitus, metabolic aberrations with life-threatening health complications like micro vascular (retinopathy, nephropathy, and neuropathy) and macro vascular complications, followed by short and long-term complications occurs. DR is the major complications of blindness in old age person. About 5.1 million adults had DR and 645,200 had vision-loss due to DR. When the blood veins inside the retina are affected by large blood levels this DR may occur. In 2011, about 116.6 million peoples were affected by DR, and it may be estimated that it can increase to 181 million by 2040. More than 3.7% of blindness all over the world also due to occurs DR.

To prevent blindness starting stage identification and classification of DR is the challenging one. Most of the people could not sense that they have DR until the disease affects their vision, which usually happens in the last stage. Finally, they might not go through treatment properly. There are a few techniques that have been performed for the classification of retinal image that utilizes conventional neural network techniques. Wilkinson classified DR into five stages based on disease severity as given in Fig. 1.

In the first stage, there is no detectable retinopathy occurs, and no damage occurs in the retina by diabetic mellitus. Mild DR is the second stage where few micro aneurysms occur. Moderate DR is the third stage where multiple micro aneurysms occur with blot hemorrhages or cotton wool spots. The fourth stage is called severe DR where intraregional micro vascular abnormalities, venous beading and cotton wool spots occurs. Proliferative DR is the final stage where new retinal blood vessels start and blood leaks into the vitreous humor of the eye, and retinal detachment appears.

Machine learning techniques come under the method of deep learning techniques. With the aim to identify the significant low-level image, deep learning method classifies with hidden technique which is opposite to the traditional conventional neural network (CNN). Deep learning utilizes the characteristics to clear the different problems in the same situation. Advantages of transfer learning follows as it saves computational time because it uses data from the previous method instead of training the new method from the elimination it also extends the knowledge from the previous one and very helpful when the dataset from the new model is small. Transfer learning contributions to the fields of classification of audio, processing of natural language and computer vision. In the ancient period, convolution neutral network (CNN) was developed for the classification of image.

To address the image classification, Krizhesky et al. introduced CNN to achieve the performance in 2011. Image classification, natural language processing and analysis of test have been done using CNN. In every case, deep learning uses the weight from the scratch requires more time and dataset which includes number of images.



**Fig. 1** Funds images: **a** normal, **b** mild, **c** moderate, **d** severe, **e** prolific DR

Deep learning makes more time consuming compared to transfer learning. To address the classification of DR, we have utilized the descriptors such as CNN, KNN, and Transfer Learning.

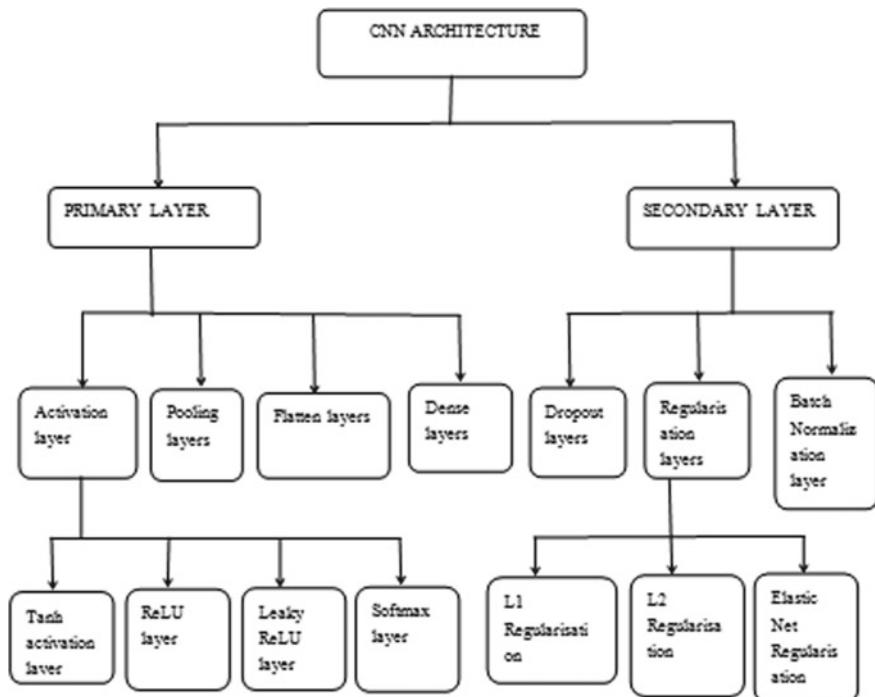
## 2 System Analysis

### 2.1 Existing Method

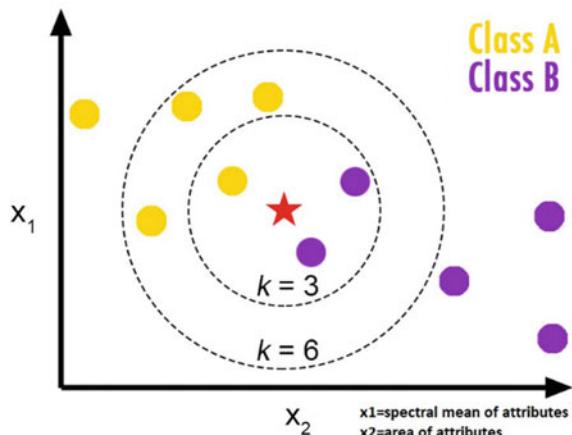
In system analysis of Diabetic Retinopathy, we discussed three existing method to classify the image of DR as given below. They are convolution Neural Network-Nearest Neighbor, and Transfer Learning.

#### 2.1.1 Convolutional Neural Network

CNN were divided into two layers as Primary and Secondary layers. The first one primary is the main layers which include convolution, activation, pooling, flatten, and dense layers. The last one is secondary are the substitutable type used to make CNN harder, and its increases the complexity [1–3] (Fig. 2).



**Fig. 2** CNN architecture

**Fig. 3** K-Nearest Neighbor

### 2.1.2 KNN

K-Nearest Neighbor is the simple algorithm in the classification of image comparing with another algorithm. It is simple as it stores the data from training models on that training stage and extracts the information to a new on the prediction stage. It measures the distance like Euclidean distance, Murkowski distance or distance. This algorithm uses K as close to the dominant models [4]. The dominant model is fixed at prediction stage as a new distance as we discussed above. A little and compact change in K results in effective structure of the image, however, larger value of K denotes less sensitive. The training dataset is classified into two types

1. Structure less NN techniques
2. Structure based NN techniques (Fig. 3).

The dataset is divided into two parts as training dataset and test dataset. The distance between the training point and sample point is estimated and calculate with lowest distance is called as the nearest neighbor. The algorithm K can be calculated using the formula given below

$$K = \sqrt{(e_1 - f_1)^2 + (e_2 - f_2)^2 + \dots + (e_n - f_n)^2}$$

### 2.1.3 Transfer Learning

Transfer learning is a deep learning technique which is used to train a CNN accurately in which its weights are not initialized from scratch. Deep learning is division of machine learning technique which trains data and computes the power by solving classification and backsliding [5–7].

Transfer learning gained a stress because of the insufficiency of defined training data to create models. Transfer learning has the advantage of minimizing the training data and training time.

In this paper, transfer learning implements the Image Net dataset which was pre-trained on deep learning model of CNN method. This pre-trained inception V3 model has been initialized and divides the fundus image into active and inactive classes. Keras library imports the pre-trained model.

A randomly taken subsample of the EyePacs DR model uses Kaggle dataset to train and test the model. Four procedures used in transfer learning, where first procedure is to eliminate the layers dividers and stabilize the network weights and uses CNN features to extract and add a fully connected layer like a support vector machine (SVM) [8]. The second procedure is to eliminate the originally connected layers by using little learning rate (LR) and add layer which is classified to fulfill the new task. The third procedure is to eliminate fully connected layer and adjust the top layer, while keeping the bottom layer stable [9]. The fourth procedure is to train the architecture from scratch and to examine the challenging dataset.

### 3 Literature Review

Based on the different aspects, this section discusses the aspects used after transfer learning which are shown as follows the dataset of the image used, the architecture used, the LR used and the fine-tuning process.

A set of weight from the last training process was learned to differentiate another dataset of the image in transfer learning. This is the reason that the author uses another transfer learning to find the DR.

The two grades DR and NO DR have been classified using Inception V3 architecture by Gulshan et al. The author contains 137,175 images which were carried out in the database. From the two datasets tested, one set with the size of 9943 have the sensitivity of 96.5% and another one set with the size of 1648 have 94.1% sensitivity [10].

Li et al. discussed the transfer learning by comparing different network architectures, including Alex Net, VGG-S, VGG16, and VGG19 to two datasets as the Messidor and DR1 datasets, respectively [11].

Mohammadian et al. compared the classification of dataset by using kaggle with inception V3 and exception [12]. To test the algorithm's performance against unseen data, the author uses with 30% of the image from the whole dataset of 35,126 images. Masood et al. classified DR into five stages by using the kaggle dataset to achieve the performance of inception V3 model. Out of 4500 images the author reduced it to 450-pixel image which gives accuracy of 47.8% [8].

Private dataset model change the Google Net architecture which was introduced by Takahashi et al. They used the model of about 9423 images and 476 images to test the process [9]. They used four-class classification technique to crop the image of  $1172 \times 1172$  pixels. The accuracy of the result was 81% and the kappa score as 0.84.

Choi et al. reviewed the issue of transfer learning on the STARE dataset, and they used image classification method to increase the size of the object to 10,000 images along with retina disorder which includes DR. The author decided for VGG19 and Alex Net architecture which was pre-trained model, and in concert it was introduced to develop the accuracy of the network and K-fold validation with  $K = 4$  was taken to ensure the results [13, 14]. In this case, VGG19 architecture achieves highest accuracy compared with Random Forest (RF) classifier.

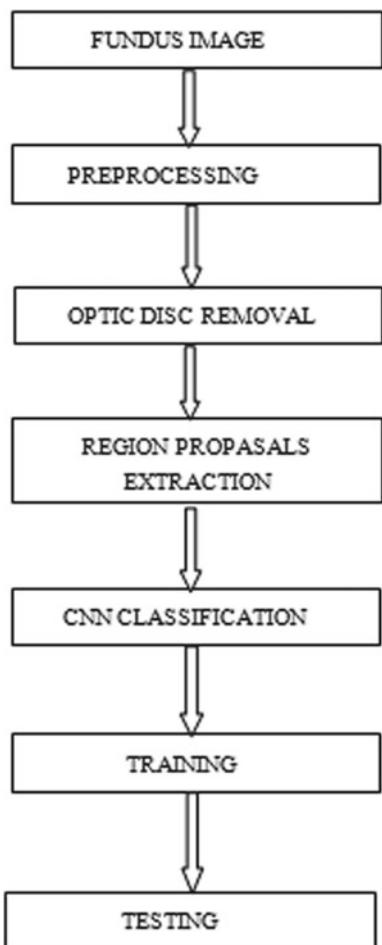
Wang et al. reviewed the transfer learning method by using three network structures as given by: Alex Net, VGG16, and InceptionNetV3. The author uses kaggle dataset to compose the algorithm into 166 images. The authors preferred to five-stage classification approach instead of binary classification approach which were used by other authors in the dataset. They employed a stochastic gradient descent optimizer with Nesterov momentum to run the convergence to the smaller one. The authors reduced the images for each structure into  $237 \times 237$  for Alex Net,  $214 \times 214$  for VGG16, and  $279 \times 279$  for InceptionV3 by using image classification method. Mansur extracted the features of transfer learning by using kaggle dataset to train the deep learning and CNN methods, while establishing the computer aided diagnosis. Lam et al. examined the sliding windows algorithm, where original images taken from small patches are used [15, 16]. These patches contain the salient characteristics such as the presence of micro aneurysms. The authors used the Kaggle dataset to develop these patches. They developed 1424 patches from 245 images and divided the patches to training and testing datasets. They tested the proposed algorithm consists of 195-pixel images by using the E-Ophtha dataset. They trained the dataset with input size  $148 \times 148$  using Google Net architecture. They normalized and resized the test images into  $2248 \times 2248$  pixel to test the model. Finally, the trained model divides the tested image to create a map with a possible score for all five grades. Tsighe et al. reviewed InceptionV3 architecture to find DR by using the Kaggle dataset. The author extract 2500 images and divide the image into  $300 \times 300$  to train the model, and 5000 images were taken to test the model. Zeng et al. described a novel Siamese-like architecture to divide the left and right fundus images. The author divides right and left eye with 28,404 images with testing of 7124 images using kaggle dataset and train the inception V3 on the image dataset. Gao et al. used normal dataset of 4476 images with four classes. The authors split images into four partitions of  $200 \times 200$  images that were the input of four Inception V3 networks, and then they integrate into a single layer as a result [17].

## 4 Proposed Method

### 4.1 RCNN

Wide applications of CNN as shown as follows, image processing, pattern recognition, and video recognition. In image classification of CNN, a picture can be taken

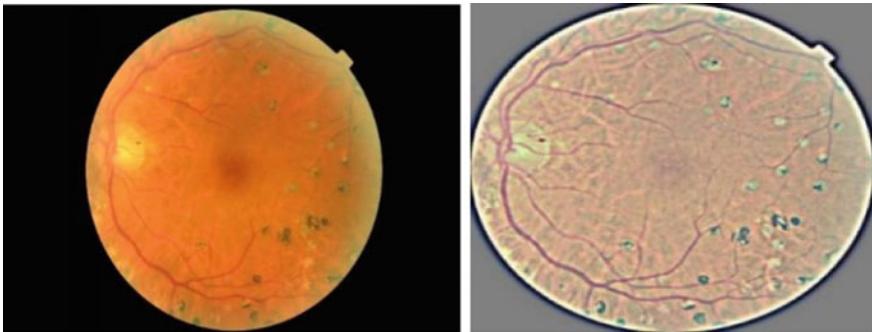
**Fig. 4** Representation of methodology



as a input, and later it is classified into suitable purpose. In R-CNN, the image is divided into various regions (or segments) and also CNN is urged to pursue in these segments [18–22]. Due to extraction of interest, accuracy of the finding the object is very high in the CNN. In image classification, CNN use image as a input and divides it into required class of object as an output (Fig. 4).

#### 4.1.1 Fundus Image

At the beginning, the fundus image should be resized into the dimension of  $120 \times 120$  pixel due to large data and giving dimensions to the picture with sensible value from fundus cameras preprocessing which is essential. If the image were not preprocessed, then the picture remains twisted and distortion occurs [23–25]. Due to various fundus



**Fig. 5** Original fundus image (Left), cropped and pre-processed fundus image (Right)

cameras, there will be an unreliable illumination in the image appears. To stabilize the illumination, standardization method will be used. The image contains red, green, and blue channels. For further process, green channel should be extracted. Green channel extraction gives higher variation among the other two channels, and it is less noise [26, 27] (Fig. 5).

#### 4.1.2 Optic Disc Removal

The removal of optic disk is very essential process as it provides higher intensity of the images. Initially, the images with range values between 0 and 255 should be changed into grayscale value and threshold method follows. Lower intensity images are changed into 0 which is black and the images having higher intensity values are changed into 255 which are white [28–31]. Until the regional image occurs, thresholding should be done continuously.

#### 4.1.3 Region Proposals Extraction

Threshold method is used to extract the image preprocessing (Fig. 6).

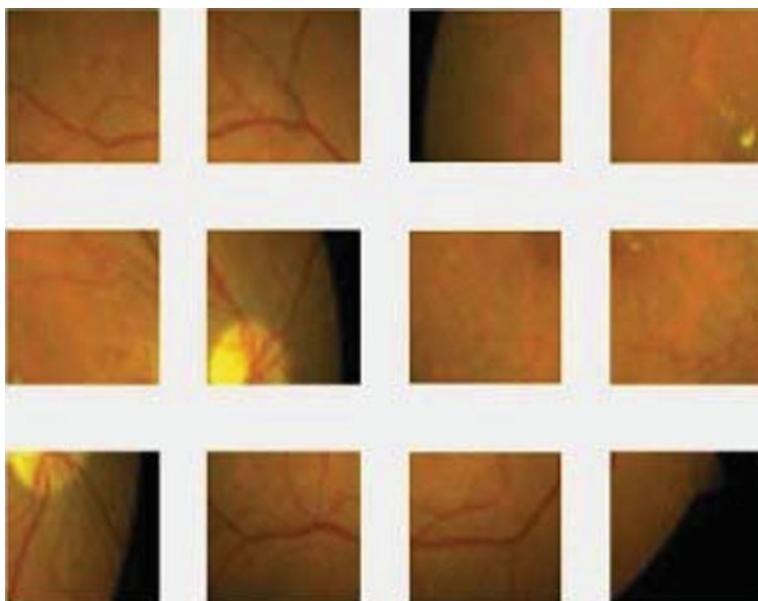
The images are split into equal segments of size  $128 \times 128$  pixel, and the imagined blocks are extracted to have the DR indication which is used for training method (Fig. 7).

