

Lecture Notes in Networks and Systems 528

P. P. Joby
Valentina E. Balas
Ram Palanisamy *Editors*

IoT Based Control Networks and Intelligent Systems

Proceedings of 3rd ICICNIS 2022

 Springer

Lecture Notes in Networks and Systems

Volume 528

Series Editor

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

Advisory Editors

Fernando Gomide, Department of Computer Engineering and Automation—DCA,
School of Electrical and Computer Engineering—FEEC, University of
Campinas—UNICAMP, São Paulo, Brazil

Okyay Kaynak, Department of Electrical and Electronic Engineering,
Bogazici University, Istanbul, Turkey

Derong Liu, Department of Electrical and Computer Engineering, University of
Illinois at Chicago, Chicago, USA

Institute of Automation, Chinese Academy of Sciences, Beijing, China

Witold Pedrycz, Department of Electrical and Computer Engineering, University of
Alberta, Alberta, Canada

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

Marios M. Polycarpou, Department of Electrical and Computer Engineering,
KIOS Research Center for Intelligent Systems and Networks, University of Cyprus,
Nicosia, Cyprus

Imre J. Rudas, Óbuda University, Budapest, Hungary

Jun Wang, Department of Computer Science, City University of Hong Kong,
Kowloon, Hong Kong

The series “Lecture Notes in Networks and Systems” publishes the latest developments in Networks and Systems—quickly, informally and with high quality. Original research reported in proceedings and post-proceedings represents the core of LNNS.

Volumes published in LNNS embrace all aspects and subfields of, as well as new challenges in, Networks and Systems.

The series contains proceedings and edited volumes in systems and networks, spanning the areas of Cyber-Physical Systems, Autonomous Systems, Sensor Networks, Control Systems, Energy Systems, Automotive Systems, Biological Systems, Vehicular Networking and Connected Vehicles, Aerospace Systems, Automation, Manufacturing, Smart Grids, Nonlinear Systems, Power Systems, Robotics, Social Systems, Economic Systems and other. Of particular value to both the contributors and the readership are the short publication timeframe and the world-wide distribution and exposure which enable both a wide and rapid dissemination of research output.

The series covers the theory, applications, and perspectives on the state of the art and future developments relevant to systems and networks, decision making, control, complex processes and related areas, as embedded in the fields of interdisciplinary and applied sciences, engineering, computer science, physics, economics, social, and life sciences, as well as the paradigms and methodologies behind them.

Indexed by SCOPUS, INSPEC, WTI Frankfurt eG, zbMATH, SCImago.

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).

P. P. Joby · Valentina E. Balas · Ram Palanisamy
Editors

IoT Based Control Networks and Intelligent Systems

Proceedings of 3rd ICICNIS 2022

 Springer

Editors

P. P. Joby
Department of Computer Science
and Engineering
St. Joseph's College of Engineering
and Technology
Palai, Kerala, India

Valentina E. Balas
Aurel Vlaicu University of Arad
Arad, Romania

Ram Palanisamy
Gerald Schwartz School of Business
Saint Francis Xavier University
Antigonish, NS, Canada

ISSN 2367-3370

ISSN 2367-3389 (electronic)

Lecture Notes in Networks and Systems

ISBN 978-981-19-5844-1

ISBN 978-981-19-5845-8 (eBook)

<https://doi.org/10.1007/978-981-19-5845-8>

© The Editor(s) (if applicable) and The Author(s), under exclusive license
to Springer Nature Singapore Pte Ltd. 2023

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

We would like to dedicate this proceeding to all members of advisory committee and program committee for providing their excellent guidance. We also dedicate this proceeding to the members of the review committee for their excellent cooperation throughout the conference. We also record our sincere thanks to all the authors and participants.

Preface

On behalf of the conference committee, it is our pleasure to extend a warmest welcome to all of the attendees of our International Conference on IoT Based Control Networks and Intelligent Systems [ICICNIS 2022]. Its main objective and feature are to bring together academia, scientists, engineers, and industry researchers to interact and share their experience and research results in most areas of control networks and intelligent systems, as well as to explore the real-time challenges and solutions adopted to it.

With a remarkable array of keynote and invited speakers from different parts of the globe, ICICNIS 2022 promises to be both interesting and informative. Delegates can participate in the wide range of technical presentation sessions to gain critical insights on the recent findings in their area of expertise. It includes a selection of 64 papers from 308 papers submitted to the conference from universities and industries all over the world.

The conference program includes invited talks, technical presentations, and conversations with eminent speakers on a wide range of control networks and information system research issues. This extensive program enables all attendees to meet and connect with one another. We guarantee you a productive and long-lasting experience at ICICNIS 2022. The conference has offered a truly comprehensive view while inspiring the attendees to come up with significant recommendations to tackle the emerging challenges. The conference will continue to thrive with your help and participation for a long period of time.

The editors would like to express their sincere appreciations and thanks to all the authors for their contributions to this ICICNIS 2022 publication. Our special thanks go to the conference organization committee, the members of the technical program committee and reviewers for their devoted assistance in reviewing papers and making valuable suggestions for the authors to improve their work. We would also like to thank the external reviewers for their assistance in the review process, as well as the

authors for sharing their state-of-the-art research results to the conference. Special thanks to Springer Publications.

Palai, India
Arad, Romania
Antigonish, Canada

Guest Editors
P. P. Joby
Valentina E. Balas
Ram Palanisamy

Contents

LabVIEW Based Anomaly Detection for Screening Diabetic Retinopathy	1
Sheena Christabel Pravin, K. Sindhu Priya, S. Suganthi, J. Saranya, and V. S. Selva Kumar	
Early Stage Parkinson’s Disease Diagnosis and Detection Using K-Nearest Neighbors Algorithm	15
S. Svetaa, S. Pavithra, and R. G. Sakthivelan	
IoT Crypto Security Communication System	27
Kiran Kumar Kommineni, G. C. Madhu, Rajadurai Narayanamurthy, and Gurpreet Singh	
CNN and XGBoost Based Hybrid Model in Classification of Fetal Ultrasound Scan Planes Images in Detection of Congenital Heart Defects	41
S. Satish and S. Sridevi	
Review on Web Application Based Infant Health Management	55
Deekshitha S. Nayak, Sannidhi Rao, Sharan Kumar, Niharika Rao, and Himanshu Bhatt	
IEEE 802.11g Wireless Protocol Standard: Performance Analysis	67
G. U. Shreelatha and M. K. Kavyashree	
Optimal Cluster Head Selection Using Vortex Search Algorithm with Deep Learning-Based Multipath Routing in MANET	81
S. Venkatasubramanian	
Secure Wireless Smart Car Door Unlocking System	99
Navneet Vinod Melarkode, Varun Niraj Agarwal, Avaneesh Kanshi, and Anisha M. Lal	

A Comprehending Deep Learning Approach for Disease Classification	113
Ankita Nainwal, Bhaskar Pant, and Garima Sharma	
Restaurant Automation Through IoT and NLP Techniques	123
Partheesh Ranjan Singh, Tejas Kumar Nazre Amarnath, Mohammed Khurram, Swathi Tripathi, Swathi Tripathi, and Deepti Chandrasekharan	
Comprehensive Assessment of Big Data in Recommendation Systems	139
Swati Dongre and Jitendra Agrawal	
Review on Application of Wireless Technology Using IoT	161
Deekshitha S. Nayak, N. Akshaya Krishna, Sahana Shetty, Sukanya D. Naik, V. Sambhram, and Krishang Shetty	
Comparative Study of Conditional Generative Models for ISL Generation	171
Marrivada Gopala Krishna Sai Charan, S. S. Poorna, K. Anuraj, Choragudi Sai Praneeth, P. G. Sai Sumanth, Chekka Venkata Sai Phaneendra Gupta, and Kota Srikar	
A Novel Approach for Segmenting Coronary Artery from Angiogram Videos	191
K. Kavipriya and Manjunatha Hiremath	
Analysis of Different Cryptographic Algorithms in Cloud-Based Multi-robot Systems	201
Saurabh Jain, Shireen Rafat Alam, and Rajesh Doriya	
An IoT Based Environment Monitoring & Controlling System for Food Grain Warehouse	217
Vrinda Parkhi, Nishant Chavhan, Sharda Chandak, Bhushan Chaware, and Prasad Bongarde	
Numerical Analysis and ANN Modeling of the Intercooled, Reheat and Regenerative Gas Turbine Cycle	229
Milind S. Patil, Shyamkumar D. Kalpande, Sanjay P. Shekhawat, and Chandrashekhar D. Mohod	
Implementation of IoT Enabled Home Automation System	251
Dikshan Shah	
Scheduled Line of Symmetry Solar Tracker with MPT and IoT	265
A. B. Gurulakshmi, Sanjeev Sharma, N. Manoj, Nikhil A. Bhinge, H. M. Santhosh, and O. M. Yogesh	

A New Method for Secure Transmission of Medical Images Using Wavelet Transform and Steganography 275
 S. Jayanth, Y. Sushwanth, Poornima Mohan, N. Tejesh, and M. Pavan

Deep Learning Based Hand Sign Recognition in the Context of Indian Greetings and Gestures 287
 Rohan Saxena, Romy Garg, Bhoomi Gupta, and Narinder Kaur

Pedestrian Detection Using MobileNetV2 Based Mask R-CNN 299
 Sonal Sahu, Satya Prakash Sahu, and Deepak Kumar Dewangan

Localization of Calcifications in Mammograms Using CNN with GAP Layer 319
 Praneeth Vykuntam, Venkata Rohith Vykuntam, Pragun Srivastav, Sri Sai Bharat Uppalapati, and Poornima Mohan

Comparative Analysis of Machine Learning and Deep Learning Algorithms for Real-Time Posture Detection to Prevent Sciatica, Kyphosis, Lordosis 331
 Palavalasa Venkata Satish and Meena Belwal

Lightweight Block Cipher for Resource Constrained IoT Environment—An Survey, Performance, Cryptanalysis and Research Challenges 347
 M. Abinaya and S. Prabakeran

Image Classification Using Quantum Machine Learning 367
 Amrit Raj and Jayakumar Vaithiyashankar

An Efficient Algorithm for Multi Class Classification in Deep Neural Network 381
 Pranamita Nanda and N. Duraipandian

Lung Cancer Classification System for CT Images using Deep Convolutional Neural Network 395
 A. Jayachandran and N. Anisha

Performance Analysis of Machine Learning Algorithms in Heart Diseases Prediction 407
 K. Nanthini, M. Pyingkodi, D. Sivabalaselvamani, Shweta Kumari, and Tarun Kumar