#### 4.1.4 CNN Classification

Image pixel value is same as the number of neurons in the input layer. The convolution layer figures out the image patches with filter as an important feature. ReLU (Rectified Linear Unit) is used in the activation layer which enforces the threshold method to



**Fig. 6** Threshold-based method

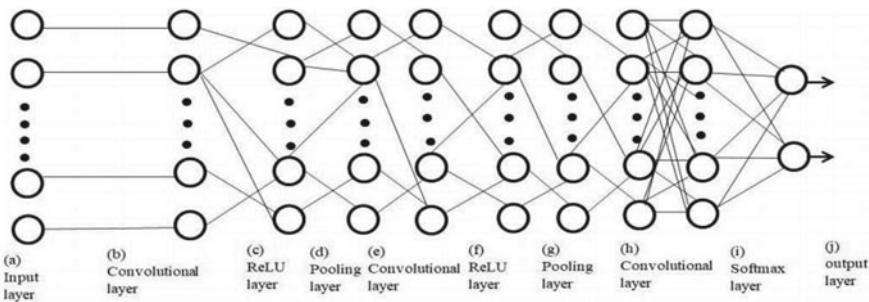


**Fig. 7** Pixel-based blocking

each input. The pooling layer changes the volume to eliminate the data and to increase the computation. Final layer of CNN is softmax layer (Fig. 8).

#### 4.1.5 Training

Multiple models in deep learning have achieved greater performance when subset of Image Net dataset was trained the model with convolution section have already trained to extract the characteristics from Image Net dataset (Table 1).

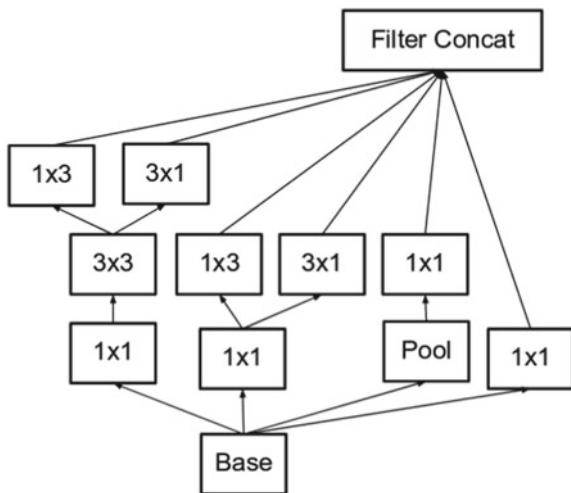
**Fig. 8** CNN layers**Table 1** Comparison of existing method and proposed method

Points of comparison	Existing method	Proposed method
Size of training data	More than 35,126 fundus images	2500 fundus images
Optimiser	CNN	RCNN
Learning rate	0.0001	0.0005
Loss function	Not specified	Cosine loss function
Data augmentation used	Yes	No
Accuracy (%)	85.7	94.6
Loss	Not specified	4.32%

In this method, we used convolution part of pre-trained Inception-V3 to extract characteristics of fundus images. Szegedy et al. introduced Inception-V3 model in inception Networks by adding convolution factorization and auxiliary classifiers. Figure 9 shows the inception modules are a group of various sized filters with same nature of convolution on an input image.

To classify the extracted features of the image, we have joined softmax layer along with activation layer. For training Stochastic Gradient Descent (SGD) and ascending learning rate of 0.0005 was used, where softmax layer taken as output layer. To determine the error in the dataset, cosine loss function used because it gives greater performance with small data.

**Fig. 9** Inception module used in Inception-V3



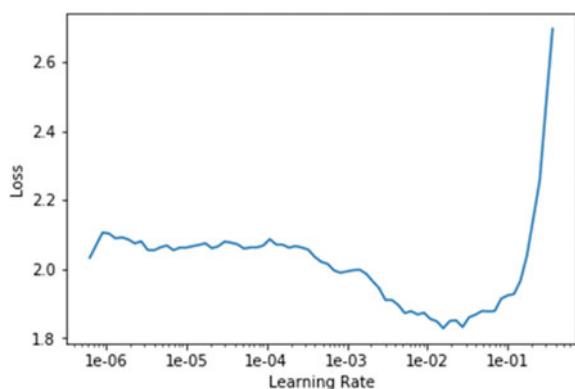
#### 4.1.6 Testing

On testing, 4000 fundus images were tested from the previous data as a result 93.9% accuracy have achieved with 2.94% loss. The trained model achieves greater efficiency compared with pre-trained inception V3 dataset and to divide kaggle dataset into binary classes.

## 5 Results

The accuracy of proposed method is related with KNN where proposed RCNN have higher accuracy (Fig. 10).

**Fig. 10** Graph comparison



## 6 Conclusion

In this paper, image classification of DR into binary classes of data have been created by using transfer learning method which have the advantage of reduced trained data compared with earlier classification technique and higher performance. In medical, image classification employs CNN to explain image with limited size of dataset which is challenging one for the skilled doctors. To overcome this challenge, transfer learning can be a suitable option with limited dataset to classify the image task. In the discipline of medical images to overcome the functions to computer fabrication complex and deep architectures are being used. To classify DR images, this paper defines CNN on focusing transfer learning techniques even many architectures and various method have been introduced, transfer learning and KNN have many advantages as we discussed here. To evaluate performance of pretrained, deep convolution network exercises have to be made in classification of DR images.

## References

1. N.H. Cho, J. Shaw, S. Karuranga, Y. Huang, J.D. da Rocha Fernandes, A. Ohlrogge, B. Malanda, IDF diabetes atlas: global estimates of diabetes prevalence for 2017 and projections for 2045. *Diabetes Res. Clin. Pract.* **138**, 271–281 (2018)
2. A. Gupta, R. Chhikara, Diabetic retinopathy: present and past. *Procedia Comput. Sci.* **132**, 1432–1440 (2018)
3. Y. Zheng, M. He, N. Congdon, The worldwide epidemic of diabetic retinopathy. *Indian J. Ophthalmol.* **60**, 428 (2012)
4. R. Bourne, G.A. Stevens, R. White, J.L. Smith, S.R. Flaxman, H. Price, J.B. Jonas, J. Keeffffe, J. Leasher, K. Naidoo et al., Causes of vision loss worldwide, 1990–2010: a systematic analysis. *Lancet Glob. Health* **1**, e339–e349 (2013)
5. P. Vashist, S. Singh, N. Gupta, R. Saxena, Role of early screening for diabetic retinopathy in patients with diabetes mellitus: an overview. *Indian J. Community Med.* **36**, 247–252 (2011)
6. T.M. Mitchell, *Machine Learning*, 1st ed. (McGraw-Hill Inc, New York, NY, USA, 1997). ISBN 0070428077
7. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning*. (The MIT Press, Cambridge, MA, USA, 2016). ISBN 9780262035613
8. Y. LeCun, B. Boser, J.S. Denker, D. Henderson, R.E. Howard, W. Hubbard, L.D. Jackel, Backpropagation applied to handwritten zip code recognition. *Neural Comput.* **1**, 541–551 (1989)
9. A. Krizhevsky, I. Sutskever, G. Hinton, ImageNet classification with deep convolutional neural networks, in *Neural Information Processing Systems*, vol. 25. (Curran Associates Inc, Lake Tahoe, NY, USA, 2012)
10. N. Kalchbrenner, E. Grefenstette, P. Blunsom, A convolutional neural network for modelling sentences. arXiv [arXiv:1404.2188](https://arxiv.org/abs/1404.2188) (2014)
11. Z. Wang, W. Yan, T. Oates, Time series classification from scratch with deep neural networks: a strong baseline, in *Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN)*, Anchorage, AK, USA, 14–19 May 2017, pp. 1578–1585
12. B. Xu, N. Wang, T. Chen, M. Li, Empirical evaluation of rectified activations in convolutional network. arXiv [arXiv:1505.00853](https://arxiv.org/abs/1505.00853) (2015)
13. G. James, D. Witten, T. Hastie, R. Tibshirani, G.J. Trevor Hastie, D.W. Robert Tibshirani, *An Introduction to Statistical Learning: With Applications in R*. (Springer Publishing Company: Berlin/Heidelberg, Germany, 2014). ISBN 9781461471370

14. S. Iofffffe, C. Szegedy, Batch normalization: accelerating deep network training by reducing internal covariate shift, in *Proceedings of the 32nd International Conference on Machine Learning*, Lille, France, vol. 37. (ACM: New York, NY, USA, 2015), pp. 448–456
15. G. Huang, Z. Liu, L. Maaten, K.Q. Weinberger, Densely connected convolutional networks, in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 21–26 July 2017, pp. 2261–2269
16. F. Chollet, Xception: deep learning with depth wise separable convolutions, in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 21–26 July 2017, pp. 1800–1807
17. R. Pires, H.F. Jelinek, J. Wainer, E. Valle, A. Rocha, Advancing bag-of-visual-words representations for Lesion classification in retinal images. *PLoS One* **9**, e96814 (2014). *Appl. Sci.* **10** (2020), 2021 23 of 24
18. X. Li, T. Pang, B. Xiong, W. Liu, P. Liang, T. Wang, Convolutional neural networks based transfer learning for diabetic retinopathy fundus image classification, in *Proceedings of the 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, Shanghai, China, 14–16 October 2017, pp. 1–11
19. S. Mohammadian, A. Karsaz, Y.M. Roshan, Comparative study of fine-tuning of pre-trained convolutional neural networks for diabetic retinopathy screening, in *Proceedings of the 24th National and 2nd International Iranian Conference on Biomedical Engineering (ICBME)*, Tehran, Iran, 30 November–1 December 2017, pp. 1–6
20. M. Anbarasan, B. Muthu, C. Sivaparthipan, R. Sundarasekar, S. Kadry, S. Krishnamoorthy, A.A. Dasel, Detection of flood disaster system based on IoT, big data and convolutional deep neural network. *Comput. Commun.* **150**, 150–157 (2020). <https://doi.org/10.1016/j.comcom.2019.11.022>
21. N.T. Le, J.-W. Wang, C.-C. Wang, T.N. Nguyen, Automatic defect inspection for coated eyeglass based on symmetrized energy analysis of color channels. *Symmetry* **11**, 1518 (2019). <https://doi.org/10.3390/sym11121518>
22. H. Takahashi, H. Tampo, Y. Arai, Y. Inoue, H. Kawashima, Applying artificial intelligence to disease staging: deep learning for improved staging of diabetic retinopathy. *PLoS One* **12**, e0179790 (2017)
23. J.Y. Choi, T.K. Yoo, J.G. Seo, J. Kwak, T.T. Um, T.H. Rim, Multi-categorical deep learning neural network to classify retinal images: a pilot study employing small database. *PLoS One* **12**, e0187336 (2017)
24. X. Wang, Y. Lu, Y. Wang, W. Chen, Diabetic retinopathy stage classification using convolutional neural networks, in *Proceedings of the 2018 IEEE International Conference on Information Reuse and Integration (IRI)*, Salt Lake City, UT, USA, 6–9 July 2018, pp. 465–471
25. M.H. Johari, H. Abu Hassan, A. Ihsan Mohd Yassin, N. Tahir, A. Zabidi, Z. Ismael Rizman, R. Baharom, N. Wahab, Early detection of diabetic retinopathy by using deep learning neural network. *Int. J. Eng. Tech.* **7**, 198–201 (2018)
26. C. Lam, C. Yu, L. Huang, D. Rubin, Retinal lesion detection with deep learning using image patches. *Investig. Ophthalmol. Vis. Sci.* **59**, 590–596 (2018)
27. C. Lam, D. Yi, M. Guo, T. Lindsey, Automated detection of diabetic retinopathy using deep learning. *AMIA Jt. Summits Transl. Sci. Proc.* **2017**, 147–155 (2018)
28. M. Tsige Hagos, S. Kant, Transfer learning based detection of diabetic retinopathy from small dataset. *arXiv arXiv:1905.07203* (2019)
29. H. Chen, X. Zeng, Y. Luo, W. Ye, Detection of diabetic retinopathy using deep neural network, in *Proceedings of the International Conference on Digital Signal Processing (DSP)*, Shanghai, China, 19–21 November 2019, vol. 2018
30. X. Zeng, H. Chen, Y. Luo, W. Ye, Automated diabetic retinopathy detection based on binocular siamese-like convolutional neural network. *IEEE Access* **7**, 30744–30753 (2019)
31. W. Zhang, J. Zhong, S. Yang, Z. Gao, J. Hu, Y. Chen, Z. Yi, Automated identification and grading system of diabetic retinopathy using deep neural networks. *Knowl. Based Syst.* **175**, 12–25 (2019). *Appl. Sci.* **10** (2020), 2021 24 of 24

# A Secure and Reliable IoT-Edge Framework with Blockchains for Smart MicroGrids



Rijo Jackson Tom , Vivia Mary John , and G. Renukadevi 

**Abstract** Growing energy demands and the rise in global warming calls for effective management of electric grid and renewable energy integration into the grid. Internet of Things, along with cloud computing has made the monitoring and control of the grid effective. To reduce latency and provide real-time monitoring of the grid Edge computing is an effective solution. Growing concerns of cyber-attacks in the grid system demands for security measures to be considered while designing the microgrid system. This work proposes a secure IoT-Edge framework for smart microgrid systems. The framework defines how edge computing will play a role in the microgrid system. The computing layers includes a blockchain-based security layer. Blockchains have proved to be an effective solution for secure transactions. There are many frameworks available for implementing blockchain. This work discusses the use cases of blockchain in energy sector and advises the use of two blockchain frameworks: Energy Web Chain for operational methods and CordaR3 based Distributed Ledger Technology for energy trading and other financial transactions in the grid.

**Index Terms** Blockchain · Cloud computing · Edge computing · Energy trading · Internet of Things · Smart grids

## 1 Introduction

Microgrid systems have become an integral part of the smart grid systems [1]. The global demand for energy increases to a new level that there is a huge need to integrate renewable energy resources into the electricity grid system [2]. The present electricity grid needs to be upgraded with the latest technologies for efficient and reliable power delivery [3]. In the next three decades, the population will rise to 9 billion. To meet

---

R. J. Tom  · V. M. John

Department of Computer Science and Engineering, CMR Institute of Technology, Bengaluru,  
India

e-mail: [rijo.j@cmrit.ac.in](mailto:rijo.j@cmrit.ac.in)

G. Renukadevi  
Chennai, India

the global need and to decrease the carbon-footprint greener energy sources has to be sought [4].

Microgrid system is a smaller grid which has its own generation, transmission, and distribution system. The microgrids are connected to the major grid and at times can separate as island power system too. The power generation is mainly from renewable energy sources. The problem with the renewable energy sources is that they are not available always are sometimes seasonal dependent. So there has mechanisms for effective forecast and managing the supply from the renewable energy sources so that it does not disturb the regular grid operations. There can be frequency fluctuations in the grid when supply is more than the demand. The fact that effective storage mechanisms are not available for energy storage [5–7].

To reduce the carbon emission and to use greener fuels, electric vehicles (EV) are introduced, and they can act as an energy storage device. Vehicle to grid is an upcoming aspect of the microgrid system that needs precise monitoring and control. In the near future investments on electric vehicles, charging stations, solar PV, and other distributed generation will increase.

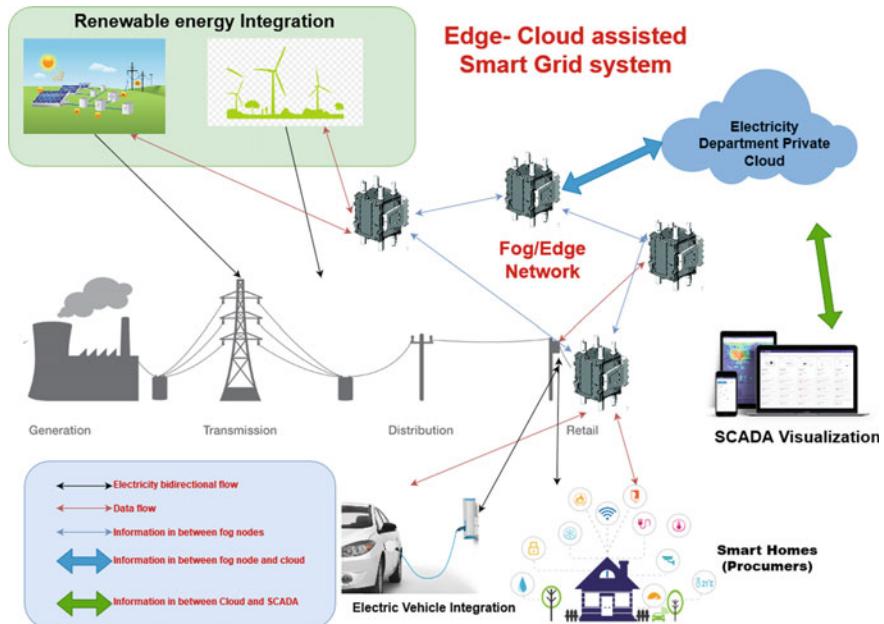
Internet of Things (IoT) and Fog/Edge Computing are the major role players in the monitoring and control of the electrical grid system. Researchers in Ardito and Procaccianti [8] have proved the effectiveness of using the Internet of Things into the electrical grid system.

Edge computing introduced by Cisco reduces the latency of analyzing the data from the end devices. The concept of edge computing is that near real-time analytics can happen at the edge and only necessary information can be passed on to the cloud does reduce the bandwidth usage in the cloud, removing the redundant data [9, 10].

The geographical distribution of the electrical grid system makes it difficult for a single centralized cloud system to monitor the grid the fair the geographically distributed nature of edge devices can make it easy for effective monitoring of the distribution system. Fog notes can now handle the data from the smart meters and also monitor the data that is emerging from the transformer's intelligent electronic devices (IEDs). Researches have shown that fog is an effective solution for or applications like smart grids [11] (Fig. 1).