Fluorescence Microscopic Image Reconstruction Using Variational Autoencoder and CycleGAN 425
 Marrivada Gopala Krishna Sai Charan, S. S. Poorna, K. Anuraj, Choragudi Sai Praneeth, P. G. Sai Sumanth, Chekka Venkata Sai Phaneendra Gupta, and Kota Srikar

Tomato Leaf Disease Detection Based on Convolutional Neural Network 437
 Jagmohan Sahu and Pavan Kumar Mishra

PDF Steganography Using Hybrid Crypto Encryption Technique 453
 Sunil Kumar Patel and Saravanan Chandran

An Efficient Classification Algorithm for Employee Well-Being Prediction Using Deep Learning 467
 S. Sunandha Shri and M. Ezhilarasan

UAV-Enabled Supply Chain Architecture for Flood Recovery in Smart Cities 483
 Theodoros Anagnostopoulos, Faidon Komisopoulos, Ioannis Salmon, and Klimis Ntalianis

Outlier-Based Sybil Attack Detection in WSN 497
 A. Jeyasekar, S. Antony Sheela, and J. Ansulin Jerusha

An Innovative Novel Method of Reducing the Impact of Traffic Jam Using the Vehicular Ad-Hoc Network 519
 Md. Nahidul Alam, Shahrukh Hossain Rian, Maruf Haider Chowdhury, Md. Rayhanul Islam, and Mahfuz Ullah

Navbot—College Navigation Chatbot Using Deep Neural Network 533
 M. Sobhana, A. Yamini, K. Hindu, and Y. L. Narayana

Heart Problems Diagnosis Using ECG and PCG Signals and a K-Nearest Neighbor Classifier 547
 Youssef Toulmi, Benayad Nsiri, and Taoufiq Belhoussine Drissi

Deep Convolutional Neural Network for Multi-class Brain Tumor Classification System in MRI Images 561
 A. Jayachandran, M. A. Sreema, S. P. Anandaraj, and T. Sudarson Rama Perumal

Application of the Particle Swarm Algorithm to the Task of Image Segmentation for Remote Sensing of the Earth 573
 Igor Ruban, Hennadii Khudov, Oleksandr Makoveichuk, Igor Butko, Sergey Glukhov, Irina Khizhnyak, Nazar Shamrai, and Temir Kalimulin

Overview of Data Center Link Load Balancing Technology Based on SDN 587
 Feifan Hao, Shan Jing, and Chuan Zhao

Wearable Band for Safety in Chemical Industries 603
 D. Diana Josephine, R. Ajay Kumar, M. Ganesamoorthi, A. Meshwin, and M. Athiq Ahmed

A Deep Learning Framework for Social Distance Monitoring and Face Mask Detection 613
 Meghana Pamarthi, Sri Latha Injam, Osman Khan Zeeshan Md., and T. Lakshmi Surekha

Information Security and Privacy in Smart Cities, Smart Agriculture, Industry 4.0, Smart Medicine, and Smart Healthcare 621
 Sanjana Prasad, Arun Samimalai, S. Rashmi Rani, B. P. Pradeep Kumar, Nayana Hegde, and Sufia Banu

Virtual Machine and Container Live Migration Algorithms for Energy Optimization of Data Centre in Cloud Environment: A Research Review 637
 Shridevi Soma and S. Rukmini

Dynamic Shortest Path Routing Algorithm to Reduce Retransmission and Congestion Avoidance for Mobile Nodes in Wireless Sensor Network 649
 S. Suma and Bharati Harsoor

A Systematic Review - Attack and Security Issues in FOG Computing 661
 C. Sabarinathan and B. Baranidharan

Examination of Water Impurities Using IoT and Machine Learning Techniques 675
 M. Pyingkodi, K. Thenmozhi, K. Nanthini, M. Karthikeyan, T. Kalpana, and P. V. Deepak

Cache Coherence for Embedded Multi-core System Architectures: A Survey and Challenges 689
 M. Thillai Rani, R. Rajkumar, K. P. Sai Pradeep, M. Jaishree, and S. TamilSelvan

Comparison of Supervised Machine Learning Algorithms for Predicting Employee Performance on Real Time Dataset 703
 Devanshu Joshi, Garima Sharma, Ankita Nainwal, and Vikas Tripathi

A Resilient and Efficient Protocol for Strengthening the Internet of Things Network Performance 715
 Salma Rattal, Isabelle Lajoie, Omar Sefraoui, Kamal Ghoumid, Réda Yahiaoui, and El Miloud Ar-Reyouchi

Early Identification of Crop Disease Using Deep Convolution Neural Networks 731
 J. Vakula Rani and Aishwarya Jakka

An Error Dependent Enhancement Method for Images Captured in Dense Fog 743
 Yucel Cimtay and Gokce Nur Yilmaz

3D Video QoE Based Adaptation Framework for Future Communication Networks	757
Gokce Nur Yilmaz and Yucel Cimtay	
Improved Lightweight Cryptography Authentication Based Secure Data Transmission in IoT Networks	769
S. Hariprasad and T. Deepa	
Modeling and Control of Induction Machine and Drive in the Combined Domain with New Chaotic Gorilla Troop Optimizer	781
Rahul Chaudhary and Souvik Ganguli	
rSense: A Novel Gesture-Based Human Assistive Device	793
Vijay A. Kanade	
Smart Boosted Model for Behavior-Based Malware Analysis and Detection	803
Saja Abu-Zaideh, Mohammad Abu Snober, and Qasem Abu Al-Haija	
Quality Rating Application for Virtual Recipes Using Facial Analysis	815
C. Shyamala Kumari, Manoharan Pon Suresh, and K. Meena	
Survey of Text Document Summarization Based on Ensemble Topic Vector Clustering Model	831
G. Bharathi Mohan and R. Prasanna Kumar	
A Hybrid Approach on Conditional GAN for Portfolio Analysis	849
Jun Lu and Danny Ding	
An Innovative Solar Power Can Satellite Model Prototype to Perceive the Environmental Data	869
Md. Nahidul Alam, Md. Zahid Hasan Buiyan, Md. Abdullah Al Hasan Anik, Atikur Rahman, and Nazifa Tahsin Lamisa	
Artificial Neural Network Based Fault Diagnosing System	885
M. Brindha, P. Nabisal Afrine, R. Priyadarshini, and P. S. Manoharan	
Smart Application for Voice Over Control on Electronic Devices Using NodeMCU	897
S. Florence, Lakshmi Narayanan, and K. Meena	
Author Index	905

About the Editors

Professor P. P. Joby is Professor and Head of Computer Science Engineering Department at St. Joseph’s College of Engineering and Technology, Palai, Kerala, India. He completed his Doctorate in Information and Communication Engineering expertise in the field of wireless sensor networks. He completed M.Tech. in Advanced Computing from Sastra University and B.E. in Computer Science and Engineering. He has many international and national publications. He is Active Member in professional bodies such as ISTE, IAENG, UACEE, and IACSIT.

Dr. Valentina E. Balas is currently Full Professor at “Aurel Vlaicu” University of Arad, Romania. She is Author of more than 300 research papers. Her research interests are in intelligent systems, fuzzy control, and soft computing. She is Editor-in-Chief to International Journal of Advanced Intelligence Paradigms (IJAIP) and to IJCSE. Dr. Balas is Member of EUSFLAT, ACM, and a SM IEEE, Member in TC-EC and TC-FS (IEEE CIS), TC-SC (IEEE SMCS), and Joint Secretary FIM.

Professor Ram Palanisamy is Professor of Enterprise Systems in the Business Administration Department at the Gerald Schwartz School of Business, St. Francis Xavier University. Dr. Palanisamy teaches courses on foundations of business information technology, enterprise systems using SAP, systems analysis and design, SAP implementation, database management systems, and electronic business (mobile commerce). Before joining StFX, he taught courses in management at the Wayne State University (Detroit, USA), Universiti Telekom (Malaysia), and National Institute of Technology (NITT), Deemed University, India. His research interest is on enterprise systems (ES) implementation; ES acquisition; ES flexibility, ES success; knowledge management systems; and healthcare inter-professional collaboration.

LabVIEW Based Anomaly Detection for Screening Diabetic Retinopathy



Sheena Christabel Pravin, K. Sindhu Priya, S. Suganthi, J. Saranya,
and V. S. Selva Kumar

Abstract Diabetic Retinopathy (DR) affects significant amount of people with acute diabetes. DR is initially asymptomatic; however, if left untreated, it can lead to low vision and visual impairment. The CNN-based EfficientNet pre-trained deep learning model is proposed in this research work to speed up the detection of DR severity levels. The retinal images were used to train the model and the images were divided into five categories viz. healthy, mild, moderate, severe, proliferate DR based on the severity level of DR. On the test set, the proposed EfficientNet achieved a classification accuracy of 0.96, while on the training set, it achieved a classification accuracy of 0.99. The proposed system has been built and evaluated on python. Also, a graphical programming environment software LabVIEW is ventured for quick and accurate detection of DR to save patient from vision loss.

Keywords Diabetic retinopathy · Deep learning · Efficientnet · Severity levels · Labview

S. C. Pravin (✉)

School of Electronics Engineering (SENSE), Vellore Institute of Technology, Chennai, Tamil Nadu, India

e-mail: sheenachristabel.p@vit.ac.in

K. S. Priya · S. Suganthi · J. Saranya · V. S. S. Kumar

Department of Electronics and Communication, Rajalakshmi Engineering College, Chennai, India

e-mail: 200811003@rajalakshmi.edu.in

S. Suganthi

e-mail: suganthi.s@rajalakshmi.edu.in

J. Saranya

e-mail: saranya.j@rajalakshmi.edu.in

V. S. S. Kumar

e-mail: selvakumar.vs@rajalakshmi.edu.in

1 Introduction

DR seems to be so crucial, and it has the potential to cause lifelong blindness. Around 415 million people across the globe are impacted with diabetes, with the figure projected to rise over 642 million by 2040. Approximately 80% of diabetic patients will experience diabetic retinopathy at some point in their lives [1]. Diabetes impacts 537 million people all over the world today. In India alone, there are over 74 million instances of diabetes, with scientists expecting a significant increase over the next decade. India is on track to become the hub of the diabetes pandemic, especially with the development of bad eating habits, obesity, and a sedentary lifestyle [2]. According to the INDIAB study, the city prevalence ranges from 10.9 to 14.2% and the rural prevalence varies from 3 to 7.8% among people aged 20 and above, with a substantially greater prevalence among people over 50 [3]. Diabetes was found to be most common in people aged 70 to 80 years old, accounting for 13.2% of the population. Association of DR with respect to duration of diabetes is shown in Fig. 1.

Diabetic retinopathy occurs when the tiny blood vessels in the retina, a region of the eye, are damaged. In Non-Proliferative DR, blood vessels get damaged and begin to leak in the retina, resulting in inadequate blood flow. The acute phase of diabetic eye disorder is known as Proliferative DR. It occurs when the retina begins to develop new blood vessels. These new blood vessels have the potential to create scar tissue. Scar tissue can wreak havoc on the macula or lead to a detached retina. Controlling blood sugar and eating a balanced diet are two possible therapies. Medication, laser surgery, and vitrectomy are all options to treat DR. Despite the alarming statistics, contemporary research suggests that at least 90% of new cases could be prevented if diabetic patients have their eyes adequately treated and monitored continually [5]. The duration of diabetes has a substantial relationship with the occurrence of retinopathy. Almost all people with type 1-diabetes and more than 60% of people with type-2 diabetes have a nominal degree of blindness after 20 years of diabetes. DR is projected as the leading source of blindness in the 20–74 age-groups [6].

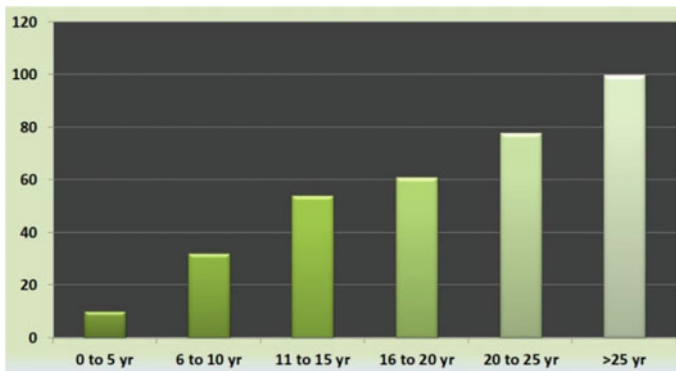


Fig. 1 Frequency of DR in patients with diabetes [4]

2 Related Work

DR seems to be more common in people who have had diabetes for a long period of time. The proposed framework is able to identify the region of hard exudate in the blood vessels by implementing a dense deep feature extraction model. The CCN-based deep learning classification is used to do early detection in a very productive and sustainable manner. For the detection of DR severity level from an input retinal image, the suggested algorithm's performance metrics were compared against Naive Bayes, SVM, and with other image processing techniques [7]. The purpose of this chapter is to reveal various digital image processing algorithms using LabView and the IMAQ vision toolbox. IMAQ vision toolbox provides a comprehensive set of digital image processing and acquisition functions that provide better performance, reduces the user computational burden, resulting in improved results in less time [8]. The proposed IOT-based computer vision system advises patients to report their health condition to the health - care server by message or email in order to categorise their state of health as emergency or serious. The suggested system is tested and analysed using several computer vision algorithms such as FRR, MAS, AWS, and ODA [9]. This study proposes an IOT based biotelemetry system for intelligent monitoring by evaluating a patient's ECG data. The data is categorised using an ANN to produce a clinical conclusion that assists the clinician in providing prompt treatment. The suggested telemetry system is compared to the standard golden technique of ECG analysis, and it is found that the proposed model is superior to the traditional model in terms of accuracy and recall [10]. The fundamental objective of this research is to create an updated pre-trained model, such as a dense convolutional neural network to classify DR severity stages. The proposed model yields performance features like 99.8% accuracy, 98.3% sensitivity, and 98.9% F measure [11]. Much research is being done in order to diagnose this DR condition early. A pre-trained model is a novel modelling approach that is progressively being deployed for DR detection and has the potential to enhance CNN's performance. Researchers employing EfficientNet-B5, a pre-trained model, in this study, got reasonably decent accuracy of approximately 94% [12]. Image processing is a technique for enhancing or extracting useful information from an image by performing related operations on it, beginning with an image and terminating also with an image or image-related features. Using the LabView vision module, researchers created and implemented an image process-based application. Image capture, colour transformation, edge recognition, and morphological operations are all performed using the vision module [13]. Tonsillitis detection module is designed in LabView and then implemented on National Instrument sbRIO 9631 FPGA kit. The Sobel operator is utilised, and the log-Gabor filter is used for optimal ridges and quick processing. It is employed in a variety of applications, including medical image processing and objects recognition. The main goal is to use the FPGA platform to process the image and use it in numerous applications [14].

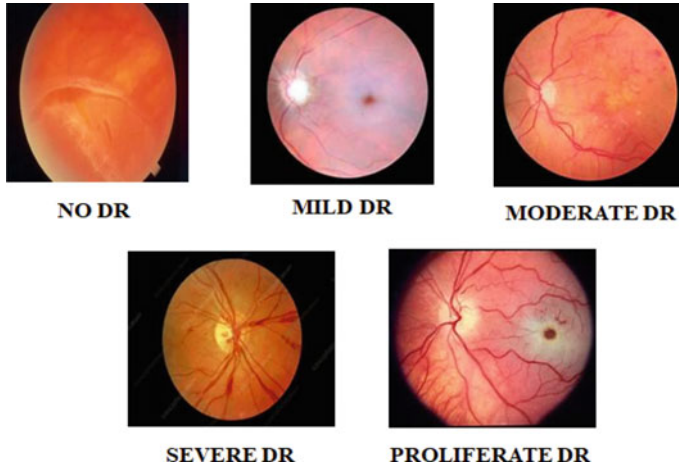


Fig. 2 Classification of dataset

The trained EfficientNet based deep learning model will be deployed as a DR detection and assessment tool to the Clinicians at a subsidized cost for rapid detection of DR. The Clinicians will be trained to use the tool eventually.

3 Dataset Description

The dataset described as taken from Kaggle is the Diabetic Retinopathy Resized dataset [15]. It was used to train the proposed EfficientNet model for screening Diabetic Retinopathy. It contains 35,126 retinal images as training set and 35,108 images as test set, recorded from the population of world. The data collection comprises of high-quality eye photographs that have been rated by qualified specialists into 5 categories (0–4) is shown in Fig. 2.

4 Data Pre-processing

Histogram Equalisation is the image processing technique used to boost image visibility in this research work. It is one of the most widely used image processing techniques for improving image visibility and quality [16]. This is accomplished by boosting the dynamic range of the target image's histogram. The HE converts the input image's grey levels into uniform grey levels in the output image. As a result, the grey levels in the final image are evenly distributed. Henceforth it is conceivable to say that the HE is used to create a uniform histogram. For each pixel, HE eventually offers a new intensity value based on its prior intensity level. The image quality

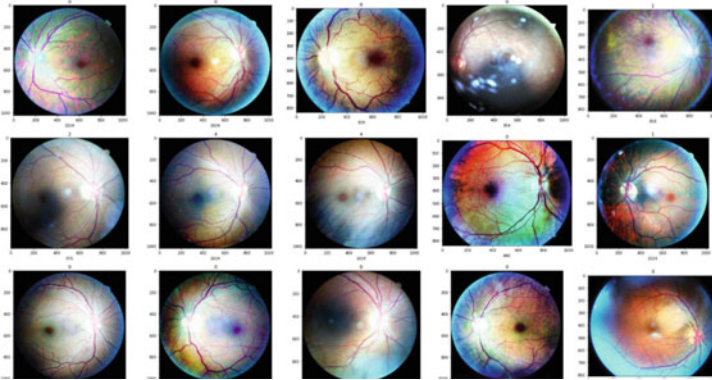


Fig. 3 Histogram equalisation

is increased and the histogram is improved because the histogram for low-contrast images is thin and centred toward the middle of the grey scale. Figure 3 represents the histogram equalisation of an image.

In the Diabetic Retinopathy, Ben Graham [17] developed a pre-processing technique in which he eliminated the local average colour from the image. For picture improvement, two-step pre-processing is used in this study, including filtering that does both scaling and circular cropping with the fundus image. Figure 4 represents the Ben Graham pre-processing method of an image.

Random flipping, random rotation, and random zoom were used as an augmentation method to enhance the effective train dataset size. Figure 5 illustrates the augmented data image.

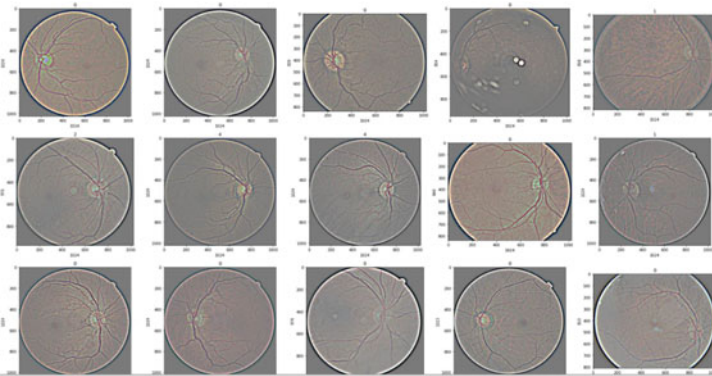


Fig. 4 Ben Graham's pre-processing

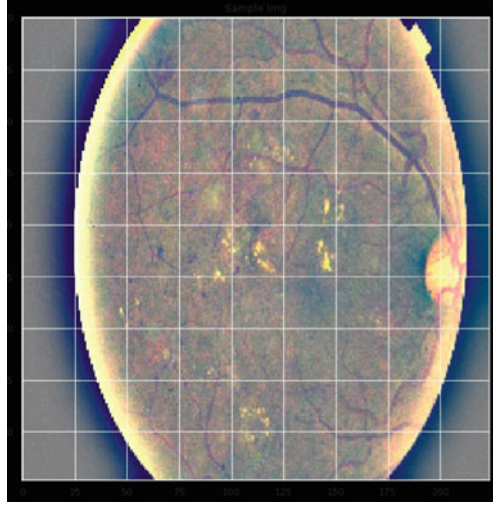


Fig. 5 Image augmentation

5 EfficientNet Model Based DR Classification

The process flow of DR severity level detection using EfficientNet is depicted in Fig. 6. The images are collected from numerous clinics over a significant period of time using fundus cameras, which will bring further differences. By evaluating data, the pictures are categorised into five levels, with 0–4 indicating the five stages of DR. The dataset’s retinal images were pre-processed using the auto-cropping technique of histogram equalisation, Ben Graham’s pre-processing and data augmentation. After pre-processing, the differentiating characteristics in images were enhanced, allowing for better training of the EfficientNet models. Finally, the potential effectiveness of a system is analysed using LabVIEW and the results produced are impressive in automated feature learning approach.