Smart grids are now more prone to cyber-attacks, and there is a need to add security measures. IEC61850 provides necessary documents for power system automation. IEC62351 part 1–8 details and need for security in data communication with respect to power systems. IEEE P2418.5 blockchain in Energy Workgroup is working on an open interoperable standard for blockchain in Energy.

Blockchain framework is distributed and decentralized system which uses a highly complex cryptographic algorithm in order to make each transaction secured. It omits the third party, transparent and also verifiable. Energy trading deals with peer-to-peer network such that hyper ledger fabric permissioned blockchain is used. The framework helps in the growth of renewable energy trading. Various types of blockchain consensus algorithm are presently based on the kind of framework used. Consensus property avoids malicious users inside the blockchain. Few types of consensus are proof of work, proof of stake, byzantine fault tolerance, and etc. The participants



**Fig. 1** IoT-Edge assisted smart grid system. Electric vehicles and smart homes communicate with the Fog/Edge routers to and the fog routers communicate with the cloud

before joining in hyper ledger fabric permissioned blockchain are verified and authenticated. In peer-to-peer energy trading model, one consumer is benefited by the other consumer who produces surplus energy.

There are many frameworks for blockchain, and many industries have come up with many frameworks for use in energy sector. In June 2019, the Energy Web Foundation announced the first open-source blockchain Energy Web Chain specifically tailored for the energy sector. Corda R3 features a private Distributed Ledger Technology for financial transactions. Energy Block Exchange (EBX) is a CorDapp that can be used for energy trading.

This work proposes a secure Edge-Cloud framework for smart microgrid systems. The Edge node has an important role in the smart grid world. They can perform real-time analytics. Provide security as a service using blockchain and have Cloud–Edge interconnection.

Thus blockchain ecosystem in energy can integrate Utilities, consumers, renewable energy sectors, energy enterprises, electric vehicles, and generation plants.

The major contributions of the paper are as follows:

- Modeling an IoT-Edge Framework for smart microgrid system. The need for Edge-Cloud computing model for an application like smart grid.

- A secure edge computing framework is proposed where the edge computing can take care of real-time analytics connectivity to cloud and inter edge communication.
- The work proposes a CordaR3 Distributed Ledger Technology as a proposed solution for financial transactions in the energy grid system.

The rest of this paper is organized as follows. Section 2 discusses the related works. The IoT-Edge Framework for microgrids is detailed in Sect. 3. In Sect. 4, secure Edge computing framework and energy trading with blockchains is explained. Section 5, concludes the work.

## 2 Related Work

This section reviews some of the recent advances in employing IoT, Fog computing, and blockchains in energy systems.

Home energy management for the optimal day-ahead appliance scheduling was proposed by Paterakis et al. [12]. Their work aimed to minimize the cost for household, meeting all the energy demands under dynamic pricing.

To maximize profit, the utilities have to reduce generation cost. A work based on introducing battery storage into smart grids was proposed by Howlader et al. [13]. Their work considered the electric vehicles and heat pumps in smart homes as sources that could be controlled for maintaining the demand. Their results show that introducing battery storage systems into the grid can improve the generation cost.

Technologies like IoT and other communications have made real-time monitoring of grid a reality. A novel method of achieving energy efficiency of SG was proposed by Chiu et al. [14]. They consider the energy buying-back and dynamic pricing can help users and utilities achieve maximum benefits. Their results showed that their model reduced peak hour loading and energy distribution.

Saleem et al. provided a detailed survey of how IoT aided SG. Their study detailed the possible advantages and the services that IoT could provide in the area of SG like connectivity, automation, and tracking [15].

Moghaddam and Leongarcia, proposed architecture for the Internet of Energy. They explained the use of Fog computing in the energy system. They designed the power system in three different layers, the lowest layer was the sensors and electronic devices which collect data, and the middle layer was the Fog layer, which directly communicates with the end-users. The last layer Fog and Cloud interacted for studying the power market. Their method did not take into consideration the individual household demands [16].

Nguyen et al. provided a mathematical model for bidding problem between the virtual power plant and the exchange of power between customers [17]. Because renewables are not always available, and there is a need for optimal usage of power, they modeled the system for day-ahead negotiation and real-time load management.

Their work showed that demand response exchange would help in effective power management.

Bellavista et al. have done an extensive study on how Fog computing can support many IoT applications. In their work, they have presented a new architecture which integrates Fog and IoT. They claim that Fog will improve the performance of IoT systems and also their architecture can support a variety of applications. The problem with IoT based applications is that each application has a different set of requirements, and they have to be met for business efficiency [18].

Yang et al. provided a detailed survey on the integration of blockchains in edge computing. Their study explains the need for technology [19]. They claim the use of blockchains into energy systems should be validated for scalability and performance before large scale deployment.

Kotsiuba et al. reviewed the challenges and effectiveness of using blockchains in energy systems. They proposed a decentralized transaction framework for the energy sector that can support advanced metering and distributed generation [20].

Aste et al. detailed the fundamental concepts behind blockchain and their advances. Their work explains how blockchain generate the necessary level of trust between unknown partners to trade [21].

Christidis et al. examined whether blockchain can be a good fit for the Internet of Things (IoT) sector. They reviewed how the distributed ledger this mechanism worked and also looked into how smart contracts allow the automation of the block creation [22]. They suggested some of the steps that should be considered for deploying blockchains in IoT.

Kang et al. proposed Peer-to-Peer (P2P) electricity trading model among Plug-in Hybrid Electric Vehicles (PHEVs) in smart grids using blockchain. Their analysis shows that their methodology improved transaction security and privacy protection [23].

Kim et al. analyzed the use of blockchain in microgrid. They studied the advantages and disadvantages of using public and private blockchain [24].

Li et al. used private blockchain technology to model a secure energy trading system called energy blockchain [25]. This energy blockchain can be used for P2P energy trading. They proposed a credit-based payment scheme for fast energy trading.

From the literature, it is understood that the combination of IoT-Edge and blockchain is an effective combination for a smart grid system. The following sections will detail a secure IoT-Edge framework for microgrids.

### 3 IoT-Edge Based Framework for Microgrid System

Smart grids are proposed for efficient power delivery to the consumers. Microgrids are self-sufficient grids that can be a part of the macrogrids or can act as an island grid. Microgrids are focused on distributed generation and integration of greener energy into the grid system. To effectively meet the supply and the demand, many architectures are proposed. Smart grid as a whole has smart homes, AMI, asset

management systems, power quality control, two-way communication between the consumers and the utilities, Demand-Side Management, Phasor Measurement Units, and so on. Many components are integrated and form a single system called smart grid. With the advent of IoT gathering data has become simpler and we have tremendous flow of data. Obtaining insights from these data and processing it will provide effective control and maintenance mechanisms. The following section explains the major components in microgrids and IoT-Edge based framework will aid such critical infrastructure.

### ***3.1 Smart Homes***

In the microgrid system, all homes that are connected to the grid are expected to provide two-way communication. All appliances in the homes are connected to a centralized computer system and can be controlled via direct load control. Smart meters are installed in every home which can communicate with the Utilities and also with the computer control of home. The smart meters are installed by the Utilities in every home. They have communication capabilities like 6LoWPAN or other wireless communication. The smart homes are the consumers, and they can play a very important role in demand reduction and energy trading. All homes are now becoming prosumers (producers + consumers); they are ready to sell the excess energy to the grid.

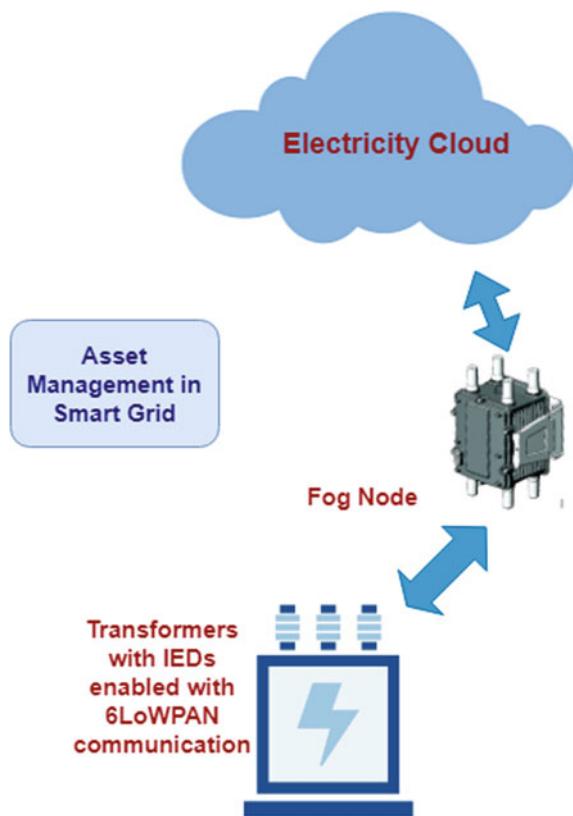
### ***3.2 Demand-Side Management***

The current electricity grid systems face the problem of supply–demand imbalance. This imbalance causes frequency fluctuation in the grid. The problem of supply and demand imbalance can be rectified if near real-time predictions of demand in an area are known. Many works have been carried out in the field of Demand-side management. This is also an important aspect of the microgrid system that has to be addressed. Machine learning has been widely used by researchers and Utilities to provide a solution for demand-side management.

### ***3.3 Transformer Health Monitoring***

Transformers are an important asset in power distribution system. The supply from substations is distributed to homes by stepping down the voltage using a transformer. Monitoring the assets of the Utilities is an important aspect that has to be considered while designing a microgrid system. There are occurrences of transformers getting burnt off or overloaded than the rated load. These overloading can reduce the health

**Fig. 2** Asset management in smart grid by IoT-Edge device



of transformers. Many solutions are provided with fuzzy logic and other algorithms for monitoring and control the function of transformer. IoT based systems can be introduced for asset management into the grid system as shown in Fig. 2.

### 3.4 Electric Vehicles and Grid

The electricity that is generated has to be used immediately. Storage of energy on a large scale has not yet been introduced into the grid. Proposals for large battery storage from Tesla are important as storage of electricity when it is cheap, and surplus is to be considered. Reduction of carbon emission is taken as a global concern due to global warming. Electric vehicles that run on batteries are introduced into the grid. They can act as an energy storage device when connected to the grid. But they also pose a problem when connected to the grid in large numbers during peak hours. Researchers work in the area of vehicle to grid to optimize the way in which electric

vehicles can be charged and turn out as an energy storage device in the electric grid network.

### **3.5 IoT-Cloud-Based System**

IoT has made possible the integration of ICT into the electrical grid system. Thus the above said smart homes, transformer health monitoring, DSM can be connected to the Internet and can be monitored by SCADA systems. Cloud-based architecture has assisted the overall integration of the power distribution system for effective monitoring and control. All metering data has to be sent to the cloud for analytics. This utilizes huge bandwidth and is found to be less efficient as the traffic from application like smart grid is tremendous. The power distribution system of a state or a region covers a large geographical distribution. This has to be addressed.

### **3.6 IoT-Edge-Cloud in Power Distribution System**

Power distribution system being largely geographically distributed and that critical infrastructures need to be analyzed and monitored with minimal delay. Cisco came up with the concept of Fog computing which brings analytics and processes that can occur in the cloud nearer to the edge devices. This work proposes a framework which exploits the use of Fog computing. After the introduction of Fog computing many Cloud service providers are providing Edge services (e.g., Azure IoT-Edge) for applications support of smart cities, water distribution systems, health care, etc. Fog computing has many advantages which makes them suitable for applications like smart grid. As Fog computing nodes are geographically distributed they can be exploited to be used in smart grid for effective monitoring and control. Fog nodes can communicate with the smart meters and other metering infrastructure by means of 6LoWPAN or RF communication and are connected to the Cloud by means of WLAN or 4G/5G. Fog computing will reduce the bandwidth usage to the cloud and reduce latency for data analysis. These Fog devices all follow IEC61850 standard.

When critical infrastructure like the smart grid is connected to the Internet, security is a major concern. This has been addressed by IEC62351 part 1 to 8 which details the need for security in data and communications with respect to power systems. The areas in the power system where security is a major concern are Energy Management System (EMS), Distribution Automation (DA), Distributed energy resources (DER), Advanced Metering Infrastructure (AMI), Demand Response (DR), Smart Home, Storage and Electric Vehicles (EV). The next section details the proposed secure framework for edge computing specifically designed for smart grids.

## 4 Secure Edge Framework with Blockchain for Smart Microgrids

Edge computing framework is a distributed computing method which supports applications that cover a large geographical area. Smart grids are systems that cover towns, cities, states, and spreads across a country. The number of systems involved and to be monitored are tremendous. IoT-Edge framework helps in supporting the infrastructure as shown in the previous sections. But securing such a system from attacks and illegal activities is an important aspect for the Utilities. The next level of cyber-attacks will take place on infrastructures like energy systems and water distribution systems. The following section details the edge computation layer that provides security as a service and is applicable for smart microgrid systems.

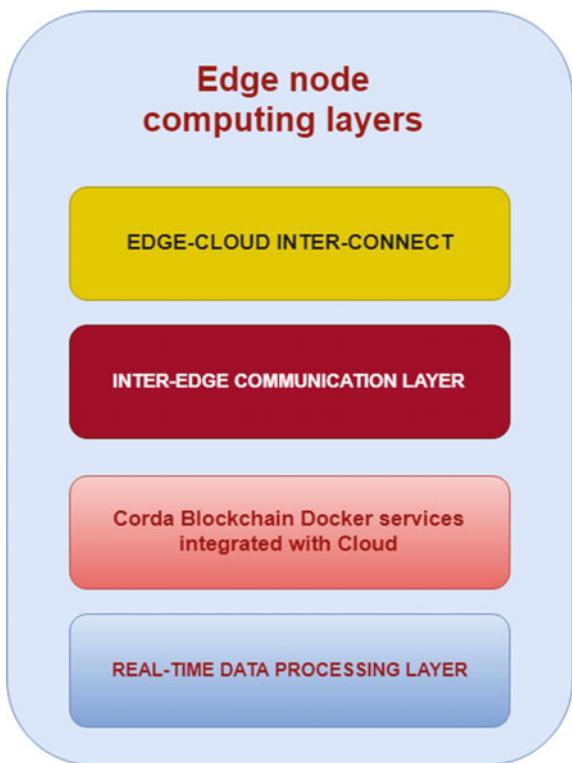
### 4.1 *Secure Edge Framework*

Edge computing deals with the distributed cloud resource and computation at the edge of the devices. On the other hand, blockchain provides secured distributed ledger. Both decentralized management and cryptographic security support each other. Combining both edge computing and blockchain reveals a reliable system in terms of network, security, and storage. The transactions are traceable since peer nodes are validated before joining into the blockchain network. Figure 3 shows the layered structure for secure smart microgrid applications. The first layer is the real-time processing layers. Once the data enters the Fog node it can be analyzed for criticality. The second layer provides security as a service layer. This layer is responsible for the security in data and control actions that are involved in the grid system. Blockchain-based security is most widely talked and has proved to be secure. Blockchain dockers can run in this layer. The next layer is the inter Edge communication layer. The nearby Fog/Edge nodes form a network and are in turn connected to the Cloud. The next layer is the Edge/Fog node connection with the Cloud.

### 4.2 *Blockchains in Edge/Fog Routers*

Blockchains are a shared distributed ledger containing digital transactions that are time-stamped and cryptographically linked. The hash of the previous block is linked with the next block so that any modification of the block can be identified. Immutability is another aspect where one-way hashing is used to make the system tamper-proof. Edge computing provides an intermediate layer in between end device and cloud infrastructure such that minimum proximity is maintained. Due to the edge server, the traveling time of data from the end device to the cloud is reduced. Such that data latency and traffic flow in the network is decreased. Edge computing provides

**Fig. 3** Edge computational layer



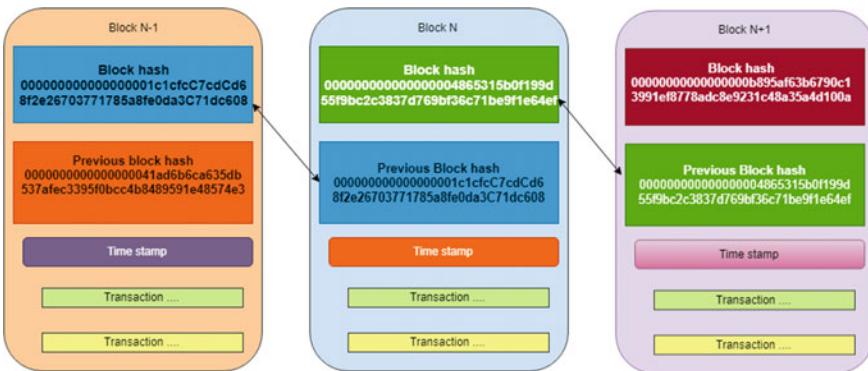
autonomous management where the data are stored locally in the end-user. Another reason for the enhanced security is due to the cryptographic hash function which is a one-way hash function and an output of 256 bits. This hash output of 256 bits practically difficult to obtain the same hash after modification.