5.1 Architecture of EfficientNet Model

The proposed EfficientNet framework for multi-class classification has been built with the following hyper-parameters: the baseline filter size was set to 32, and the filter size was raised by a factor of two at each block as shown in Fig. 7. Batch normalisation was used with dropout at the hierarchical levels in each block. Furthermore, the LeakyReLU activation function was used instead of the normal ReLU activation function to minimize the declining gradient effect [18, 19]. The number of neurons in the fully connected layers FC1, FC2, and FC3 was set to 2048, 512, and 128 consecutively for feature compression. The Max pooling layer, which is commonly

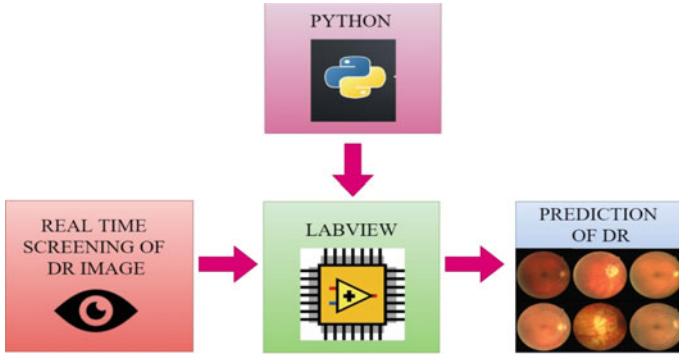


Fig. 6 Process flow for EfficientNet model based DR classification

added to EfficientNet models, decreases the dimensionality of the image by reducing the levels of pixels in the preceding layer’s output. After the max pooling layer, a dropout layer is introduced to improve the performance of the proposed EfficientNet model by removing data overfitting. The sigmoid layer was introduced at the output layer with 5 class probabilities instead of softmax layer. As a result, the proposed model EfficientNet minimizes the overall computational complexity, maintaining higher classification accuracy.

Figures 8 and 9 depicts the graphic representation of the EfficientNet’s accuracy and loss in relation to the number of epochs. The blue colour legend characterizes the classification accuracy and loss over the testing data, while the orange colour legend depicts the classification accuracy and loss measured over training dataset. The training and testing sets measures were calculated to be 0.99 and 0.96 respectively.

The other performance metrics such as confusion matrix, model loss, precision, recall and F1-score were also measured to evaluate the proposed model against the baseline models. The confusion matrix is plotted in Fig. 10.

A comparative study was undertaken to relate the proposed model’s performance with the contemporary models in the literature. The performance summary is displayed in Table 1.

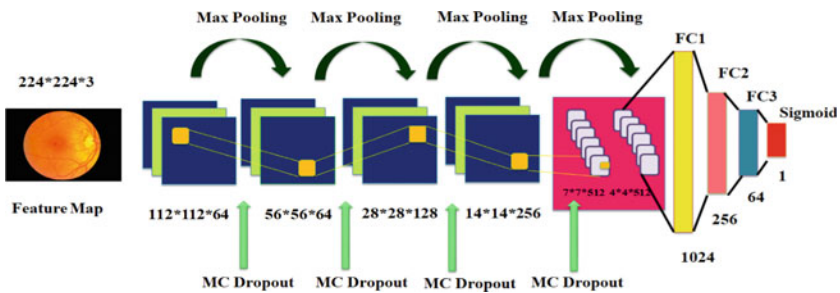


Fig. 7 Block diagram for EfficientNet

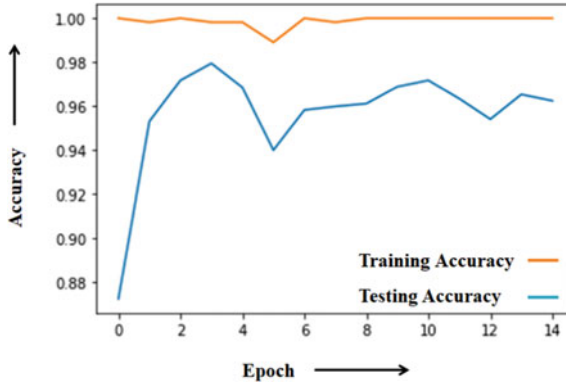


Fig. 8 Accuracy model for EfficientNet

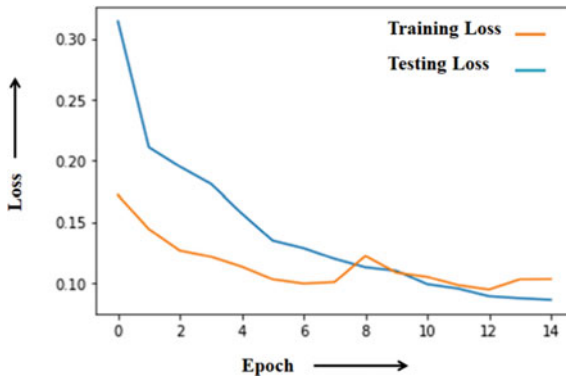


Fig. 9 Loss model for EfficientNet

6 Real Time DR Diagnosis

Real time DR diagnosis was experimented over LabVIEW, which is a graphical programming environment that lets users to develop and evaluate any complex system in less time than a text-based programming environment. The graphical applications in LabVIEW are referred to as Virtual Instruments (VI). The control palette is accessible only on front panel and contains the VI functions to build the block diagram. The block or graphical component is performed when data is available at all inputs. Following the completion of the execution, the data is delivered to output terminals and then sent to the next block in the dataflow route.

Figure 11 illustrates the process flow to run a python script in LabVIEW using python 3.6. Initially select a string and path from the control palette on the front panel. Select file input output from the function palette and then interface read to text file block in the block diagram. Path control will feature a window on the front

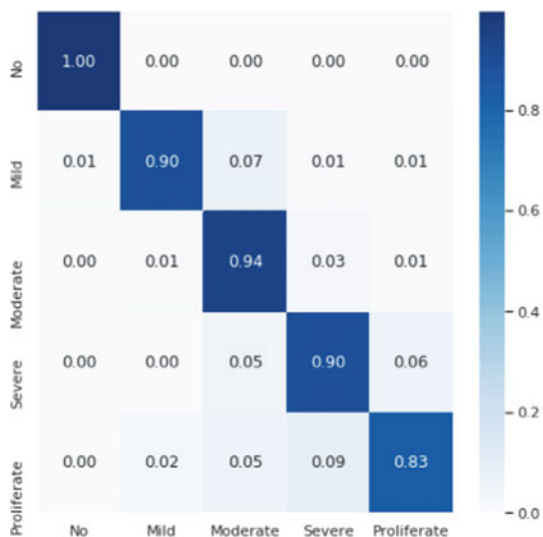


Fig. 10 Confusion matrix for EfficientNet

Table 1 Characteristics of included studies

Author	Year	Dataset	Model	Accuracy	Recall	Specificity
Sheena Pravin et al. (Proposed)	2022	Kaggle	EfficientNet	96	93	94
Mirza Mohd Shahriar et al. [12]	2021	Kaggle	EfficientNet B5	94	83	92
Sungeetha et al. [7]	2021	Kaggle	DeepFeature CNN	97	92	92
U. Sathish et al. [5]	2020	Kaggle	CNN	75	30	95
Nitin ShivSharan et al. [11]	2021	Kaggle	DenseNet-BC	99	98	–

panel and a browsing button will allow user to browse the file on your computer in the window. Finally running this VI opens the python script.

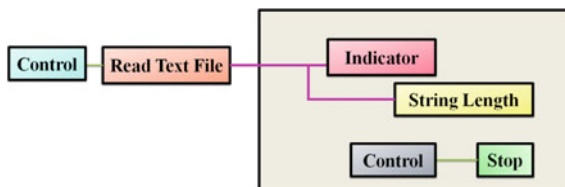


Fig. 11 Accessing python script in LabView

The average pixel value at each grayscale level is counted and graphed via a histogram is shown in Fig. 13. A histogram is a visual representation of an image that aids in the identification of numerous elements namely the background, objects, and noise. Figure 12 depicts user to enumerate region from the retinal image to calculate a histogram.

Anomalies are detected from the retinal images using an image processing tool kit. Testing is done on real-time pictures obtained with a fundus camera. Figure 15 illustrates given DR image is proliferative by its diagnosis value 4. The suggested system may be used as a stand-alone tool for DR analysis via retinal image processing. The blocks for the detection of anomaly are depicted in Fig. 14.

The process of acquiring an image stream is simple and can be broken down into three parts. The first step is to interface the camera using image acquisition toolbox and set up the exact data containers for the images. A loop is used to take a single frames from the camera, and finally, camera is closed and cleaned