Blockchains are classified broadly into public (permission less) and private (permissioned). Bitcoins the first introduced cryptocurrency using blockchains uses public blockchain architecture. The block can be added to a chain only after acceptance from network members. This is called *reaching consensus*. This procedure is different for each blockchain framework, and they determine the performance metrics of a blockchain mechanism. Only after the block is accepted and hashed it can become a part of the chain, a process called *finality*. Table 1 shows the differences between a private ledger and a public ledger system. Some researchers are still in debate as to how private permissioned ledger system can be called a blockchain.

The different consensus mechanisms used in different blockchain methodology are Proof of Work (PoW), Proof of Stake (PoS), Practical Byzantine Fault Tolerance (PBFT), Delegated Proof of Stake (DPoS), Federated Byzantine Agreement (FBA), Proof of Authority (PoAu), Proof of Elapsed Time (PoET), Proof of Activity (PoAc), Proof of Burn (PoB) and Proof of Capacity (PoC) [27] (Fig. 4).

**Table 1** Difference between public and private ledger models

Metrics	Public ledger	Private ledger
Accessibility	Read and write by public	Read and write possible only by invited members
Network Actors	Not known to each other	Knows each other
Native token	Yes	Not needed
Consensus mechanisms	Proof of work Proof of stake Proof of space and so on	Legal contract Proof of authority
Speed	Low	High
Examples	Bitcoin Ethereum Monero Bash etc	R3 (banking) EWF (energy) Corda (insurance) HyperLedger

**Fig. 4** Blockchain: This shows how each block is linked to the next block using the previous block hash

Blockchain mechanisms can be made to run in edge node by means of docker mechanism. Cloud services like Azure provide mechanisms for implementing blockchain dockers in resource constraint devices. Blockchain implementation is highly resource consuming, and consensus mechanisms like PoW are not practically applicable for applications like smart grids [28, 29]. The Edge node is connected to the cloud by means of backhaul network like 4G/5G or WLAN. Introduction of 5G worldwide will aid for the implementation of such security architectures.

### 4.3 Use Cases of Blockchains in Energy

Blockchains have a number of use cases in the energy sector and broadly classified into operations and business side as given below:

Microgrid management, data communication, Distribution Automation, electric vehicles, security, energy trading, billing, sales, and market. We look into Energy trading as a use case in this work.

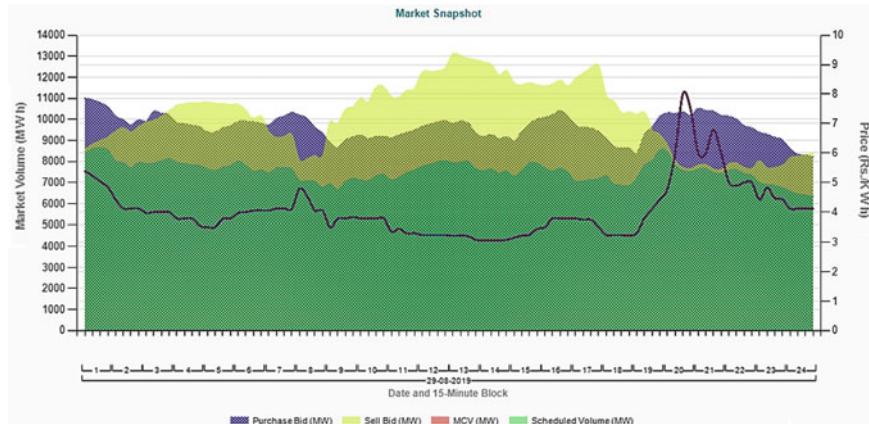
In this work, authors propose the use two different blockchain mechanism one for operational applications and another for business applications within the energy domain:

- (1) Energy Web Chain (EWC) blockchain mechanisms with Dapps for the use in smart microgrid systems for applications like Advanced Metering Infrastructure (TEO/Engie), Electrical vehicle interconnection to the grid (Open Charging Network), solar and wind farm integration(Origin PTT). EWC is an open-source blockchain that uses the Proof of Authority as the consensus mechanism. They use a native token for transaction payments. They use a mechanism called EW tokens for payment of energy trading. On average, the time taken for creating a block is five seconds.
- (2) Corda R3 Distributed Ledger Technology (DLT) can be used for Energy trading. Where we need not distribute the ledger to all the users in the network. The mechanism is few nodes can connect to a notary and can get approved of a transaction. They use CorDapps for distributed applications. They are written in Java or Kotlin. The features of DLT are that they are cryptographically secured, immutable, they run on smart contracts, they have shared ledgers, and also have the distributed consensus mechanism [26].

Corda works on shared ledger. It is a permissioned network, and no broadcast of data is done. The data is shared on a need to know basis. The identity of the sender and the receiver is known. There is no central ledger, but each node shares shared facts. These properties make us believe that Corda framework will provide a good framework for energy trading. A CorDapp called Energy Blok Exchange (EBX) can be used for energy trading, especially when energy trade between different Utilities has to be done. The Corda is now getting integrated with Azure cloud blockchain services.

### 4.4 Electricity Trading

Energy trading and marketing is a process of buying the electricity and also selling it to the place where the demand for energy is needed. In energy trading, apart from electricity other products can also be traded such as wind power, natural gas, and crude oil. The supply and the demand of energy decide the price of the wholesale. For example, in northern America during winter season or summer season the usage



**Fig. 5** Energy Exchange real-time information of India from iexindia

of heater and air conditioner is high which leads to increase in the price of energy. The price may fluctuate on a daily basis also. For example, in day time the usage of electricity is high compared to night time [30–34]. Due to this the price of electricity is high during peak hours and low during mid-day and night time as shown in Fig. 5.

The identities used by the trading agent for trading have to be preserved by using some contract among the trading agents during the bidding, negotiation, and transaction [35]. The majority of the financial infrastructure in energy trading is centralized, trusted third party. The maintenance of the account provides security, bill payment process is maintained by the centralized finance infrastructure. The centralized trading has disadvantage such as (1) single point failure and (2) lack of privacy and being anonymous.

**Single point failure:** Since the financial infrastructure is centralized, the failure of the system leads to an obstacle in the process of authentication and payment transaction.

**Lack of privacy and being anonymous:** the centralized system can reveal the activities of the agent or users by tracking the energy usage pattern of the corresponding agent or users.

To resolve these two demerits, a finance infrastructure should be strong in both availability/reliability and security. Blockchain implementation using Corda DLT can be effective in overcoming such drawbacks.

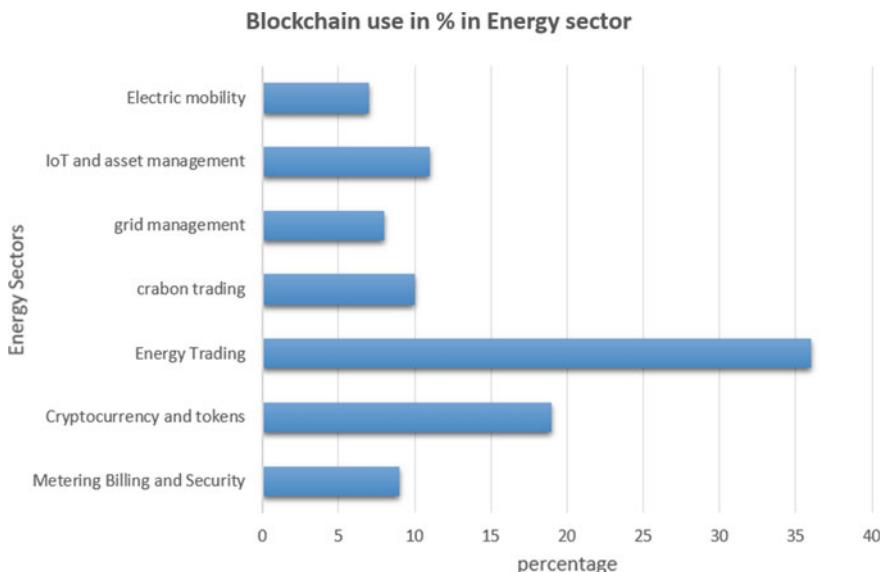
#### 4.5 Implementation Using Corda

The implementation was done using Corda R3 open-source on Intel i5 machine with 12 GB RAM.

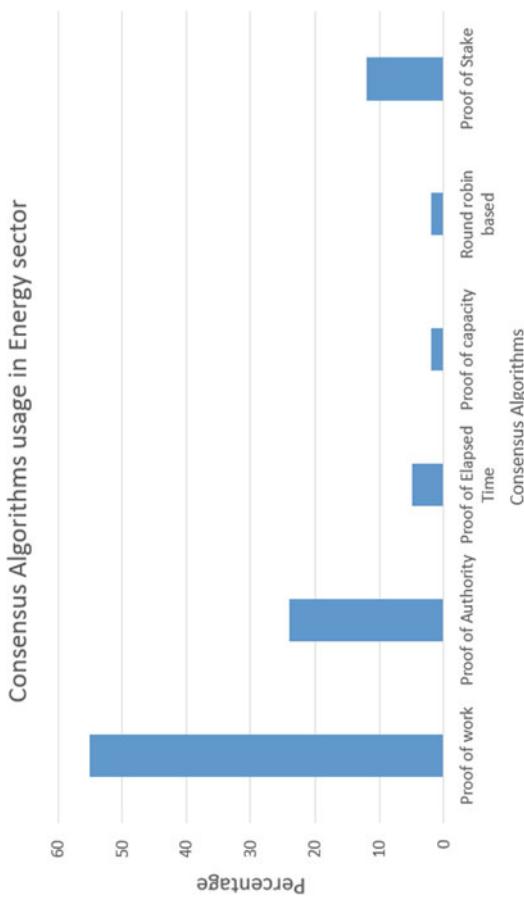
Consider one party needs to sell his renewable energy to the Utility they can transact using the Corda network.

The key concepts of Corda are State object, transactions, consensus, and flow. The state represents the immutable agreements between two parties at a specific point of time. Corda uses an unspent transaction output (UTXO) model for transactions. Transactions took the input state and created the output state. This output state replaces the present input state in the Ledger.

A Corda network comprises of a doorman, two or more Corda nodes, a network map service, one or more notary service, zero or more oracles. The *doorman* enforces the rules that are to be followed for a node to be accepted in a network. Once a node is accepted, it is certified with root authority signed TLS certificate. *Nodes* are Java Virtual Machine run time running Corda with network identity. Nodes have two interfaces one is network layer which connects with the peers the other is the RPC which connects with the cloud. Network map service contains the IP addresses of the nodes. A *notary* verifies the uniqueness of a transaction. Oracles signs a transaction if they state the fact and that fact is true. Corda network is a fully connected graph, and peers communicate through AMQP over TLS. Figure 6 shows the sectors in the microgrids where blockchain technology will be effective. Figure 7 shows the most widely used consensus mechanisms used based on previous works and in industry. Proof of work though most widely used consumes more energy and is computationally complex. A block to get generated by proof of work takes more time. This work proposes the use of blockchain technology that uses Proof of Authority as the consensus mechanism. The model of energy trading between supplier and Utility is detailed.



**Fig. 6** Energy sectors where blockchains are widely preferred



**Fig. 7** Blockchain consensus algorithms most widely used

Energy trading of renewable energy depends on *{energy pricing, Power Availability, Demand}*.

Consider a renewable source  $RE_x$  is ready to trade energy with another Utility  $U$  who wants to buy electricity. The doorman accepts them as nodes, and they are digitally certified.

Let  $RE_\alpha = \{RE_1, RE_2, \dots, RE_n\}$  be the set of energy trade contracts over which a Utility node needs to come to a consensus.

*Renewable energy node transaction for energy trading:*

The RE sells electricity when the production is more. This makes the Utilities buy electricity because the price of the electricity is cheaper and the energy source is non-fossil fuel. The two parties need to come to a contract agreement. The contract agreement between RE and Utility is given by

$$E_t^{RE \rightarrow U}(RE_i^j) = \left\{ \varsigma_t^{RE \rightarrow U}(RE_i^j) \right\}_{j=1}^m \quad (1)$$

where,

$$\varsigma_t^{RE \rightarrow U}(RE_i^j) = \min(RE_i^j) + \theta_t(\max(RE_i^j) - \min(RE_i^j)) \quad (2)$$

$\theta_t = \left( \frac{\omega_{end} - \omega}{\omega_{total}} \right)^{\alpha_i^j}$ ,  $\omega_{end}$ —ending time of negotiation,  $\omega_{total}$ —is the total time taken for concluding a negotiation,  $\omega$ —current time.  $\alpha_i^j$  is the speed of achieving the transaction.

The score of the RE's agreement is found by

$$\Theta E_t^{RC \rightarrow U} = \sum_{j=1}^m \Theta(\varsigma_t^{RE \rightarrow U}(RE_i^j)) * \theta_i^j \quad (3)$$

where,

$$\Theta(\varsigma_t^{RE \rightarrow U}(RE_i^j)) = \left( \frac{\max(RE_i^j) - \varsigma_t^{RE \rightarrow U}(RE_i^j)}{\max(RE_i^j) - \min(RE_i^j)} \right) \quad (4)$$

*Utility node payment transaction for energy purchase:*

Once the energy is purchased from the renewable energy node, the payment has to be settled. The transaction agreement between the Utility and the renewable energy node is given by

$$\Upsilon_t^{U \rightarrow \text{RE}}(\psi_i^t) = \{\varsigma_t^{U \rightarrow \text{RE}}(\psi_i^j)\}_{j=1}^m \quad (5)$$

where  $\psi_i^t$  is the cost Utilities has to pay for each energy purchase.

$$\varsigma_t^{U \rightarrow \text{RE}}(\psi_i^j) = \min(\psi_i^j) + \varepsilon_t (\max(\psi_i^j) - \min(\psi_i^j)) \quad (6)$$

$\theta_t = \left( \frac{\omega_{\text{end}} - \omega}{\omega_{\text{total}}} \right)^{\alpha_i^j}$ ,  $\omega_{\text{end}}$ —ending time of negotiation,  $\omega_{\text{total}}$ —total time is taken for concluding a negotiation,  $\omega$ —current time.  $\alpha_i^j$  is the speed of achieving the transaction.

The score of the Utility's agreement is found by

$$\Theta \Upsilon_t^{U \rightarrow \text{RE}} = \sum_{j=1}^m \Theta(\varsigma_t^{U \rightarrow \text{RE}}(\psi_i^j)) * \theta_i^j \quad (7)$$

where,

$$\Theta(\varsigma_t^{U \rightarrow \text{RE}}(\psi_i^j)) = \left( \frac{\max(\psi_i^j) - \varsigma_t^{U \rightarrow \text{RE}}(\psi_i^j)}{\max(\psi_i^j) - \min(\psi_i^j)} \right) \quad (8)$$

Once the scores are calculated for the proposals  $E_t^{\text{RE} \rightarrow U}$  ( $\text{RE}_i$ ) and  $\Upsilon_t^{U \rightarrow \text{RE}}(\psi_i)$ , the consensus is concluded when the two parties come to an agreement for the best deal. Once the transaction between the two parties is done, the *verification consensus* is done where all the participating peers and checks for the mandates in the contract. The *notary* service provides the *uniqueness consensus* to eliminate the double-spending problem. The notary confirms that an input state is used only once for a transaction. The *finality* is reached once the *notary* signs the transaction.

## 5 Conclusion

Microgrids form an integral part of the smart grids. For effective monitoring and control of the grid IoT integrated with Edge and Cloud is an effective solution. The distributed nature of the Edge computing nodes makes it suitable for monitoring geographically distributed power distribution systems. To overcome the security issues in smart grids a secure framework incorporating blockchains was proposed. The distributed Ledger Technology can be exploited for different applications in smart grid-like billing, asset management, grid automation, advanced metering infrastructure, electric vehicle integration, renewable energy integration, energy trading, market, and Utilities. The newly introduced Energy Web chain uses a public

and private blockchain mechanism and is specifically tailored to fit the different applications in the smart grid. A lightweight blockchain like Corda can be used for financial transactions in the smart grid regime. Though some of the researchers claim that blockchain is not an effective solution for application like smart grid, incorporation of blockchain into the electric grid system is still in its nascent stage.

## References

1. G. Bedi, G. Venayagamoorthy, R. Singh, R.R. Brooks, Review of the Internet of Things (IoT) in electric power and energy systems. *IEEE Internet Things J.* **5**(2), 847–870 (2018)
2. J. He, S. Member, J. Wei, K. Chen, Z. Tang, Multitier fog computing with large-scale IoT data analytics for smart cities **5**(2), 677–686 (2018)
3. A.H. Ngu, M. Gutierrez, V. Metsis, Q.Z. Sheng, IoT middleware : a survey on issues and enabling technologies **4**(1), 1–20 (2017)
4. A. Zanella, N. Bui, A. Castellani, L. Vangelista, M. Zorzi, Internet of Things for smart cities. *IEEE Internet Things J.* **1**(1), 22–32 (2014)
5. F. Montori, L. Bedogni, L. Bononi, A collaborative Internet of Things architecture for smart cities and environmental monitoring. *IEEE Internet Things J.* **5**(2), 592–605 (2018)
6. J.-C. Kim, S.-M. Cho, H.-S. Shin, Advanced power distribution system configuration for smart grid. *IEEE Trans. Smart Grid* **4**(1), 353–358 (2013)
7. N. Saputro, K. Akkaya, Investigation of smart meter data reporting strategies for optimized performance in smart grid AMI networks. *IEEE Internet Things J.* **4**(4), 894–904 (2017)
8. L. Ardito, G. Procaccianti, A survey on smart grid technologies in Europe, in *The Second International Conference on Smart Grids, Green Communications and IT Energy-aware Technologies A* (2012), pp. 22–28
9. R.M. Cisco, M.Y. Upc, M. Nemirovsky, Fog computing. *Cloud Assist. Serv. Eur. Conf. Bled 2012*, 1–15 (2012)
10. W. Shi, J. Cao, Q. Zhang, Y. Li, L. Xu, Edge computing: vision and challenges. *IEEE Internet Things J.* **3**(5), 637–646 (2016)
11. N. M. Gonzalez et al., Fog computing: data analytics and cloud distributed processing on the network edges, in *35th International Conference of the Chilean Computer Science Society (SCCC)* (2016), pp. 1–9
12. Z. Zhao, L. Wu, G. Song, Convergence of volatile power markets with price-based demand response. *IEEE Trans. Power Syst.* **29**(5), 2107–2118 (2014)
13. N.G. Paterakis, S. Member, O. Erdinç et al., Optimal household appliances scheduling under day-ahead pricing and load-shaping demand response strategies. *IEEE Trans. Ind. Inf.* **11**, 1509–1519 (2015). <https://doi.org/10.1109/TII.2015.2438534>
14. H.O.R. Howlader, H. Matayoshi, T. Senju, Distributed generation integrated with thermal unit commitment considering demand response for energy storage optimization of smart grid. *Renew Energy* **99**, 107–117 (2016). <https://doi.org/10.1016/j.renene.2016.06.050>
15. T.C. Chiu, Y.Y. Shih, A.C. Pang, C.W. Pai, Optimized day-ahead pricing with renewable energy demand-side management for smart grids. *IEEE Internet Things J.* **4**, 374–383 (2017). <https://doi.org/10.1109/JIOT.2016.2556006>
16. Y. Saleem, S. Member, N. Crespi, et al. Internet of Things-aided smart grid : technologies, architectures, applications, prototypes, and future research directions. *arXiv* **170408977**, 1–30 (2017)
17. Y. Moghaddam, A. Leon-garcia, A fog-based internet of energy architecture for transactive energy management systems. *IEEE Internet Things J.* **5**, 1055–1069 (2018). <https://doi.org/10.1109/JIOT.2018.2805899>

18. P. Bellavista, J. Berrocal, A. Corradi et al., A survey on fog computing for the Internet of Things. *Pervasive Mob. Comput.* **52**, 71–99 (2019). <https://doi.org/10.1016/j.pmcj.2018.12.007>
19. F. Ruizhe Yang, Y. Richard, P. Si, Z. Yang, Y. Zhang, Integrated blockchain and edge computing systems: a survey, some research issues, and challenges. *IEEE Commun Surv Tutorials* **21**(2), 1508–1532 (2019)
20. I. Kotsiuba, et al., Using blockchain for smart electrical grids, in *IEEE International Conference on Big Data (Big Data)* (2018), pp. 1–9
21. T. Aste, P. Tasca, T. Di Matteo, Blockchain technologies: the foreseeable impact on society and industry. *Comput. IEEE Comput. Soc.* **50**, 18–28 (2017)
22. K. Christidis, M. Devetsikiotis, Blockchains and smart contracts for the Internet of Things. *IEEE Access* **4**, 2292–2303 (2016). <https://doi.org/10.1109/ACCESS.2016.2566339>
23. J. Kang, Y. Rong, X. Huang, S. Maharjan et al., Enabling localized peer-to-peer electricity trading among plug-in hybrid electric vehicles using consortium blockchains. *IEEE Trans. Industr. Inf.* (2017). <https://doi.org/10.1109/TII.2017.2709784>
24. G. Kim, J. Park, J. Ryoo, A study on utilization of blockchain for electricity trading in microgrid. *IEEE Int. Conf. Big Data Smart Comput.* (2018). <https://doi.org/10.1109/BigComp.2018.00141>
25. Z. Li, J. Kang, R. Yu et al., Consortium blockchain for secure energy trading in industrial Internet of Things. *IEEE Trans. Ind. Inf.* <https://doi.org/10.1109/TII.2017.2786307>
26. R.G. Brown, The corda platform: an introduction, white paper (2018)
27. Andoni et al., Blockchain technology in the energy sector: a systematic review of challenges and opportunities. *Renew. Sustain. Energy Rev.* Elsevier (2018)
28. I. Alqassem, D. Svetinovic, Towards reference architecture for cryptocurrencies: Bitcoin architectural analysis, in *IEEE iThings, GreenCom, and CPSCom* (2014), pp. 436–443
29. Z. Zheng, S.H. Xie N. Dai, H. Wang, Blockchain challenges and opportunities: a survey (2017). <https://www.henrylab.net/wp-content/uploads/2017/10/blockchain.pdf>. Accessed 5 Jun 2018
30. I.O.T.A, IOTA (2017) URL <https://www.iotatoken.com>
31. G. Briscoe, T. Aste, Blockchains: distributed consensus protocols, transparency and business models (Univ. College London; submitted to *J. Digital Economy*, 2017)
32. Introduction to Smart Contracts—Solidity 0.2.0 Documentation. Available: <http://solidity.readthedocs.org/en/latest/introduction-to-smart-contracts.html> Accessed on 15 Mar. 2016
33. W. Tushar et al., Three-party energy management with distributed energy resources in smart grid. *IEEE Trans. Ind. Electron.* **62**(4), 2487–2498 (2015)
34. Z. Zheng et al., *Blockchain Challenges and Opportunities: A Survey* (Workshop Paper, IOP Publishing, Geneva, Switzerland, 2016)
35. J. Matamoros, et al, Microgrids energy trading in islanding mode, in *IEEE SmartGridComm* (2012), pp. 49–54

# Sentiment Analysis on the Performance of Engineering Students in Continuous Assessment Evaluation System: A Non-parametric Approach Using Kruskal-Wallis Method



A. Vijay Bharath, A. Shanthini, and A. Subbarayan

**Abstract** Sentiment Analysis is implicitly related to Continuous Assessment system in educational evaluation. Continuous Assessment system is a systematic and objective process of determining the extent to which changes have taken place in the student's performance in the various areas of educational objectives viz., Cognitive, Affective and Psychomotor. In this paper, an attempt is made to study in detail the different components of Continuous Assessment system in relation to the objectives noted above in an empirical manner. In the empirical analysis, the internal marks of a sample of students in a higher education institution is used. A nonparametric approach namely Kruskal–Wallis one way ANOVA is used for the analysis of the sampled data. The inferences drawn from the study clearly indicates that the planning mechanism adapted by educational authorities in evolving Continuous Assessment system for engineering education is fulfilled in a successful manner.

**Keywords** Sentiment analysis · Nonparametric · One way ANOVA · Kruskal–Wallis · Continuous Assessment · Ranking

## 1 Introduction

Sentiment Analysis (SA) primarily involves identification and classification of opinions. The sentiments expressed by the users may be of the type namely positive, negative and neutral or emotions of the type viz., happy, sad, angry or disgusted towards a particular subject. The importance of SA in different fields is felt by researchers. SA plays an important role in the evaluation of progress of students stage by stage in respect of their learning process. Institutes of higher learning and universities adopts different evaluation system.

---

A. Vijay Bharath (✉) · A. Shanthini

Department of Data Science and Business Systems, SRM Institute of Science and Technology, Tamilnadu Chennai, India

e-mail: [vijaybha@srmist.edu.in](mailto:vijaybha@srmist.edu.in)

A. Subbarayan

DIRECTORATE OF RESEARCH, SRM INSTITUTE OF SCIENCE AND TECHNOLOGY, TAMILNADU CHENNAI, INDIA

## ***1.1 Sentiment Analysis and Its Relation to Continuous Assessment***

Adedibu [1] defined Continuous Assessment as a systematic and objective process of determining the extent to which changes have taken place in the students' performance in the various areas of educational objectives (cognitive, affective and psychomotor). Assessment within the cognitive domain is related with the method of information and understanding. The emotional space applies to characteristics such as states of mind, thought processes, intrigued and other identity trials. Assessment in psychomotor space includes in surveying the learners capacity to utilize his or her hands (eg., Handwriting, Constructions and Projects).

## ***1.2 Continuous Assessment and Its Features***

Continuous Assessment is a planned process of identifying, gathering, and also interpreting information about the performance of learners. It includes four steps, creating and collecting prove of accomplishment, assessing this prove against results, recording the discoveries of the data to understand and thereby help the learners' development and progress the method of learning and teaching. Continuous Assessment is an integral part of teaching and learning. It is transparent and learners centered in approach. Learners are assessed holistically.

The term CA is used to describe the assessments that are completed during the course module. The method is also referred to as curriculum integrated assessment or embedded assessment. The reason for doing Continuous Assessment is to secure/enable a continuous and independent work rate and learning for students in the course.

### **Features of Continuous Assessment**

- (a) They are regular and frequent in nature.
- (b) The Continuous Assessment procedures is a compelling instrument to decide and create competencies.
- (c) The strategy is comprehensive, aggregate, demonstrative, developmental, direction oriented and orderly in nature.

### **Purposes of CA**

- (a) Enhance student learning.
- (b) Improve the faculties teaching skills.
- (c) Improves the education and organization evaluation framework

### ***1.3 Basic Components of CA System***

The students Continuous Assessment mainly consists of the following components namely,

- (i) Assignments
- (ii) Quizzes
- (iii) Student seminars
- (iv) Project reports
- (v) Periodical Tests
- (vi) Video tape
- (vii) Audio tape
- (viii) Photographs
- (ix) Artifacts and products of students science

The above aspects constitute Continuous Assessment (CA) of evaluation and its relation to SA. In the present study, the focus is on the integration of SA and CA evaluation system.

### ***1.4 Motivation of the Present Study***

The educational planners formulate courses for different engineering disciplines catering to the needs of the industry and society at large. In this context they emphasize more or two components of evaluation namely Continuous Assessment and external assessment. The Continuous Assessment focuses on learning capabilities of the students stage by stage. The author is of the firm opinion that Continuous Assessment system is equally important with external assessment. It may be noted there are only limited studies relating to assessment of Continuous Assessment. The author of this paper would like to investigate in detail the Continuous Assessment component related to Sentiment Analysis.

### ***1.5 Objectives of the Study***

- (i) To formulate nonparametric Analysis method for data relating to categories of subjects offered under different engineering disciplines.
- (ii) To apply Kruskal-Wallis one way Analysis of Variance structure for the data used.
- (iii) To compare the difference in Continuous Assessment marks under different categories of subjects.
- (iv) To integrate the Sentiment Analysis aspects with the policies of educational planners for effective evaluation of Continuous Assessment marks.

In Sect. 2, we have given an updated review of literature relating to Sentiment Analysis and its associated aspects. Section 3 deals with the methodology adapted for the study. Details relating to nonparametric methods are discussed. In Sect. 4, we have studied in detail the data structure considered for the study. Detailed computational aspects are included in this section. The results and conclusions are presented in Sect. 5 incorporating Sentiment Analysis aspects.

## 2 Review of Literature Incorporating Continuous Assessment and Sentiment Analysis

Bordovsky et al. discussed in length the educational process quality management with special emphasis on Continuous Assessment System [2]. Belova developed prognostic model of system of quality management of the educational activities in University [3]. Hernandez examined the aspects of Continuous Assessment in Higher education [4]. In this study, the author compared Continuous Assessment with External Assessment and pointed out the challenges of implementation in respect of both the assessment systems. Guzman et al have made an empirical study in respect of Sentiment Analysis of Commit Comments in GitHub [5]. The authors have mentioned that emotions have an high impact in productivity, Task Quality, Creativity and Group rapport. These aspects are implicitly related to Continuous Assessment studies.

Dhagat and Mukherjee focused on the need of changes that can make the method of Continuous assessment more fruitful and accessible to the accessors [6]. They have also suggested ways and means for the effective adoption of Continuous Assessment in higher Education. Getient Seifu examined the status of implementation of CA in mettu University and pointed out the challenges of the adaption of CA [7]. The author also given recommendations for improving the implementation of CA in Universities. Onihunwa et al. investigated Continuous Assessment in determining the academic performance of computer science student in a college [8]. In this investigation they found that students final grades is a function of scored obtained in the Continuous Assessment. They suggested further studies can be conducted for the gender-based attitude to Continuous Assessment and its effect on student Academic Performance.

It is important to note that project work is an important component of Continuous Assessment system. Hengrik Gustavsson and Marcus Brohede studied in detail approaches for the study in Continuous Assessment in large Software Engineering Project [9]. The authors have pointed out that the self-assessment tools allow each student to gage individual progress of the students continuously based on week time schedule.

### 3 Methodology for the Study

#### 3.1 Nonparametric Methods

Parametric statistical methods are based on stringent assumptions about the population from which the sample has been drawn. Particularly the assumptions like form of the probability distribution, accuracy of observations etc., are more common. Also, the parametric methods are applicable primarily to the data which are measured in interval or ratio scale. In practice, however, stringent assumptions are seldom fully valid. Moreover, the measurements are often made on nominal or ordinal scale.

If the assumptions do not hold good or the data do not meet the requirement of parametric statistical methods, nonparametric methods come to the rescue of the researcher. Nonparametric methods entail very mild assumptions like continuity and symmetry of the distribution. Also most of the nonparametric methods are applicable for ordered statistics.

Nonparametric test may be quite powerful even if the sample size is small. It may also be noted that nonparametric test are inherently robust against certain violation of assumptions.

#### 3.2 Kruskal-Wallis Test

It is vital to note that the Mann-Whitney looks for contrasts in median values between two samples only. Kruskal-Wallis test is utilized for differences in median values between more than two samples and the test has the additional merits.

This test is used when the assumptions for Analysis of Variance (ANOVA) are not met. It is also called the one way ANOVA on ranks. The ranks of the data value are used in the test rather than actual points.

The test determines whether the median of two or more groups. The hypothesis of the test are:

Null Hypothesis:— $H_0$ : Population median are equal.

Alternate Hypothesis:— $H_1$ : Population median are not equal.

#### Assumptions

1. The test is more commonly used when we have three or more levels.
2. Observations should be independent, i.e., there is no relationship between in each group or between groups.
3. All groups should have same shape distribution.

#### Description of Kruskal-Wallis Test

Let H statistic is given by [10]:

$$H = \frac{12}{n(n+1)} \sum_{j=1}^C \frac{R_j^2}{n_j} - 3(n+1) \quad (1)$$

$n$  = Sum of sample sizes for all samples.

$c$  = Number of samples.

$R_j$  = Sum of Ranks in the  $j$ th sample.

$n_j$  = Size of the  $j$ th sample.

In case of occurrences of ties in the rank the correction factor is given by [11]:

$$C = 1 - \frac{\sum_{j=1}^g t_j^3 - t_j}{N^3 - N} \quad (2)$$

$g$  = is the number of tied groups.

$t_j$  = is the number of tied data in the  $j$ th group.

For large samples, the calculated value of  $H$  is compared with  $\chi_{0.05}^2$  with  $(k-1)$  degrees of freedom.

The corrected  $H$  is given by:

$$H_c = H/C.$$

Comparing  $H$  value with critical Chi-Square ( $\chi_e^2$ ) value:

- (i) If  $\chi_e^2 < H$  statistic, reject the null hypothesis that the medians are equal.
- (ii) If  $\chi_e^2 > H$  statistic, there is enough evidence to suggest that medians are unequal.

## 4 Data Structure Used for Sentiment Analysis

The data structure used for Sentiment Analysis is discussed in length in the following sections. The data structure incorporates four different components which are distinct in nature.

### 4.1 Data Description

In respect of Sentiment Analysis for nonparametric data structure, we have used the four categories of subjects which are included in the curriculum viz.

- (i) Professional Core (PC)
- (ii) Engineering Science (ES)
- (iii) Basic Science and (BS)
- (iv) Mandatory Course (M)

This professional course plays an important role in the curriculum of the study. The students necessarily have to complete this course to move to the next level for getting the degree. The professional course forms the basis for foundational

knowledge and skills which they acquire during their study period. The engineering experts emphasize that this professional core provides strong foundation.

Engineering science course encompasses different scientific principles and associated mathematics that underlie engineering. This course integrates viz., biological, chemical, mathematical and physical sciences. It also emphasizes on arts, humanities and social sciences to tackle challenges posed by global society for well-being.

Basic science course relates to the basic discovery and inventions in scientific research. The basic science course explores the knowledge in a particular field. In addition to this, the course has the benefits viz., contributions to the culture, spin-offs and stimulation of industry and enormous economic and practical importance.

The mandatory course plays an important role among the students in respect of creative arts, literacy etc. It may be noted that this is a popular course among the engineering students for self-development.

## 4.2 Procedure for Computing H Statistics

We have formed the following basic table for computing H statistics in Table 1 and ranking for the same in Table 2.

It is important to note that the size of the sample varies under different categories of subjects ( $n_1 \neq n_2 \neq n_3 \neq n_4$ ).