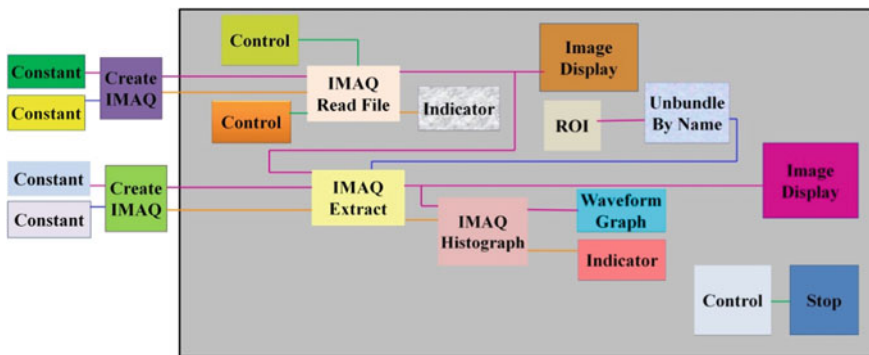


Fig. 12 IMAQ histogram

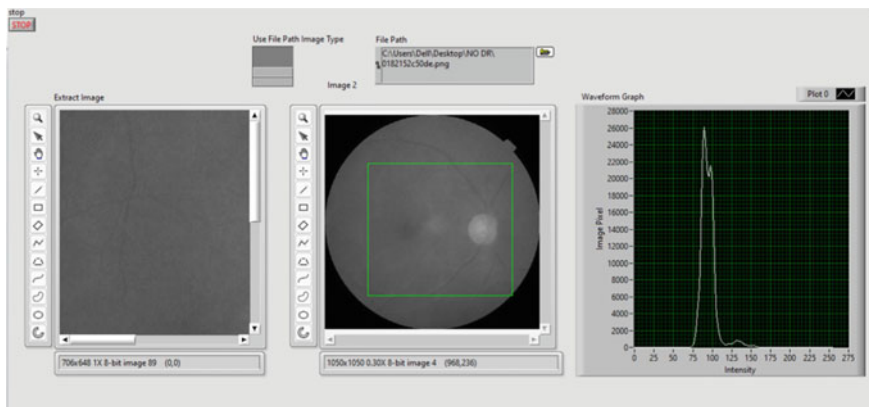


Fig. 13 Image histogram in LabView

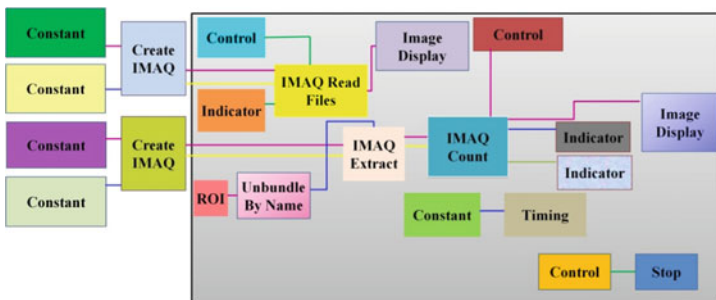


Fig. 14 Diagnosis of DR image

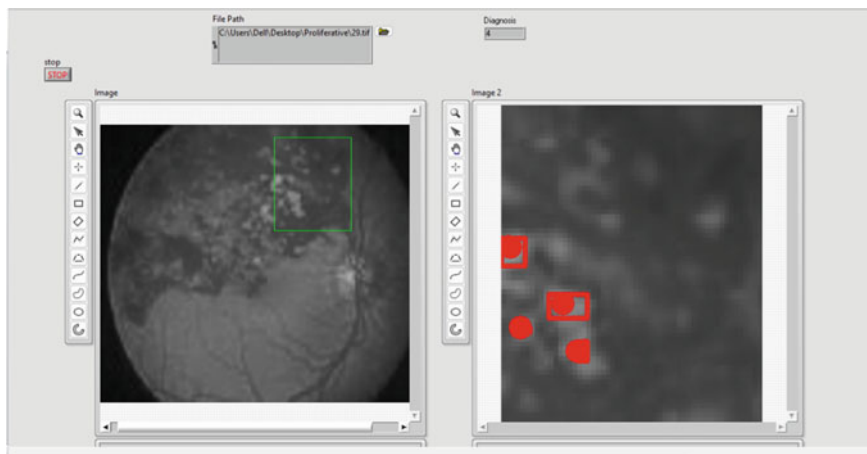


Fig. 15 Extracting diagnosis value with its ROI in front panel

for further processing. Figure 17 illustrate if the patient is impacted by DR, Region of Interest (ROI) enumerate the region in the retinal image that contains anomalies using LabVIEW Ngene deep learning toolkit and the round LED tool flashes green for moderate DR, yellow for no DR and red for proliferative DR. Figure 16 depicts the block diagram to diagnose DR.

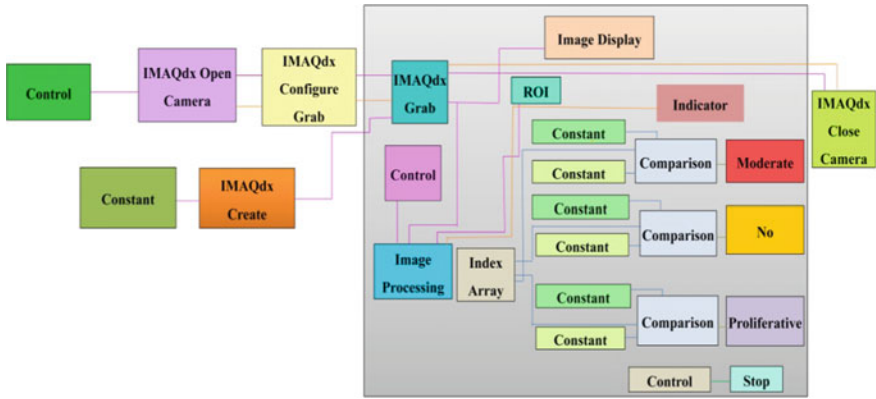


Fig. 16 Real time screening of DR image

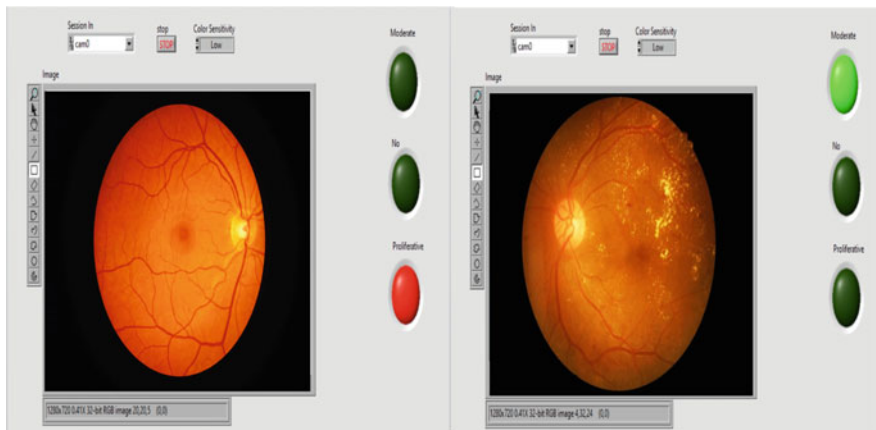


Fig. 17 Diagnosis of moderate and proliferative DR in front panel

7 Conclusion

In this research work, current tools such as LabView, Python 3.6 are employed to diagnose severity level of diabetic retinopathy. The software-implemented approach may be used to aid ophthalmologists in the early diagnosis of DR. The system’s performance is tested for the publicly accessible database Kaggle, and its correctness is validated by computing its accuracy and loss value. In future, experimental studies on multiple retinal datasets would be considered to estimate the specific outcomes. Further, the size of the training and test datasets would be varied to see how it adversely affected the model’s prediction accuracy. An investigation on the severity level prediction of other diseases such as glaucoma and macular degeneration would be thoughtfully experimented with the proposed model.

References

1. Bommer C, Sagalova V, Heesemann E, Manne-Goehler J, Atun R, Barnighausen T, Davies J, Vollmer S (2018) Global economic burden of diabetes in adults: projections from 2015 to 2030. *Diab Care* 41(5):963–970. <https://doi.org/10.2337/dc17-1962>, Epub 2018 Feb 23, PMID: 29475843
2. Diabetic Retinopathy is on the Rise in Young People. Here's how you can control it!. <https://www.firstpost.com/health/diabetic-retinopathy-is-on-the-rise-in-young-people-heres-how-you-can-control-it-10231861.html>
3. Government Survey Found 11.8% Prevalence of Diabetes in India, <https://www.livemint.com/science/health/government-survey-found-11-8-prevalence-of-diabetes-in-india-11570702665713.html>
4. Gupta RP, Kotecha M, Bansal P (2013) Frequency of diabetic retinopathy in patients with diabetes mellitus and its correlation with duration of diabetes mellitus. *Med J Dr. D.Y. Patil Univ* 6(4):366
5. Sathish U, Abdul SK, Rishindray V, Mukhesh SH (2020) Detection of diabetic retinopathy using CNN. *IRE J* 3(11). ISSN: 2456–8880
6. Schwartz SG, Mieler WF (2013) Retinal and choroidal manifestations of systemic medications. In: Arevalo J (eds) *Retinal and Choroidal manifestations of selected systemic diseases*. Springer, New York
7. Sungeetha A, Sharma R (2021) Design an early detection and classification for diabetic retinopathy by deep feature extraction based convolution neural network. *J Trends Comput Sci Smart Technol* 3(2):81–94
8. Posada-Gomez R, Osvaldo O, Martinez A, Portillo-Rodriguez O, Alor-Hernandez G (2011) Digital image processing using LabView. In: *Practical applications and solutions using LabView* ™
9. Balasubramaniam V (2022) IoT based biotelemetry for smart health care monitoring system. *J Inf Technol Digital World* 02(03):183–190
10. Sathesh A (2020) Computer vision on IOT based patient preference management system. *J Trends Comput Sci Smart Technol* 02(02):68–77
11. Shivsharan N, Ganorkar S (2021) implementation of the modified pre-trained DenseNet model for the classification of grades of the diabetic retinopathy. In: *Inventive communication and computational technologies. Lecture notes in networks and systems book series, LNNS*, vol 311
12. Maswood MMS, Hussain T, Khan MB, Islam MT, Alharbi AG (2020) CNN based detection of the severity of diabetic retinopathy from the fundus photography using EfficientNet-B5. In: *11th IEEE annual information technology, electronics and mobile communication conference*. <https://doi.org/10.1109/iemcon51383.2020.9284944>
13. Nelutla A (2018) Image processing techniques using LabView. *Int J Latest Technol Eng Manag Appl Sci* 7(8). ISSN 2278–2540
14. Kumbhalwar V, Dixit S (2016) Labview design for edge detection using log gabor filter for disease detection. *Int J Res Dev Technol* 5(5). ISSN (O) 2349–3585
15. DiabeticRetinopathy (Resized). <https://www.kaggle.com/datasets/tanlikesmath/diabetic-retinopathy-resized>
16. Shakya S, Joby PP (2021) Heart disease prediction using fog computing based wireless body sensor networks (WSNs). *IRO J Sustain Wirel Syst* 3(1):49–58
17. Pravin SC, Palanivelan M (2021) A hybrid deep ensemble for speech disfluency classification. *Circ Syst Signal Process* 40(8):3968–3995

18. Pravin SC, Palanivelan M (2021) Regularized deep LSTM autoencoder for phonological deviation assessment. *Int J Pattern Recogn Artif Intell* 35(4):2152002
19. Pravin SC, Palanivelan M (2021) Acousto-prosodic delineation and classification of speech disfluencies in bilingual children. In: *Proceedings of the 12th international conference on soft computing and pattern recognition (SoCPaR 2020)*. SoCPaR 2020. *Advances in intelligent systems and computing*, vol 1383. Springer, Cham. https://doi.org/10.1007/978-3-030-73689-7_59

Early Stage Parkinson's Disease Diagnosis and Detection Using K-Nearest Neighbors Algorithm



S. Svetaa, S. Pavithra, and R. G. Sakthivelan

Abstract This research article proposes a novel method to detect and recognize the symptoms of Parkinson's disease in its early stages. Parkinson's disease generally refers to a neurological and neurodegenerative movement disorder, which affects millions of people all over the world. Loss of automatic movements and rigidity of muscles are common symptoms, which eventually lead to difficulty in walking, balancing, and coordination disorders. Over time, these symptoms deteriorate the patient's physical, emotional, and mental health. There is currently no way to completely cure the Parkinson's disease. However, a therapeutic care can be provided to reduce the prognosis of the disease before the patient's condition deteriorates. This necessitates the need to implement machine learning approaches for Parkinson's disease detection. The proposed machine learning algorithm will be trained by utilizing the audio recordings the patient's speaking abilities. Following the training process, introducing further input signal will tend to identify and confirm the existence of Parkinson's disease by allowing the medical practitioners to take appropriate action and adopt different medical technologies and approaches to avoid and, more significantly, prevent the patient from becoming further affected. Further, the algorithms such as K-Nearest Neighbors, Logistic Regression, Naive Bayes, and Random Forest will be examined to investigate and analyze different properties and determine the appropriate algorithm that may give superior accuracy in the early identification and confirmation of Parkinson's disease.