**Table 1** Basic data for computation of  $H$  statistic

Samples	Internal marks obtained
Sample I	$S_{11}, S_{12}, \dots, S_{1n_1}$
Sample II	$S_{21}, S_{22}, \dots, S_{2n_2}$
Sample III	$S_{31}, S_{32}, \dots, S_{3n_3}$
Sample IV	$S_{41}, S_{42}, \dots, S_{4n_4}$

**Table 2** Ranking based on samples

Samples	Ranks assigned	Sum of ranks	Sample size
Sample I	Ranking as per data	$R_1$	$n_1$
Sample II	„	$R_2$	$n_2$
Sample III	„	$R_3$	$n_3$
Sample IV	„	$R_4$	$n_4$
Total	....	....	$N = n_1 + n_2 + n_3 + n_4$

### 4.3 Empirical Analysis

For each subject we noted the continuous assessment marks obtained by the student in Table 3. For example, in respect of Professional Core subject relating to computer science and engineering we have considered four sections (Sample I–IV). Ranking has been assigned to the observations depending on the magnitude of the marks obtained Table 4. An illustrative example is presented below.

A similar procedure has been followed for further analysis of the data in respect of ES, BS and Mandatory courses from Tables 5, 6, 7, 8, 9, 10.

In respect of electronics and communication engineering and mechanical engineering we have analyzed the data as per the same procedure stated above for all the categories of subjects. The computations are listed from Tables 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26.

**Table 3** Internal marks in respect of samples (PC)

Samples	Internal marks obtained
Sample I	22.15, 19.48, 21.85, 20.98, 23.58, 21.48, 20.45, 11.63, 19, 23.6
Sample II	22.6, 17.22, 12, 13.92, 12.17, 16.82, 20.82, 19.45
Sample III	23.05, 19.42, 18.27, 16, 16.75
Sample IV	19.85, 22.15, 21.4, 17.7, 11.75, 10.1, 11.65, 21.85, 19.97, 21.58, 8.1, 16.3, 16.55, 15.47, 20.43, 21.43, 10.8, 22.67, 13.6, 23.45

**Table 4** Assignment of ranks and computational details (PC)

Samples	Ranks assigned	Sum of ranks	Sample size
Sample I	36.5, 23, 34.5, 29, 42, 32, 27, 4, 20, 43	$R_1 = 291$	$n_1 = 10$
Sample II	38, 17, 7, 10, 8, 16, 28, 22	$R_2 = 146$	$n_2 = 8$
Sample III	40, 21, 19, 12, 15	$R_3 = 107$	$n_3 = 5$
Sample IV	24, 36.5, 30, 18, 6, 2, 5, 34.5, 25, 33, 1, 13, 14, 11, 26, 31, 3, 39, 9, 41	$R_4 = 402$	$n_4 = 20$
Total			$N = 43$

**Table 5** Internal marks in respect of samples (ES)

Samples	Internal marks obtained
Sample I	17.27, 12.22, 18.02, 17.65, 22.3, 18, 15.87, 11.12, 13.32, 17.75
Sample II	17.9, 12.98, 8.08, 10.9, 11.88, 14.7, 15.73, 18.2
Sample III	20.98, 20.73, 17.18, 14.1, 15.5
Sample IV	18.05, 22.15, 22.45, 14.75, 18.05, 10.9, 10.6, 23.9, 19.7, 21.4, 10.6, 20.2, 22.15, 18.6, 18.4, 19.15, 12.1, 22.2, 17.35, 22.25

**Table 6** Assignment of ranks and computational details (ES)

Samples	Ranks assigned	Sum of ranks	Sample size
Sample I	19, 9, 25, 21, 41, 24, 17, 6, 11, 22	$R_1 = 195$	$n_1 = 10$
Sample II	23, 10, 1, 4.5, 7, 13, 16, 28	$R_2 = 102.5$	$n_2 = 8$
Sample III	35, 34, 18, 12, 15	$R_3 = 114$	$n_3 = 5$
Sample IV	26.5, 37.5, 42, 14, 26.5, 4.5, 2.5, 43, 32, 36, 2.5, 33, 37.5, 30, 29, 31, 8, 39, 20, 40	$R_4 = 534.5$	$n_4 = 20$
Total			$N = 43$

**Table 7** Internal marks in respect of samples (BS)

Samples	Internal marks obtained
Sample I	41.1, 35.05, 41.15, 36.95, 45.45, 40.45, 41.35, 23.3, 44.35, 48.5
Sample II	42.75, 38.05, 22.15, 33.9, 26.25, 33.6, 46, 40.6
Sample III	47.05, 40.1, 38.4, 32.95, 39.05
Sample IV	47.9, 43.4, 48.1, 40.9, 27.85, 25.4, 26.2, 49.15, 45.65, 47.35, 17.7, 36.95, 39.75, 29.1, 36.3, 43.05, 20.7, 45.5, 30.15, 49.4

**Table 8** Assignment of ranks and computational details (BS)

Samples	Ranks assigned	Sum of ranks	Sample size
Sample I	26, 14, 27, 16.5, 33, 23, 28, 4, 32, 42	$R_1 = 245.5$	$n_1 = 10$
Sample II	29, 18, 3, 13, 7, 12, 36, 24	$R_2 = 142$	$n_2 = 8$
Sample III	37, 22, 19, 11, 20	$R_3 = 109$	$n_3 = 5$
Sample IV	39, 31, 40, 25, 8, 5, 6, 42, 35, 38, 1, 16.5, 21, 9, 15, 30, 2, 34, 10, 43	$R_4 = 450.5$	$n_4 = 20$
Total			$N = 43$

**Table 9** Internal marks in respect of samples (M)

Samples	Internal marks obtained
Sample I	73, 71, 75, 85, 76, 60, 67, 64, 80, 73
Sample II	79, 65, 55, 66, 54, 68, 71, 69
Sample III	87, 86, 68, 67, 68
Sample IV	72.5, 79, 87, 70, 49, 53.25, 59.75, 82, 84, 78.5, 43.75, 65, 63.25, 60.5, 72.5, 68.5, 66.5, 79, 64, 92

**Table 10** Assignment of ranks and computational details (M)

Samples	Ranks Assigned	Sum of ranks	Sample size
Sample I	28.5, 24.5, 30, 39, 31, 7, 16.5, 10.5, 36, 29.5	$R_1 = 252.5$	$n_1 = 10$
Sample II	33.5, 12.5, 5, 14, 4, 18.5, 25.5, 22	$R_2 = 135$	$n_2 = 8$
Sample III	41.5, 40, 20.5, 17.5, 19.5	$R_3 = 139$	$n_3 = 5$
Sample IV	26.5, 34.5, 42.5, 23, 2, 3, 6, 37, 38, 32, 1, 13.5, 9, 8, 27.5, 21, 15, 35.5, 11.5, 43	$R_4 = 429.5$	$n_4 = 20$
Total			$N = 43$

**Table 11** Internal marks in respect of samples (PC)

Samples	Internal marks obtained
Sample I	19.7, 20, 20.4, 18.8, 17.4, 21.5, 12.3, 10.7, 21.7, 16.3
Sample II	9.63, 16.65, 12.58, 11.08, 11.55, 15.38, 16.28, 20.4
Sample III	9.3, 23.3, 15.05, 12.05, 12.25
Sample IV	16.57, 15.72, 7.4, 3.87, 19.52, 11.07, 2.93, 15.48, 7.07, 10.54, 14.74, 13.51, 19.38, 18.42, 18.13, 9.67, 8.8, 19.62, 12.57, 16.01

**Table 12** Assignment of ranks and computational details (PC)

Samples	Ranks assigned	Sum of ranks	Sample size
Sample I	37, 38, 39.5, 33, 30, 41, 16, 10, 42, 27	$R_1 = 313.5$	$n_1 = 10$
Sample II	7, 29, 18, 12, 13, 22, 26, 39.5	$R_2 = 166.5$	$n_2 = 8$
Sample III	6, 43, 21, 14, 15	$R_3 = 99$	$n_3 = 5$
Sample IV	28, 24, 4, 2, 35, 11, 1, 23, 3, 9, 20, 19, 34, 32, 31, 8, 5, 36, 17, 25	$R_4 = 367$	$n_4 = 20$
Total			$N = 43$

**Table 13** Internal marks in respect of samples (ES)

Samples	Internal marks obtained
Sample I	33, 40.5, 42.3, 32.1, 42, 40.2, 21.9, 26.3, 45.5, 34.6
Sample II	11.1, 35.6, 25.6, 26.8, 21.1, 25.2, 32.5, 32.9
Sample III	17.6, 45, 36.5, 30.7, 15.2
Sample IV	41.15, 36.9, 20.45, 26.65, 44.2, 37, 17.95, 37.15, 23.65, 30.25, 33.05, 37.4, 38.45, 41.95, 42.4, 28.65, 28.8, 42.84, 37.9, 39.25

**Table 14** Assignment of ranks and computational details (ES)

Samples	Ranks assigned	Sum of ranks	Sample size
Sample I	21, 34, 38, 18, 37, 33, 7, 11, 43, 23	$R_1 = 265$	$n_1 = 10$
Sample II	1, 24, 10, 13, 6, 9, 19, 20	$R_2 = 102$	$n_2 = 8$
Sample III	3, 42, 25, 17, 2	$R_3 = 89$	$n_3 = 5$
Sample IV	35, 26, 5, 12, 41, 27, 4, 28, 8, 16, 22, 29, 31, 36, 39, 14, 15, 40, 30, 32	$R_4 = 490$	$n_4 = 20$
Total			$N = 43$

**Table 15** Internal marks in respect of samples (BS)

Samples	Internal marks obtained
Sample I	43.6, 37.7, 26.3, 28.45, 33.4, 42.25, 21.05, 25.65, 41.9, 34.8
Sample II	18.2, 35.85, 18.1, 26.25, 21.35, 29.25, 36.6, 43.05
Sample III	15.35, 43.9, 41.15, 27.85, 20.8
Sample IV	33.15, 31.25, 16.2, 20.5, 43.45, 33, 16.4, 29.3, 17.6, 32.4, 20.4, 27.85, 41.8, 38.65, 34.65, 23.9, 34.85, 49.1, 35.1, 31.2

**Table 16** Assignment of ranks and computational details (BS)

Samples	Ranks assigned	Sum of ranks	Sample size
Sample I	41, 33, 15, 18, 26, 38, 10, 13, 37, 28	$R_1 = 259$	$n_1 = 10$
Sample II	6, 31, 5, 14, 11, 19, 32, 39	$R_2 = 157$	$n_2 = 8$
Sample III	1, 42, 35, 16.5, 9	$R_3 = 103.5$	$n_3 = 5$
Sample IV	25, 22, 2, 8, 40, 24, 3, 20, 4, 23, 7, 16.5, 36, 34, 27, 12, 29, 43, 30, 21	$R_4 = 426.5$	$n_4 = 20$
Total			$N = 43$

**Table 17** Internal marks in respect of samples (M)

Samples	Internal marks obtained
Sample I	68, 71.25, 62.75, 65.75, 61.75, 79.5, 67.75, 54.75, 67.75, 68.5
Sample II	69, 66.5, 62.75, 73, 55.25, 67.5, 65.75, 77.5
Sample III	55.75, 86, 66, 55.25, 59.25
Sample IV	76, 79.5, 75, 52.25, 83.5, 66.25, 62.5, 64.25, 69.25, 58.25, 70.25, 65.25, 73, 70.5, 73, 56.25, 67, 66, 64, 71

**Table 18** Assignment of ranks and computational details (M)

Samples	Ranks assigned	Sum of ranks	Sample size
Sample I	26, 33, 11.5, 16.5, 9, 40.5, 24.5, 2, 24.5, 27	$R_1 = 214.5$	$n_1 = 10$
Sample II	28, 21, 11.5, 35, 3.5, 23, 16.5, 39	$R_2 = 177.5$	$n_2 = 8$
Sample III	5, 43, 18.5, 3.5, 8	$R_3 = 78$	$n_3 = 5$
Sample IV	38, 40.5, 37, 1, 42, 20, 10, 14, 29, 7, 30, 15, 35, 31, 35, 6, 22, 18.5, 13, 32	$R_4 = 476$	$n_4 = 20$
Total			$N = 43$

**Table 19** Internal marks in respect of samples (PC)

Samples	Internal marks obtained
Sample I	30.2, 20.25, 16.65, 14.35, 40.3, 24.65, 9.3, 9.7, 30.1, 13.75
Sample II	28.65, 32.1, 27.75, 40.45, 12.8, 23.2, 21.25, 28.75
Sample III	21.5, 23.05, 30.05, 12.55, 34.1
Sample IV	29.4, 8.9, 21.7, 40.6, 41.5, 42.2, 28, 20.9, 32.3, 12.1, 16.4, 24.9, 38.5, 14.6, 39.3, 10.9, 44.5, 20.9, 11.8, 26.3

**Table 20** Assignment of ranks and computational details (PC)

Samples	Ranks assigned	Sum of ranks	Sample size
Sample I	32, 14, 13, 10, 38, 22, 2, 3, 31, 9	$R_1 = 174$	$n_1 = 10$
Sample II	27, 33, 25, 39, 8, 21, 17, 28	$R_2 = 198$	$n_2 = 8$
Sample III	18, 20, 30, 7, 35	$R_3 = 110$	$n_3 = 5$
Sample IV	23, 1, 19, 40, 41, 42, 26, 15.5, 34, 6, 12, 29, 36, 11, 37, 4, 43, 15.5, 5, 24	$R_4 = 464$	$n_4 = 20$
Total			$N = 43$

**Table 21** Internal marks in respect of samples (ES)

Samples	Internal marks obtained
Sample I	35.3, 27.1, 26.35, 19.5, 43.4, 31.45, 14.2, 12.9, 37.75, 24
Sample II	21.15, 31.65, 19.25, 42.6, 7.45, 22.25, 25.8, 28.2
Sample III	24.7, 22.8, 42.35, 17.5, 37.6
Sample IV	30, 14.5, 26.3, 28.8, 42.2, 33.7, 20.5, 17.9, 26.2, 16.6, 12.6, 26.7, 26, 17.7, 34.3, 4.8, 43.9, 22.3, 9.5, 25.1

**Table 22** Assignment of ranks and computational details (ES)

Samples	Ranks assigned	Sum of ranks	Sample size
Sample I	36, 28, 26, 13, 42, 32, 6, 5, 38, 19	$R_1 = 245$	$n_1 = 10$
Sample II	15, 33, 12, 41, 2, 16, 22, 29	$R_2 = 170$	$n_2 = 8$
Sample III	20, 18, 40, 9, 37	$R_3 = 124$	$n_3 = 5$
Sample IV	31, 7, 25, 30, 39, 34, 14, 11, 24, 8, 4, 27, 23, 10, 35, 1, 43, 17, 3, 21	$R_4 = 407$	$n_4 = 20$
Total			$N = 43$

**Table 23** Internal marks in respect of samples (BS)

Samples	Internal marks obtained
Sample I	41.9, 24.15, 28.6, 28.4, 45.75, 38.45, 18, 18.7, 40.95, 22.55
Sample II	40, 43.2, 36.15, 42.2, 19.45, 29.15, 29.25, 33.6
Sample III	25.3, 32.6, 35.65, 7.8, 35.45
Sample IV	35.75, 13.25, 27.85, 40.65, 48.95, 47.65, 33, 29.45, 38.15, 18.65, 15.8, 33.75, 42.35, 26.3, 46.6, 21.4, 49.1, 27.4, 6.35, 32.15

**Table 24** Assignment of ranks and computational details (BS)

Samples	Ranks assigned	Sum of ranks	Sample size
Sample I	35, 11, 17, 16, 39, 31, 5, 7, 34, 10	$R_1 = 205$	$n_1 = 10$
Sample II	32, 38, 29, 36, 8, 18, 19, 24	$R_2 = 204$	$n_2 = 8$
Sample III	12, 22, 27, 2, 26	$R_3 = 89$	$n_3 = 5$
Sample IV	28, 3, 15, 33, 42, 41, 23, 20, 30, 6, 4, 25, 37, 13, 40, 9, 43, 14, 1, 21	$R_4 = 448$	$n_4 = 20$
Total			$N = 43$

**Table 25** Internal marks in respect of samples (M)

Samples	Internal marks obtained
Sample I	66, 72.5, 60.25, 61.75, 61.5, 82.5, 30.75, 50, 71, 63
Sample II	73, 69.5, 74, 72, 67, 54, 57, 65
Sample III	61, 64, 74, 59, 72.75
Sample IV	66.25, 53.75, 73.5, 71.25, 72, 71, 56.25, 60.25, 70, 59.75, 56.25, 54.25, 66, 61, 70.5, 60.5, 79, 65.25, 64.25, 62

**Table 26** Assignment of ranks and computational details (M)

Samples	Ranks assigned	Sum of ranks	Sample size
Sample I	24.5, 36, 11.5, 17, 16, 43, 1, 2, 31.5, 19	$R_1 = 201.5$	$n_1 = 10$
Sample II	38, 28, 40.5, 34.5, 27, 4, 8, 22	$R_2 = 202$	$n_2 = 8$
Sample III	14.5, 20, 40.5, 9, 37	$R_3 = 121$	$n_3 = 5$
Sample IV	26, 3, 39, 33, 34.5, 31.5, 6.5, 11.5, 29, 10, 6.5, 5, 24.5, 14.5, 30, 13, 42, 23, 21, 18	$R_4 = 421.5$	$n_4 = 20$
Total			$N = 43$

## 5 Summary of Results and Conclusion

We have carried out the computations relating to  $H$  statistic in respect of the departments viz., computer science and engineering, electronics and communication engineering and mechanical engineering for the categories of subjects considered in the above analysis. The details relating to  $H$  statistic and Chi-square are presented in Table 27.

We have drawn the inferences based on comparing  $H$  statistic with  $\chi^2$  value. The inferences drawn are noted below.

- (a) It is observed that the value of  $H$  statistic is smaller than  $\chi^2_{0.05}$  and this leads to the inference that there is no significant difference in the continuous assessment marks obtained by the students between the categories of subjects.
- (b) It is also observed that the basic aspect of Sentiment Analysis relating to the existence of difference in the marks obtained by the students do not differ between the categories of subjects offered by the departments. The educational planners have carefully evolved the categories of subjects though there exists difference between different engineering streams.

**Table 27** Final computations for different streams of engineering disciplines and categories of subjects in CA system

Departments	Test statistics and $\chi^2_{0.05}$	Categories of subjects			
		PC	ES	BS	M
CSE	$H$	4.3801	7.5137	1.6442	5.8951
	$\chi^2_{0.05}$	7.8200	7.8200	7.8200	7.8200
ECE	$H$	7.4598	6.9780	1.3623	1.7309
	$\chi^2_{0.05}$	7.8200	7.8200	7.8200	7.8200
MECH	$H$	1.9085	1.0189	1.3440	1.0150
	$\chi^2_{0.05}$	7.8200	7.8200	7.8200	7.8200

## References

1. Adedibu, A. A., Continuous Assessment in 6-3-3-4 system of education, in *Towards implementing 6-3-3-4. System of Education in Nigeria*, ed. by G. O. Akpa, S. U. Udo (Jos, Tech source) (1988)
2. G.A. Bordovsky, A.A. Nesterov, Yu TS Educational process quality management: monograph. St.Petersburg: RGPU named after. A.I. Hertsen (2001), pp. 359
3. S.N. Belova, Prognostic model of the internal system of quality assessment of the educational activities of a university. Scientific Notes: The online academic journal of Kursk State University (2008), Direct access: <http://cyberleninka.ru/article/n/prognosticheskaya-model-vnutrenneye-sistemy-otsenivaniya-kachestva-obrazovatelnogo-protsessa-v-vuze>
4. R. Hernandez, Does the continuous assessment in higher education support student learning? High. Educ. **64**(4), 489–502 (2012). <https://doi.org/10.1007/s10734-012-9506-7>
5. E. Guzman, D. Azócar, Y. Li, Sentiment analysis of commit comments in GitHub: an empirical study, in *Proceedings of the 11th Working Conference on Mining Software Repositories* (New York, NY, USA, 2014), pp. 352–355
6. S. Dhagat, P. Mukherjee, Continuous assessment—a new revolution in higher education. SIT J. Manag. **5**(1), 29–51 (2015)
7. W. Getinet Seifu, Assessment of the implementation of continuous assessment: the case of METTU university. Euro. J. Sci. Math. Edu. **4**(4), 534–544 (2016)
8. J. Onihunwa, O. Adigun, E. Irunokhai, Y. Sada, A. Jeje, O. Adeyemi, O. Adesina, Roles of continuous Assessment in Determining the Academic Performance of Computer Science Students in Federal College of Wildlife Management. Am. J. Eng. Res. (AJER) **7**(5), 07–20 (2018)
9. H. Gustavsson, M. Brohede, Continuous assessment in Software Engineering Project Course Using Publicly Available data from GitHub. Assoc. Comput. Mach. ACM ISBN 978-1-4503-6319-8/19/08. <https://doi.org/10.1145/3303446.3306446.3340820>.
10. W.H. Kruskal, W. Allen Wallis, Use of ranks in one criterion variance analysis. J. Am. Stat. Assoc. **47**(260), 583–621 (1952)
11. Y. Dodge, Kruskal-Wallis test, in *The Concise Encyclopedia of Statistics* (Springer, 2008), pp. 288–290. [https://doi.org/10.1007/978-0-387-32833-1\\_216](https://doi.org/10.1007/978-0-387-32833-1_216)

# Centralized and Decentralized Data Backup Approaches



Rahul Kumar and K. Venkatesh

**Abstract** In this digital era, data are very insecure, and data are being compromised with several risks, potential attackers, several methodologies, and mechanisms have been evolved to ensure backup of critical data. The reliable data backup technology ensures the reliability and availability of data over the network. Demand of safety and security for information and storage is increasing over the world day by day. Data are generated from a wide range of domain such as information technology sector, health and education sector, defense, banking, e-commerce, and telecommunication. Therefore, the backup of data plays a vital role to maintain the confidentiality, integrity, and availability of data for the end users. Decentralized data backup using blockchain technology can provide data confidentiality, data integrity, authentication, and authorization with digital signatures. This paper presents a review of various data backup techniques using centralized backup systems and discusses the problems and resolution associated with them and finally, how to resolve those issues with the help of decentralized data backup mechanisms using blockchain technology.

**Keywords** Backup · Backup techniques · Blockchain backup approach · Confidentiality · Availability

## 1 Introduction

In 1998, the victor of the ACM Turing Award Jim Gray proposed “Moore’s law”. He anticipated that the worldwide measure of data would twofold at regular intervals (approximately 18 months). References [1, 2] explain that the information backup requires more storage equipment, yet in addition implies more power and energy to

---

R. Kumar · K. Venkatesh (✉)

School of Computing, SRM Institute of Science and Technology, Kattankulathur, Chennai, India  
e-mail: [venkatek2@srmist.edu.in](mailto:venkatek2@srmist.edu.in)

R. Kumar  
e-mail: [rk5186@srmist.edu.in](mailto:rk5186@srmist.edu.in)

store backed up data. The globalizations of business have likewise raised progressively greater levels of popularity on the reinforcement, how to successfully backup the information has gotten one of the main points of contention. The requirement for backup fundamentally emerges from various sector's IT methodology guidelines. They require backup of data that will help to recover the data in case of a disaster or data loss. With the existence of big data and data science, backup of data is getting increasingly significant.

In paper [3, 4], authors have discussed about ways to handle delay sensitive data over a network which could be useful while taking a backup of large data or heterogeneous data types which in result will require less time to take a copy of the original data and sending it over to a secondary storage device. In case, if the delay occurs due to traffic then data packets can be prioritized based on their types and their sensitivity. In paper [5, 6], authors have discussed about improving the QoS while transmitting data over a network which will eventually be helpful to backup and recovery operation. In addition, with this, author has also discussed about the SDN which is very helpful to manage the network routes and devices virtually which in result will help to make efficient use of network hardware and will be cost-effective. In this present era, cloud computing plays a vital role to manage the data remotely and it is easily accessible everywhere. Authors in [7] have discussed about deployment of a private cloud which can be used to deploy a backup application in cloud to perform the backup and recovery operation remotely by utilizing the attributes of cloud computing.

Paper [8] refers that the data deduplication is utilized to discover repeatedly used data and store a sole duplicate, at that point utilizes a pointer of duplicate data and rest duplicate data is replaced with that pointer, at last just a single data block exists for the original information. Deduplication technique is used to manage the information, for instance, reinforcing same information from various places (e.g., email and social networking sites, various institutions, sectors); the greater part of which are "static" and incessant changed information (e.g., datasets, records, accounting pages), just as duplicate data.

This paper has been divided into following sections. Section 1 includes introduction; Section 2 includes survey of already existing centralized backup and decentralized backup approaches and their issues; and Sect. 3 includes conclusion and future scope.

## 2 Background

### 2.1 Centralized Backup

Paper [9–11] has discussed centralized backup approaches which is a technique that includes consequently copying information from point A and sending it over to point B or at several points. Centralized backups are used to reduce the administration costs

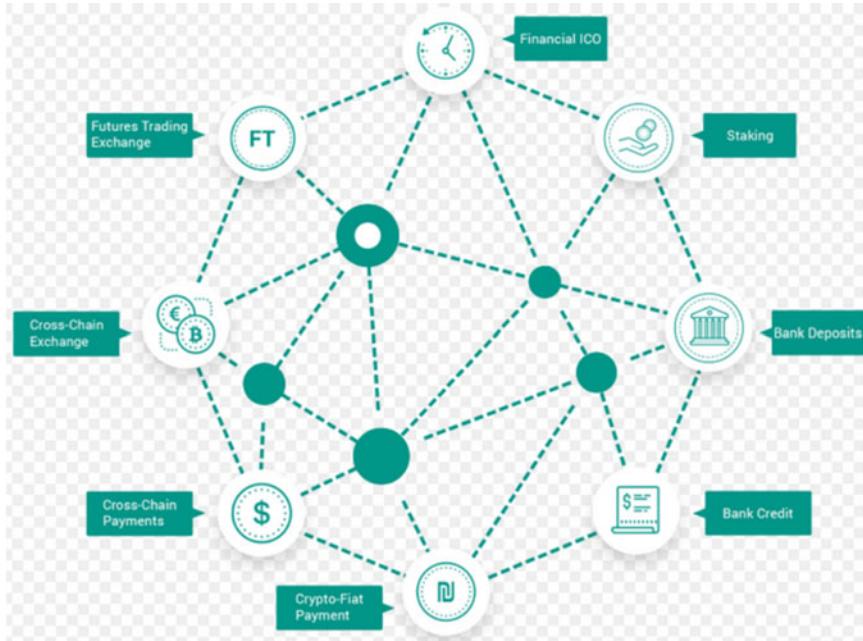
**Fig. 1** A centralized backup use case



by automatically managing backups at remote sites by taking a backup copies of full backup, incremental backup, differential, and cumulative backups. In layman term, a centralized backup is an alternative to a local backup. In general, centralized backup mechanism takes longer time to backups and restores a copy of original data which in result requires more bandwidth. A centralized data backup system could suffer data loss due to central server failure or dependency on third party can raise various concerns related to service level agreement, interoperability, policies, etc. (Fig. 1).

## 2.2 Decentralized Backup

As discussed in cryptokami Webpage [12], blockchain is a technology which is distributed in nature and works on peer to peer network. Blockchain used to create D-apps (decentralized applications). Blockchain technology is a great advancement in computer science technology. Blockchain is an incredible mechanism that can guarantee hundred percent correct transactions, no data loss, no human errors and makes sure that backups are secure and protected from any tampering or deletion. Tomaso Aste [13] has discussed the future of blockchain. With the existence of digital technology, data are more vulnerable to attackers and intruders, but blockchain powers define the accessibility of data to a user. Blockchain solves the issue of data privacy and data accessibility by following the stringent rules and by giving the role-based access at each level such that user, super user, and admin. as mentioned in Xenyta [14]. Blockchain has become an inevitable part of data protection division because it touches upon each and every pieces of data for backup and recovery purpose as discussed by authors in paper [15]. Currently, there are quite a few decentralized cloud-based data backup mechanisms like Filecoin, MaidSafe, Siacoin, and Storj. All these big players claim that the decentralized backup systems are superior to existing centralized backup systems in terms of uptime, security, and costs. All such



**Fig. 2** A decentralized backup use case

data protection division platforms rent the unused hard drive storage spaces from other hardware service providers and in some cases individual entities as well. In fact, any individual user can rent their unused storage spaces to earn money. For example, Uber and Ola make use of others vehicle when their owner doesn't make use of their four wheelers frequently. Similarly, blockchain technology-based backup and recovery service providers want to make the use of full potential of extra hard drive spaces which are potentially available for rent (Fig. 2).

### 3 Literature Review

#### 3.1 *Research and Implementation of a Data Backup and Recovery System for Important Business Areas, 2017*

Paper [1] mentions that a significant method for ensuring data protection, data backup mechanism plays a significant role. To tackle the issue of data availability and data integrity, researchers have found several techniques to backup and recovery heterogenous data types such as database, file systems, virtual machines, and operating systems. This paper has proposed business function which consists of backup

and recovery module and security enhancement function which consists of authentication, backup data management module, and data protection module. In existing data backup and recovery systems, along with data reliability and data availability, this paper has mainly focused on data confidentiality requirements by providing proper user authentication and authorization techniques. This paper uses the mainstream USB key which is added with a unique PIN code to achieve a two-factor mechanism-based authentications. Before recovering the backed up data and performing the recovery operation it performs an identity authentication a target machine wherein recovery operation is going to be performed by using the machine fingerprint technology by providing security identifiers (SID) and universally unique identifier (UUID), and finally by using the hash-based algorithms to get a machine fingerprint which is unique for each and every machine. By fortifying the user validation, the authenticity between the destination machine and data to be recovered, the cycle control during activities and the review of the key activity, the framework can prevent administrators from sponsorship up information to a non-authenticated storage system or disks.

### ***3.2 Heterogeneous Data Backup against Early Warning Disasters in Geo-Distributed Data Center Networks, 2018***

Paper [2], has discussed on how to manage data cost-efficiently for backup and recovery purpose to achieve either the maximum capacity of a backup or the minimum cost required to perform a backup. This paper mainly focuses on taking the backup of heterogeneous data types to a common place; it chooses different set of data candidate from data center network (DCN) service providers where all the users adhere to some service policy, safety norms which prefers that a particular set of data might be backed to a specific set of destination. In computer network, different set of data need to compete for required transmission bandwidth because different set of data would require different routes while transmitting the data over the network. This paper also has discussed about critical node for backup and recovery purposes which is kept dedicatedly and comes alive whenever a disaster scenario hits. A disaster scenario may consist of several node so in case of a node failure or overload, the workload can be shared among other nodes to maintain the high availability of data. This paper has finally concluded that the maximum data backup (MDB) backups schemes can be used to obtain the maximum amount of data which can be backed up and fair data backup (FDB) schemes can help to maximize the same backup proportion for heterogenous data types. In the proposed paper, author has performed several comparisons between MDB and FDB on maximum amount of data for several use cases for different time units, and both backup schemes have achieved optimal integer linear program (ILP) value for different set of data for backup and recovery scenario with efficient heuristics. In ILP, wherein, objective functions and objectives other than integer constraints are linear.

### ***3.3 Incremental Local Data Backup System Based on Bacula, 2018***

Paper [16, 17] authors have discussed how to protect data safely by preventing the data corruption, human errors, hardware failures, data tampering, and several other physical man-made and natural disaster scenarios. Along with this, it has focused on how to efficiently take incremental backups by only taking the backup of those data which is available after taking the last backup of data. This paper discusses on techniques to store the data efficiently because large amount of storage space is required to store the data which in result is very expensive in terms of time and space and thus resulting in big investment with less profit. It has also discussed about diff algorithm which are used on the file content. Diff algorithm differentiates the previous backed up data and recent incremental changes. This paper has implemented incremental backup and recovery mechanisms in local environment for file level data and databases. As part of data backup before taking incremental backup, first backup should always be a full backup and corresponding ones will be incremental backups. In database type data backup scenario, database backup module takes backup of only a single database by converting the database data into a SQL file. This paper has mentioned a problem while taking database backup wherein due to a defect in diff algorithm there are more recursive calls when the amount of data is large which is used to calculate the longest common subsequence which in result consumes more CPU.

### ***3.4 Lightweight Backup and Efficient Recovery Scheme for Health Blockchain Keys, 2017***

In paper [18], authors proposed a methodology for lightweight and efficient backup and recovery scheme for keys of health blockchain. Body sensor network (BSN) is used for the design of this backup and recovery scheme. This scheme can be used effectively for the protection of messages on health blockchain. It addresses privacy issues by building a key management for blockchain and BSN. The system consists of a blockchain and a BSN which is made up of wearable nodes implanted on a user's body. One of the nodes is gateway device. The implanted nodes collect signals from user's body and send them to gateway device. The gateway device forward this signal to respective hospitals which make a healthcare alliance. Each of these hospitals in the alliance provide a blockchain. When a hospital receives a signal from gateway device, its validity is checked using consensus mechanism. Upon successful validation, the data are stored on blockchain. Blockchain nodes associated with other hospitals in the alliance also store this data thereby solving monopoly problem of data. Data vulnerability problem is solved here by storing the data in multiple blockchain nodes. For the privacy problem, following approach is proposed: Before sending the signals from user's body to the gateway device, a key is produced by the blockchain

node. This key is used to encrypt the signals which is then sent to the gateway device. The user's data are safe in this case as the blockchain nodes and the hospital alliance are unaware of the encryption key. For data recovery, the key is produced again using user's body signals using which data are decrypted. This ensures that only user has access to his health data.

### ***3.5 Decentralized Distributed Blockchain Ledger for Financial Transaction Backup Data, 2019***

In [19, 20], authors proposed a methodology to secure financial transaction data by designing a backup mechanism in a decentralized network using blockchain technique. Because of decentralized blockchain, a copy of data resides in each node which ensures that failure in one node does not affect the entire system. In this application, a financial transaction block is constructed which consists of transactions data along with their timestamps. Each of these blocks, which hold the data that needs to be protected, are chained together to the previous block. The system consists of n number of participating nodes. To ensure validity of participating node, two authentication schemes are identified. In a normal authentication scheme, a node is validated for legitimacy. In a mutual authentication scheme, two nodes communicate with each other in order to identify each other. Since the verification of participating nodes is achieved through consensus, the transaction blocks stored in blockchain are traceable and immutable. With this approach, transaction data are verified transparently.