Keywords Parkinson's disease · Vocal skills · K-Nearest Neighbors · Logistic Regression · Naive Bayes · Random Forest

S. Svetaa (✉) · S. Pavithra · R. G. Sakthivelan
Department of Data Science, Rajalakshmi Engineering College, Thandalam, Chennai 602105,
Tamil Nadu, India
e-mail: 201011005@rajalakshmi.edu.in

S. Pavithra
e-mail: pavicse06@gmail.com

R. G. Sakthivelan
e-mail: sakthivelan.rg@rajalakshmi.edu.in

1 Introduction

Parkinson's disease [5, 9] could be a hereditary, neurological, and sensory system disorder, which impairs the human ability to control the body's motions and cause tremors, rigid muscles, and impairment in bodily balance and coordination. Parkinson's infection tends to develop once the neurons (neural cells) of a nearby region, referred to as locus niger, settle within the neural structure of the human brain and begin to paralyse or injure themselves. The locus niger present within the neural network is the vital region that synthesizes a crucial neurochemical substance called Dopastat, which is used to transfer the signals (messages) between neurons' synapses to numerous areas of the body.

In Parkinson's disease (PD) is the world's second-most prevalent disorder, affecting various individuals worldwide. Custom-tailored and subject-specific treatment (prescription or profound cerebrum incitement (Deep Brain Stimulus)) is a ground-level basic diagnosis carried out for diagnosing the Parkinson's disease. However, this process is dependent on the precise evaluation of cardinal Parkinson's disease symptoms such as bradykinesia (slowness of movement)-one of the cardinal signs of Parkinson's disease, inflexibility of body joints, and body tremors [3, 7, 8].

A typical Parkinson's disease evaluation will be carried out by using a semiconductor diode on a series of patients. If the patients are in remote locations, doctors will use mobile phones (also known as telemedicine or tele-treatment, etc.) to determine whether they can acquire objective conduct data semi-consistently (on and off method), track the fluctuations of the disease in their body, and avoid relying on others.

Several doctors utilize the Hoehn and Yahr scale to identify different stages of Parkinson's disease. This scale categorizes the symptoms into five stages and assists healthcare providers in determining the severity of disease signs and symptoms [6, 12].

2 Related Work

Kenneth A. Loparo et al. presented a method [11] to avoid overstimulation, which might be harmful to patients. The PD state is represented by distorted thalamic relay reliability, and the network model consists of connected biophysically driven spiking neurons called Tyrosine Hydroxylase (TH). The results show that the TH's distorted relay reliability can be efficiently restored while consuming less DBS energy using an RL-based closed-loop control approach. The PD state is represented as distorted thalamic relay reliability, and the network model is made out of linked biophysically spiking neurons.

Tobias Piroth et al. developed a system [10] for recording temporary DBS-induced modifications, as well as a technique for evaluating fine motor function. Its underlying machine learning model is interpretable, displaying the relative importance of

Table 1 Dataset information

Dataset type	Dataset characteristic	Attribute characteristics	Attributes	Split	
				Train	Test
Parkinson disease	Multivariate	Integer, Real	924	739	185

different features. Single DBS-sensitive movement components may be identified and described by using this technique. The proposed assessment strategy will help in the advancement of data-driven discovery in PD-based neural biomarkers. As a result, the source code for the paradigm is now made open source.

According to John Prince et al. [13], the use of multi-channel CNNs and the development of models using a large cohort of participants were partially responsible for the increase in accuracy.

3 Dataset Description

For the proposed study, the Parkinson disease dataset was obtained from the UCI Repository[14]. The database consists of 924 samples. The information for carrying out this research study was gathered from 708 people with Parkinson's disease (509 men and 199 women), aged in the range of 36 to 93 in Fig. 4. Several speech signal processing methods are used to determine different attributes. On this dataset, the supervised learning algorithms were used to categorize the records into two categories: Parkinson affected individuals or non-affected individuals, as indicated by the values of class label 1 or 0. Table 1 displays the attribute data. Table 2 illustrates the characteristics of Parkinson's illness [3].

4 Data Pre-processing

To increase detection capabilities, the data has been pre-processed to eliminate unrefined medical data by eliminating missing values. Further, the data modifications and manipulation processes are carried out to make the data useful for performing analytics with the help of machine learning models.

During data pre-processing, null values are validated and filled using two methods: random sampling for a greater number of null values present and a mean and mode sampling strategy for a lesser number of null values present. After dealing with null values, this study proceeds to incorporate categorical feature encoding.

Table 2 Dataset description table

S. no	Categories	Features	dtype
1	Vocal fundamental frequency	Avg: MDVP:Fo(Hz), Max: MDVP:Fhi(Hz), Min: MDVP:Flo(HZ)	Float64
2	Frequency parameters	MDVP:Jitter(%), MDVP:Jitter(Abs), MDVP:RAP, MDVP:PPQ, Jitter:DDP	Float64
3	Amplitude parameters	MDVP:Shimmer, MDVP:Shimmer(db), Shimmer:APQ3, Shimmer:APQ5, MDVP:APQ, Shimmer:DDA	Float64
4	Harmonicity parameters	Noise-to-HarmonicRatio (NHR), Harmonic-to-Noise Ratio (HNR)	Float64
5	Other parameters	Recurrence Period Density Entropy (RPDE), Dentrended Fluctuation Analysis (DFA), D2 - Correlation Dimension, Spread 1,2 - Non Linear measure of fundamental frequency, Pitch Period Entropy (PPE)	Float64

5 Data Exploration

This stage is used to visually explore the dataset in order to comprehend numerous biological parameters. As depicted in Fig. 1, the dataset includes 670 male patients and 254 female patients, wherein 509 male patients and 199 female patients are affected individuals.

In the considered dataset, people aged in between 66 and 73 had the highest frequency of Parkinson's disease within the 36–93 age range, as shown in Fig. 2.

6 Model Building

6.1 Training and Testing

The dataset has been divided into 80% (739 samples) for training and 20% (185 samples) for testing.

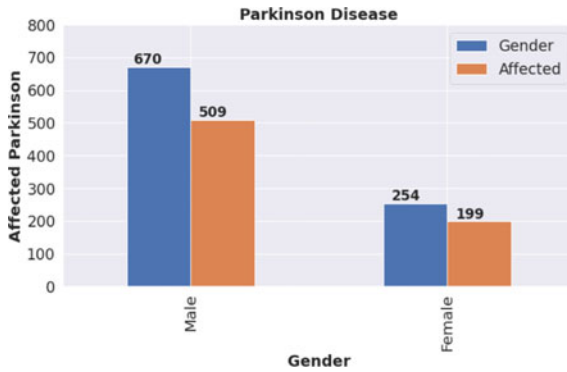


Fig. 1 Affected by Parkinson’s disease in gender

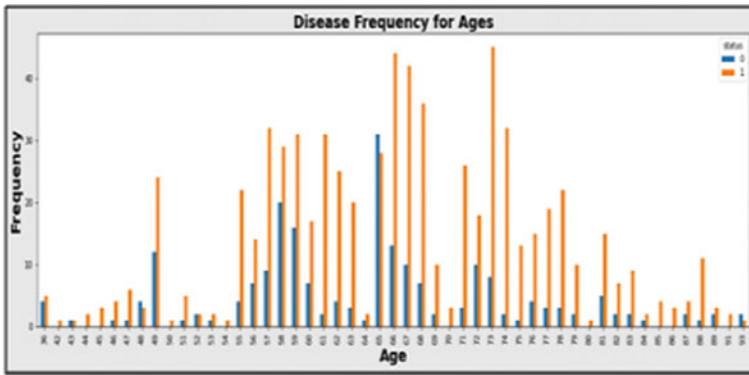


Fig. 2 Frequency histogram of the dataset

6.2 Classification

The supervised machine learning algorithms employed in this study are mentioned in this section. The pre-processed datasets will be available once the victimisation strategies for all of the feature choice strategies are implemented. These datasets are used to train and test different machine learning techniques. Training enables the techniques to obtain a comprehensive and distinct collection of knowledge from the input dataset. Figure 3 shows the use of K-Nearest Neighbors (KNN), logistic regression, Naive Bayes, and the random forest square measure.

K-Nearest Neighbors (KNN)

The K-Nearest Neighbor or KNN algorithmic program generates a limit for knowledge categorization. When fresh information points become available, the algorithmic program might attempt to anticipate their proximity to the nearest boundary. As a consequence, a higher k-value indicates the power tool curves of separation

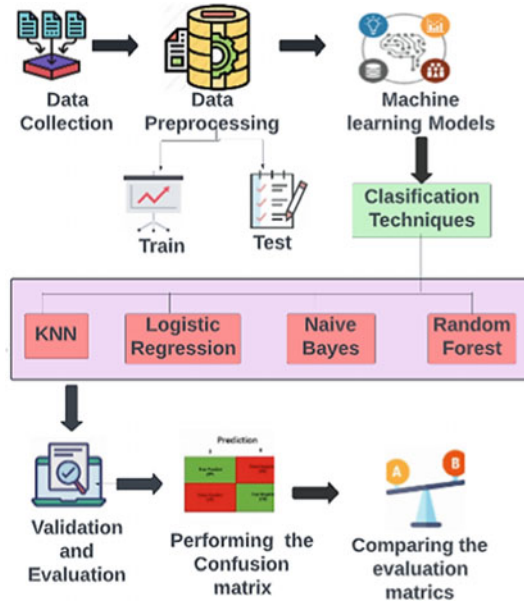


Fig. 3 Proposed methodology

		Actual Values	
		0	1
Predicted Values	0	True Negative (TN)	False Positive (FP)
	1	False Negative (FN)	True Positive (TP)

Fig. 4 Confusion matrix

result in fewer intricate models, whereas a lower k-value tends to overfit the information, resulting in more sophisticated models. It is critical to understand the right k-value when analysing a dataset to avoid overfitting and underfitting challenges. For a long term forecasting, the model will use the previous data (or train the model).

When compared to [4], the proposed study has extended the dataset and received a training data accuracy of 96% and a testing data accuracy of 97%.

Logistic Regression

The purpose of logistic regression is to discover a relationship between a set of characteristics and the chance of an event occurrence.

For example, if we have the number of hours spent studying as a characteristic and need to predict whether a student will pass or fail a test, the response variable has two values: pass and fail.

In Binomial Logistic Regression, the response variable has two values: 0 and 1, or true and false. Multinomial logistic regression is used when the response variable includes three or more possible values.

Naive Bayes

The Naive Bayes classifier is a probabilistic algorithm that computes a set of possibilities by considering the frequency and groupings of values present in a given record.

Random Forest

Random Forest (RF) is a type of feature selection approach, which is used to obtain the required ranks from tree-based models [1]. The feature importance is determined by taking the mean of each variable's improvement. This is frequently accomplished by sorting the attributes before selecting a set. Random Forest attributes are more likely to be related to those obtained from other classes. Furthermore, irrespective of the options' quality, if two alternatives are significantly related with every AdaBoost variation, each of their scores lower. Furthermore, the relevance of the option may be calculated by mistreating the mean by lowering impurity (Gini index) [2].

Confusion Matrix

Based on the classification results obtained from different classifier algorithms, a distinct confusion matrix is constructed to determine performance indicators such as accuracy, precision, recall, and f1 score. The classification results are displayed as a matrix in the confusion matrix mentioned in Fig. 4.

It is not essential to show that a person has Parkinson's disease. We produced the correlation matrix after performing exploratory data analysis, which correlates the data set's properties. We are building a confusion matrix using the KNN, Logistic Regression, Nave Bayes, and Random Forest technique and the testing data set. The confusion matrix yields more advanced metrics such as accuracy, precision, recall and f1 score which can aid us in making decisions throughout the classification process.

The confusion matrix for KNN, Logistic Regression, Nave Bayes, and Random Forest was covered in the following section in the Figs. 5, 6, 7 and 8 to calculate the evaluation matrix of these four supervised classification techniques.

The confusion matrix shown in Figs. 5, 6, 7 and 8 was created to support the findings of the simplest acting cross-validation set. During this work, we tended to use four completely different supervised classification algorithms to sight brain disease. As a result, many quantitative ways are utilized to gauge the classifiers' performance. For example, accuracy, precision, recall, and f1.

As a consequence, the calculating method for evaluating the concerns is as follows:

$$Accuracy = (Correct\ Predictions)/(Total\ Predictions) \quad (1)$$

$$Precision = (True\ Positive)/(True\ Positive + False\ Positive) \quad (2)$$

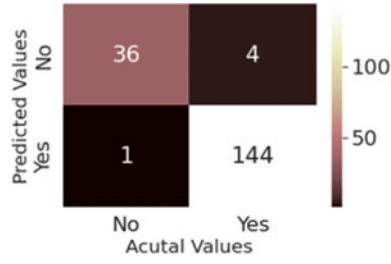


Fig. 5 The confusion matrix for KNN

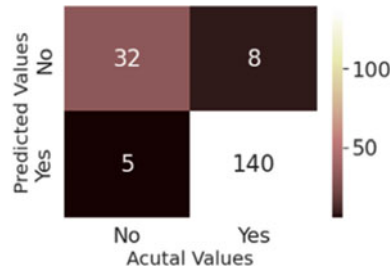


Fig. 6 The confusion matrix for Logistic Regression

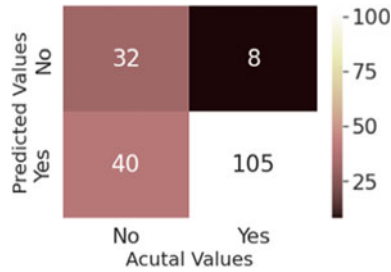


Fig. 7 The confusion matrix for Naive Bayes

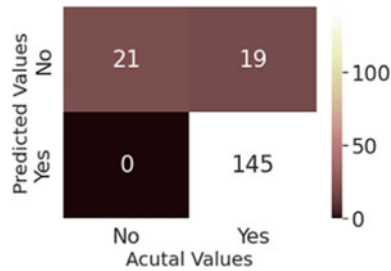


Fig. 8 The confusion matrix for Random Forest

$$Recall = (TruePositive)/(TruePositive + FalseNegative) \tag{3}$$

$$F1 - Score = 2 * (Precision + Recall)/(Precision + Recall) \tag{4}$$

We conducted various research in this part to evaluate the metric values of the four methods for identifying Parkinson’s illness. The study of four ways of exposing Parkinson’s disease data was conducted. Figure 9 demonstrates the performance of four supervised classification approaches using the information from Table 3.

“Train” and “test” datasets are split in an 8:2 ratio. The model is tested on the test dataset after it has been fitted to the training dataset to estimate its performance on new data rather than previously trained data. The confusion matrix, as well as numerous model performance measures such as accuracy, precision, recall, F1-Score, and roc, are then used to compare and contrast the four distinct supervised machine learning classifiers to find the best and most accurate model for predicting Parkinson’s disease. After that, the findings are compared to previous studies. The Confusion Matrix is the most widely used tool for predicting Parkinson’s disease in this field of research.

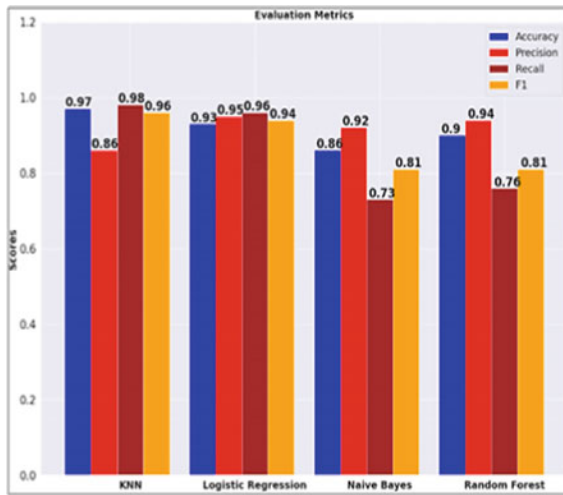


Fig. 9 Performance of four supervised classification techniques

Table 3 Classification performance measurements

Parameters	Accuracy	Precision	Recall	F1-score	Roc
KNN	0.97	0.86	0.98	0.96	0.94
Logistic regression	0.93	0.95	0.96	0.94	0.94
Naive Bayes	0.86	0.92	0.73	0.81	0.76
Random forest	0.90	0.94	0.76	0.81	0.76

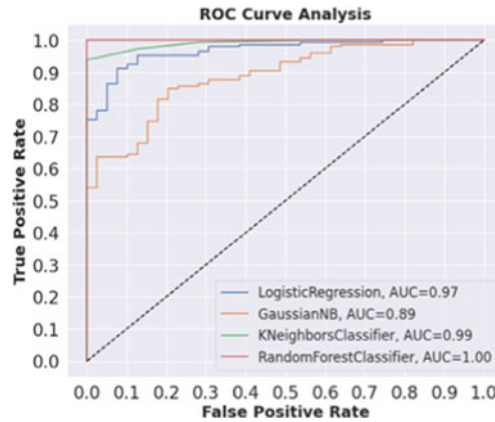


Fig. 10 Roc curve for four supervised classification techniques

It's used to figure out what the different classifiers are up to. This sort of matrix shows how well the classifier compared the correctly predicted cases to the badly predicted ones. As a consequence, a confusion matrix is a table that shows the estimated model and the dataset's actual values. It is used to evaluate model performance in machine learning classification-based issues. As illustrated in Fig. 9, it is made up of four distinct combinations of expected and actual values: True Negative, True Positive, False Positive, and False Negative. The creation of the confusion matrix is built on the foundation of these four features.

In this case, KNN beat Logistic Regression, Naive Bayes, and Random Forest in terms of accuracy in this case. The results clearly show that the KNN achieved maximum accuracy (97%). The highest level of Precision (94%) was achieved by Random Forest. The most precise method was logistic regression, which produced a recall of 95%. In terms of the f1 score and roc score (i.e., 0.94 and 0.76), Naive bayes and Random forest had the weakest result. Finally, in terms of overall performance, KNN comes out on top.

The graphical representation shows the ROC curve in the above Fig. 10 between True Positive Rate (TPR) and False Positive Rate (FPR) at various threshold levels. The AUC value generated from the ROC curve is 1.00 (100%), indicating that the model deliver 100% accurate prediction on Parkinson's disease.

7 Conclusion

In this paper, Parkinson's disease detection strategies are studied by using supervised classification approaches such as K-Nearest Neighbors (KNN), Logistic Regression, Naive Bayes, and Random Forest, and the optimal choice is determined. This also allows for selecting effective treatment in a relatively short period of time by reducing

the time required to track the Parkinson's disease in a long term. Since the determination should be achievable by only using a few approaches, there is a part of the extension to figure the innovation. To develop an outline for Parkinson's disease, the AI computations can be applied in the near future.