### ***3.6 Secure Deduplication for Images Using Blockchain, 2020***

In paper [21], authors proposed a Web-based platform in which data deduplication and blockchain are integrated to make storage of images effective, secured, and efficient. Data deduplication is an approach for duplicate data elimination to improve storage efficiency. In this work, SHA256 algorithm is used for identification of duplicate images. When a user uploads an image to be stored in cloud, the hash digest of the image is created. The image is then stored if the digest is unique. To secure images from unauthorized access, blockchain technology is used. Using blockchain, data can be shared between peers without any third-party intermediaries. Images uploaded by the user and the corresponding hash value of the image are considered as blocks. These blocks are broadcasted for copyright purpose. For the validation of the blocks, a difficulty parameter is calculated based on protocol agreed upon by all the peers and is stored for future purpose. This calculated number is broadcasted to all the peers. Thus, the ownership of the image is validated.

### ***3.7 Data Backup and Recovery Techniques in Cloud Computing, 2018***

In paper [22, 23], author has proposed that to maintain a large amount of data we need a service to perform the recovery operation sophistically on a cold backup and hot backup. A cold backup is done passively, while a hot backup is done actively. Cold and hot backup operation occurs whenever a service failure is detected inline. Author has also discussed about a seed block algorithm which is used to take remote backup of an original data in a smart manner. This algorithm performs its operation in two ways. Firstly, it gathers the information remotely to take the backup, and secondly, it recovers all the data from a remote location which might be deleted or could falls under a potential data loss use case. As mentioned, seed block algorithm also reduces the time factor to recover a set a data as compared to other algorithms. To protect data remotely from disaster, author has proposed a scenario wherein author is using advanced recovery methods line ARIES, concurrency control methods and making use of parallel streams efficiently to perform the similar operation to parallelly to save time which thus enhances the data backup and recovery performances. Each backup system contains at least one backup therefore organizing a backup copy is very important in terms of storage space. Thus, a data model is presented to organize data efficiently in each storage space. This paper has also discussed about parity cloud service (PCS) technique which is a convenient method to recover a data with high probability. All the abovementioned backup and recovery techniques have its own advantage and disadvantage but all of them serve their best performance in a given scenario.

### ***3.8 Research and Implementation of Data Storage Backup, 2018***

Paper [24] discusses the fact that due to advancement in data science and big data applications, need for data backup and storage spaces have become significantly important. Nowadays, industry is looking for a simple, reliable, safe, and flexible backup mechanism to provide backup and recovery operation for data, systems, databases, and applications. Given paper has discussed about backup schemes which could be performed in hybrid network environments such as SAN and NAS. It also mentions about a system backup which can take automatic backup without any human intervention by following a backup strategy and protection life cycle to protect the data. Data backup is mainly referred to backup up business data, mission critical data etc. Depending upon the scenario, data backup can be a full backup, cumulative backup, incremental backup, log backup, and a transactional backup. Depending upon the policy life cycle, a data backup can be taken by transmitting a backup copy to the backup server and storing the data to some physical location locally or remotely such as disk, hard drives, tape drives, or data domains. As discussed in the

paper, whenever a disaster occurs, agent installs the restore software in the agent, wherein all the required data are restored and is sent back to the physical or virtual storage space. Paper [25] has discussed about the government-related data which is very confidential by nature that in a disaster-related scenario such data are been prioritized to take a backup and to recover such data to a remote locations safely by maintaining the security standards and role-based access to a respective users.

### ***3.9 Decentralizing Privacy: Using Blockchain to Protect Personal Data, 2015***

In [26], author proposes a management system for personal data which is decentralized to ensure user's control over their personal data. Focus in this work is to solve the privacy issues on mobile platforms where applications collect data in the background without user's control. Here, blockchain is used as a manager to control access to data thereby eliminating the need for third-party services. Blockchain is designed to accept two types of transactions: access control and data storage transactions. When a user accesses the system for the first time, a unique identity is created along with set of permissions. These are sent to blockchain as an access control transaction. Data, which is collected on phone, are first encrypted and then sent to blockchain as a data storage transaction. The user can then query the data whenever required. Blockchain verifies the user based on digital signature. User can change the permissions at any time by sending new permissions to blockchain as an access control transaction. Thus, user personal data are protected using blockchain without third-party services.

### ***3.10 General Data Protection Regulation Complied Blockchain Architecture for Personally Identifiable Information Management, 2018***

Paper [27] focuses on personally identifiable information (PII) collected for market research and analysis is being misused massively. PII includes name, address, email, unique identification number, and fingerprints. In [28], authors proposed a private blockchain-based architecture that uses distributed ledgers to secure PII lifecycle. When an enormous amount of PII is collected from user, an organization uses this data for market analysis. The proposed system delivers the PII from users to controller. A contract is signed between user and controller during the flow of information. The contract and information other than PII are stored in blockchain. Smart contracts which includes terms and conditions are created based on the consensus between user and controller. This contract is added as a new block to blockchain. PII of users is stored in controller's local databases. Hash of the local database where PII is stored is also added to blockchain as a new block. Once all the data are added to blockchain,

the data along with contracts are broadcasted among all the nodes. For sharing data, nodes need to get user's consent for accessing personal information. The controller shares this data with processor for analysis, when all nodes agree for data sharing. Thus, the proposed work reduces the risk of PII leak.

## 4 Conclusion

In this information era, due to existence of digital solution, almost all the data are being stored digitally in storage devices either locally or remotely. As a result, data have become gold for each and every institutions, sectors, and defense academies, and these data have become so valuable for the continuity of their business operation. Similarly, for research organizations, relevant data are archived for reusability. Therefore, to backup such valuable data which could be image, text, a virtual machine, operating system, etc., researchers and individual contributors have invented various backup and recovery techniques which can be used to backup, clone, recover, and restore original data to multiple remote locations to ensure availability of data across the globe over a network. This paper has an analysis of several centralized backup mechanisms and decentralized backup mechanism using blockchain technology where centralized backup schemes use a traditional approach to backup data using a central server where various cost-effective techniques, deduplication methodology, diff algorithms are used to store a copy of data efficiently, while decentralized backup schemes have used blockchain technology to store a copy of data to multiple locations by maintaining the data privacy, data security, data confidentiality and before backup makes sure that original data are coming from the legitimate source by maintaining a digital signature between sender and receiver. Decentralized backup mechanisms using blockchain also makes sure that no original data are been tampered by maintaining a hash table for each data transactions. Several players in decentralized backup technology have claims that blockchain technology-based backup and recovery products are far better than traditional centralized backup and recovery mechanisms.

## 5 Future Scope

At present, all the existing players like CryptoKami, and MaidSafe have public blockchain-based backup systems to store data on peer to peer network which in most of the cases run on Ethereum platform. In future, we will try to find the methodologies and requirements to implement a private blockchain-based decentralized data backup system using Hyperledger fabric [15] which is a framework which helps us to create a private blockchain network. We will also have a comparison between centralized and decentralized backup system on several parameters in terms of time and space to back up a fixed data unit. A proposed methodology in future scope will focus

on a backup mechanism which will serve the purpose of private organizations like schools, colleges, hospitals, and banks.

## References

1. H.L. Jianping Zhang, Research and implementation of a data backup and recovery system for important business areas, in *International Conference on Intelligent Human-Machine Systems and Cybernetics, Mianyang* (2017)
2. R.W.S.X.L.B.W., X.J. Lisheng Ma, Heterogeneous data backup against early warning disasters in geodistributed data center networks. *J. Opt. Commun. Netw.*, **10**, 376–385 (2018)
3. K. Venkatesh, L.N.B. Srinivas, M.B. Mukesh Krishnan, A. Shanthini, QoS improvisation of delay sensitive communication using SDN based multipath routing for medical applications. *Future Gen. Comput. Syst. [Int.]* **93**, 256–265 (2019)
4. V.A. Reddy, K. Venkatesh, L.N.B. Srinivas, Software defined networking based delay sensitive traffic engineering of critical data in internet of things. *J. Comput. Theor. Nanosci. [Int.]* **17**(1), 48–53 (2020)
5. R. Kumar, K. Venkatesh, SDN-based QOS-aware multipath routing mechanism using open-stack. *Int. J. Pure. Appl. Math. [Int.]* **118**(20A), 357–363 (2018)
6. V. Reddy, K. Venkatesh, Role of software-defined network in industry 4.0”, in *EAI/Springer Innovations in Communication and Computing* (2020), pp.197–218
7. B. Sukesh, K. Venkatesh, L.N.B. Srinivas, 5 a custom cluster design with Raspberry Pi for parallel programming and deployment of private cloud, in *Role of Edge Analytics in Sustainable Smart City Development: Challenges and Solutions [International]* (2020), pp.273–288
8. F.E. Guo, Building a high-performance deduplication system, in *11th USENIX Annual Technical Conference* (Portland, 2011)
9. L.Z.M.X. Quanqiu Xu, YuruBackup: a space-efficient and highly scalable incremental backup system in the cloud. *Int J Parallel Prog* **43**, 316–338 (2015)
10. S.C. Jun Liu, HVF: An optimized data backup and recovery system for hard disk based consumer electronics, in *Optik* (Elsevier, 2015), pp. 251–257
11. W.S.B.W.T.T.X. J., N.S.L. Ma, ε time early warning data backup in disaster-aware optical inter-connected data center networks. *Opt. Commun. Netw.* **9**, 536–545 (2017)
12. Triadi, cryptokami-blockchain-decentralized-backup-system, 2017. Available: <https://steemit.com/cryptokami/@triadi/cryptokami-blockchain-decentralized-backup-system>. Accessed October 2020
13. P.T.D.M. Tomaso Aste, in *Blockchain Technologies: The Foreseeable Impact on Society and Industry* (IEEE, London, 2017)
14. M.C. Xenya, Decentralized distributed blockchain ledger for financial transaction backup data, in *International Conference on Cyber Security and Internet of Things (ICSIoT)* (Ghana, 2019)
15. M.T.H. e. al., Bubbles of Trust\_a decentralized Blockchain-based authentication. *Comput Secur IEEE* (2018)
16. Y. C. et al., Incremental local data backup system based on bacula, 2018, in *IEEE International Conference of Safety Produce Informatization (IICSPI)* (2018), pp. 429–432
17. X.L.L.Z.Z.H. Zhao, Data integrity protection method for microorganism sampling robots based on blockchain technolgy, in *Natural Science Edition* (2015)
18. Y.Z.Y.P.R.X. Huawei Zhao, Lightweight backup and efficient recovery scheme for health blockchain, in *IEEE 13th International Symposium on Autonomous Decentralized Systems* (Jinan, 2017)
19. S.F.H.W.N.A.A.J. Kishigami, The blockchain-based digital content distribution system, in *IEEE 5th International Conference on Big Data and Cloud Computing* (Dalian, 2016)
20. N.K.S.S.R.S. Rishabh Mehta, Decentralised image sharing and copyright protection using blockchain and perceptual hashes, in *11th International Conference on Communication Systems & Networks* (Delhi)

21. R.G.K., S.C.R. Aparna, in *Secure Deduplication for Images using Blockchain* (IEEE, Bengaluru, 2020)
22. K. Laxmi, K. Deepika, N. Pranay, V. Supriya, Data backup and recovery techniques in cloud computing. *Int J Sci Res Comput Sci, Eng Inf Technol* **9**, 1002–1005 (2018)
23. Vijaykumar Javaraiah, Brocade Advanced Networks and Telecommunication Systems (ANTS), Backup for cloud and disaster recovery for consumers and SMBs, in *IEEE 5th International Conference* (2011)
24. Y. Zhao, N. Lu, Research and implementation of data storage backup, in *IEEE International Conference on Energy Internet* (Beijing, 2018), pp. 181–184
25. B. Zhu, B. Liu, Research and implementation of e-government data disaster recovery in administrative units. *Comput. Knowledge Technol.* **21**, 5059–5056 (2011)
26. G. Zyskind, O. Nathan, Alex ‘Sandy’ pentland decentralizing privacy: using blockchain to protect personal data, in *IEEE CS Security and Privacy Workshops*(2015)
27. N. Al-Zaben et al., General data protection regulation complied blockchain architecture for personally identifiable information management, in *International Conference on Computing, Electronics & Communications Engineering (iCCECE)* (Southend, 2018), pp. 77–82
28. N. Heath, Difference-between-private-public-blockchain, 05 September 2018. Available: <https://www.intheblack.com/articles/2018/09/05/difference-between-private-public-blockchain#:~:text=A%20public%20blockchain%20is%20an,or%20participate%20in%20the%20network.&text=A%20private%20blockchain%20is%20an,write%20or%20audit%20the%20blockchain>. Accessed 17 July 2020

# Author Index

## A

- Adak, Shilpi, 525  
Amutha, B., 637  
Ananthapadmanaban, K. R., 33, 43  
Ananya, P. R., 551  
Annapurani, K., 191, 205, 581  
Anumula, Kalyan, 355  
Arivazhagan, N., 323  
Arunnehru, J., 179  
Arun Prasath, G., 191  
Avadhanam, Meghana, 119  
Ayatti, Sudha, 19

## B

- Bangarashetti, Sadhana P., 19  
Bharadwaj, Tarun, 95  
Bogdanović, Milena, 237  
Brijpuriya, Shatakshi, 167  
Brindha, M., 155

## D

- Dixit, Vaibhav, 401  
Dorathi Jayaseeli, J. D., 213  
Dusane, Palash, 83

## F

- Fancy, C., 271  
Fathima Najiya, M., 323  
Ferni Ukrit, M., 571, 631

## G

- Ganesan, S. Selva, 505

Gautham, S. V., 457

- Godwin, J., 287  
Gouri, K., 19  
Grace, Anne Lourdu, 593  
Gupta, Nikita, 373

## H

- Harish Kumar, J., 105  
Hayani, Bashar, 227  
Hemavathi, D., 457, 605

## I

- Iniyan, S., 335  
Iswariya, M. S., 205

## J

- Jebakumar, R., 335  
Jha, Vishal, 401  
John, Vivia Mary, 413, 419, 651

## K

- Karthick, T., 1  
Kesana, Satwika, 119  
Kheerthana, R., 505  
Kirthiga Devi, T., 105  
Kumar, Aditya, 271  
Kumar, Rahul, 687  
Kunchur, Pavan, 19  
Kuppusamy, Palanivel, 65

**L**

Lavanya, R., 539  
Lekha, K., 257

**M**

Malathi, D., 213  
Malathy, C., 245  
Metilda Florence, S., 295  
Mittal, Shreya, 95  
Muruganandam, S., 33, 43

**N**

Naga Malleswari, T. Y. J., 95, 119, 257  
Nimala, K., 389  
Niranjana, G., 51  
Nithiya, S., 305, 621  
Nithyakani, P., 505, 571

**P**

Pachisia, Vedika, 551  
Parekh, Kunal H., 373  
Parimala, G., 305, 621  
Patnaik, Satish Chandra, 561  
Patyal, Jayvardhan Singh, 561  
Paulraj, M., 155  
Pillaithambi, Nikil, 539  
Ponnusamy, Vijayakumar, 237  
Poovammal, E., 227  
Prince, Michelle Catherina, 443

**R**

Raam, V. S. Bharat, 539  
Rajalakshmi, M., 167, 437, 581  
Rajaram, V., 437  
Ramamoorthy, S., 143  
Ramani, K., 51  
Ramprasath, M., 1  
Rani, L. Paul Jasmine, 437  
Rawat, Sachin, 485  
Raymond, Joseph, 355, 363  
Razia Sulthana, A., 621  
Renukadevi, G., 651  
Renukadevi, P., 133  
Rifas, S. Mohammed, 419  
Roy, Mukul Lata, 213

**S**

Sabhanayagam, T., 349  
Saini, Akash, 515

Sangeetha, M., 1

Saranya, K., 155  
Saranya, P., 561  
Saveetha, D., 313  
Saxena, Sagar, 389  
Selvaraj, P., 363  
Shanthini, A., 427, 671  
Shrikrishna, A., 505  
Silpa, C., 51  
Sindhu, C., 525  
Sindhu, S., 469, 605  
Sornalakshmi, K., 515, 605  
Sreehari, T., 469

Srinivasan, Sujatha, 33, 43  
Srinivas, L. N. B., 443  
Srividhya, S., 305, 437  
Subbarayan, A., 671  
Subburaman, Dhivya, 495  
Sugirtham, A. Veronica Nithila, 245  
Sujatha, G., 83, 605  
Sujithra, J., 631  
Suresh Joseph, K., 65  
Surya, S., 143  
Syed Mohamed, M., 133

**T**

Teja, Burla Sai, 413  
Thenmozhi, M., 593  
Tigga, Soumya Celina, 525  
Tom, Rijo Jackson, 651  
Trajanović, Miroslav, 237

**U**

Udhaya Mugil, D., 295  
Umamaheswari, K. M., 561  
Usharani, R., 427  
Ushasukhanya, S., 551

**V**

Vadivu, G., 485  
Vamsi, R. V. Krishna, 495  
Vanusha, D., 637  
Venkatesh, K., 687  
Vijay Bharath, A., 671  
Viji, D., 373, 401  
Vishal Balaji, D., 179

**W**

Wadhwa, Anurag, 505

**Y**

- Yadav, Nisha, [257](#)  
Yallamandhala, Pavithra, [287](#)  
Yuvaraman, S., [313](#)

**Z**

- Zdravković, Nemanja, [237](#)