References

1. Chen R, Dewi C, Huang S, Caraka R (2020) Selecting critical features for data classification based on machine learning methods. *J. Big Data* 7:1–26
2. Antony L et al (2021) A comprehensive unsupervised framework for chronic kidney disease prediction. *IEEE Access* 9:126481–212650
3. Wan S, Liang Y, Zhang Y, Guizani M (2018) Deep multi-layer perceptron classifier for behavior analysis to estimate Parkinson's disease severity using smartphones. *IEEE Access* 6:36825–36833
4. Tiwari H, Shridhar SK, Patil PV, Sinchana KR, Aishwarya G (2021) Early prediction of parkinson disease using machine learning and deep learning approaches. *EasyChair Preprint* 4889:1–14
5. Cao Z, John AR, Chen H-T, Martens KE, Georgiades M, Gilat M, Nguyen HT, Lewis SJG, Lin C-T (2021) Identification of EEG dynamics during freezing of gait and voluntary stopping in patients with parkinson's disease. *IEEE Trans Neural Syst Rehabil Eng* 29:1774–1783
6. Chen OY, Lipsmeier F, Phan H, Prince J, Taylor KI, Gossens C, Lindemann M, de Vos M (2020) Building a machine-learning framework to remotely assess Parkinson's disease using smartphones. *IEEE Trans Biomed Eng* 67(12):3491–3500
7. Mikos V, Heng CH, Tay A, Yen SC, Chia NSY, Koh KML, Au WL (2019) A wearable, patient-adaptive freezing of gait detection system for biofeedback cueing in Parkinson's disease. *IEEE Trans Biomed Circ Syst* 13(3):503–515
8. Prince J, Andreotti F, De Vos M (2018) Multi-source ensemble learning for the remote prediction of Parkinson's disease in the presence of source-wise missing data. *IEEE Trans Biomed Eng* 66(5):1402–1411
9. Cao Z, John AR, Chen HT, Martens KE, Georgiades M, Gilat M, Lin CT (2021) Identification of EEG dynamics during freezing of gait and voluntary stopping in patients with Parkinson's disease. *IEEE Trans Neural Syst Rehabil Eng* 29:1774–1783
10. Castano-Candamil S, Piroth T, Reinacher P, Sajonz B, Coenen VA, Tangermann M (2019) An easy-to-use and fast assessment of patient-specific DBS-induced changes in hand motor control in Parkinson's disease. *IEEE Trans Neural Syst Rehabil Eng* 27(10):2155–2163
11. Lu M, Wei X, Che Y, Wang J, Loparo KA (2019) Application of reinforcement learning to deep brain stimulation in a computational model of Parkinson's disease. *IEEE Trans Neural Syst Rehabil Eng* 28(1):339–349
12. Mikos V, Heng CH, Tay A, Yen SC, Chia NSY, Koh KML, Au WL (2019) A wearable, patient-adaptive freezing of gait detection system for biofeedback cueing in Parkinson's disease. *IEEE Trans Biomed Circ Syst* 13(3):503–515
13. Prince J, Andreotti F, De Vos M (2018) Multi-source ensemble learning for the remote prediction of Parkinson's disease in the presence of source-wise missing data. *IEEE Trans Biomed Eng* 66(5):1402–1411
14. Index of /ml/machine-learning-databases/parkinsons/telemonitoring (uci.edu)

IoT Crypto Security Communication System



Kiran Kumar Kommineni, G. C. Madhu, Rajadurai Narayanamurthy, and Gurpreet Singh

Abstract There are several uses for the Internet of Things networks, which link many devices. Massive volumes of sensitive data are collected and processed by IoT devices. An IoT network's most important characteristic should be data security. This paper introduces IoT-Crypto, a cryptographic-enabled communication protocol for the Internet of Things. Devices' limited capabilities, the necessity to decrease data transmission capacity, and compatibility with the Internet are all addressed by this technology. The IoT-Crypto certificate format and trust architecture are based on real-world corporate relationships and are unique and lightweight. A secure communication protocol based on an encrypted DTLS connection is also specified in the document. IoT-Crypto is presented in the context of similar systems, and its unique features and implementation details are discussed. The IoT-Crypto network's test and experiment results show that it is safe and reliable. For decoding BLE IPSE standards profiles, the IoT is used as a text network; the feature work is based on the same issue.

Keywords IoT · Security system · Crypto network · BLE IPSE standard profile

K. K. Kommineni (✉)

Chalapathi Institute of Engineering and Technology, Lam, Guntur, Andhra Pradesh, India
e-mail: kommineni.kiran11@gmail.com

G. C. Madhu

Department of ECE, Sree Vidyanikethan Engineering College (Autonomous),
Tirupati, Andhra Pradesh, India

R. Narayanamurthy

Swiss School of Business Management (SSBM), Geneva, Switzerland

G. Singh

University Institute of Computing, Chandigarh University, Mohali, Punjab, India

1 Introduction

The Internet of Things (IoT) integrates physical objects that can function properly and communicate independently to optimize and enable new services [1]. IoT links numerous gadgets to the internet [2]. The architecture of the underlying network determines how IoT devices perform, operate, and connect.

IoT technologies include M2M, WSN, D2D, medium-to-low-energy wireless LAN, and RFID. Due to the automated integration of applications into the Internet of Things, security risks may occur. IoT devices and networks need a security and serviceability plan. Unprotected IoT infrastructure is subjected to criminal behaviour [5], and security weaknesses including jamming, replay, DDoS, malware propagation and penetration [6], sinkhole attack [7], mischievous sequence [8], sensor assault [9], sensor assault and routing episode.

Recent improvements in communication technologies and device manufacturing are projected to enhance resource management and ubiquitous sensing in smart cities, health care, intelligent power grids, industrial automation, and intelligent spaces [7]. It will revolutionize human life quality, but it will also bring new security challenges such as access control, configuration, privacy and data management/storage. [8]. IoT challenges include privacy and security. The Internet of Things is heterogeneous, which increases the risk of security issues. Authorization and authentication systems may solve security and privacy problems in resource-constrained IoT. Due to the fast proliferation of intelligent IoT devices, customers are worried about security and privacy. IoT devices cause cyberattacks on networked systems. These IoT devices are exposed to attacks and threats.

Cybercriminals are developing new methods to steal sensitive data from IoT-based applications, including industrial, agricultural, smart home, healthcare, transportation, and bright house. Lack of standardization, acceptance of approved procedures, and limited hardware and software security resources create loopholes in IoT networks. Large devices and heterogeneity in complicated networks make connecting IoT devices more difficult.

Cyberattacks and intrusions cause IoT infrastructure breakdowns, and smart-device failures affect system performance. New risk disclosure will affect IoT device performance and network management.

IoT networks are less reliable and more vulnerable to attacks than traditional systems. Securing IoT networks involves several difficulties.

One such IoT solution is IoT-Crypto. Custom-built mechanics and features make up the bulk of its design. They are designed to address several operational issues previously discussed concerning IoT networks.

As shown in Fig. 1, the IoT-Crypto system's overall design is based on existing solutions disclosed in scientific literature and is commercially accessible. The three-tiered architecture is used by several of them. Because of its adaptability and scalability, the design has found use in various fields. In addition, this makes it easier to understand and visualise the network's topology. Thus, the IoT-Crypto network has three layers: cloud, gateway and IoT devices.

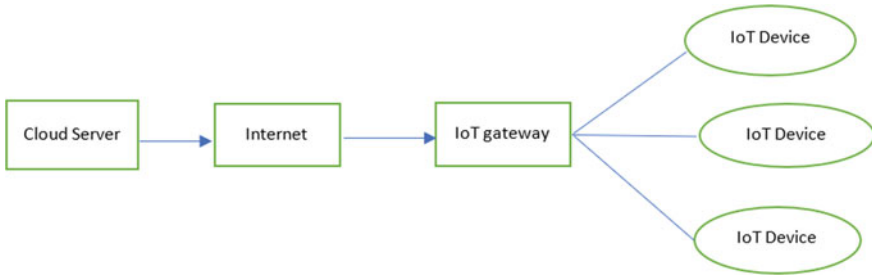


Fig. 1 Architecture of IoT crypto system

2 Implementation Details

The majority of IoT-Crypto software are built in C. Code written in C may be compiled and executed on almost any hardware platform, from microcontrollers to large servers. Maintaining code that performs visual operations would be complicated, and maintaining code written in many programming languages employs language-specific libraries. The programme is built on the Mbed TLS open-source C library. It's constantly being worked on and improved. There are just a few external dependencies, making it stand out from the crowd. Even if they aren't accessible on a given platform, Mbed TLS-based software may operate on any device, meeting the criteria of the introduction.

2.1 Protocol Stack

IoT-Crypto employs IP as part of TCP/network IP's layer. IPv4 and IPv6 compatibility is conceivable; using the same protocol version in a heterogeneous IoT network may ensure network visibility. Network access may employ any protocols or standards.

MbedTLS lies between TCP/transport IP and application layers. IoT networks have constrained devices, as said. TCP isn't optimal for them, but UDP is. Less overhead (memory, CPU, bits transferred, latency) reduces device stress. Because MbedTLS uses UDP-based DTLS instead of TCP-based TLS, DTLS doesn't change TCP's core characteristics (retransmission of lost packets and error correction). IoT-Crypto supports application-level TCP/IP stacks.

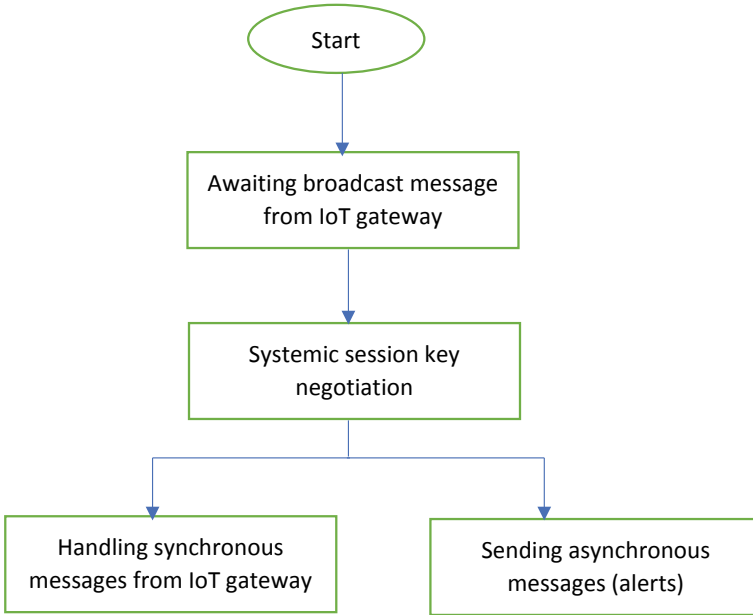


Fig. 2 Software operation system-Crypto system

2.2 *IoT-Crypto Device Software*

The IoT-Crypto programme is easy to use (Fig. 2). The gateway sends an initialization message after the device is switched on. Gateway id and network address are supplied. IoT devices and gateways negotiate. After crucial negotiation, the machine handles synchronous (request/response) gateway communications such as sensor reading requests and instructions. Sensor-based alerts from IoT devices may be sent to the gateway asynchronously. Both tasks are done concurrently yet separately. Developers must supply message handler and sender APIs for IoT-Crypto.

2.3 *IoT-Crypto Cloud Server Software*

There are no devices connected to the IoT-Crypto cloud server since it is situated on a completely separate network as shown in Fig. 3. Its software has two primary functions:

1. Users may engage with the IoT-Crypto network via a standard business application that exposes HTTP APIs and processes requests from users and other systems. IoT-Crypto network can also be interfaced using a cryptographic application.

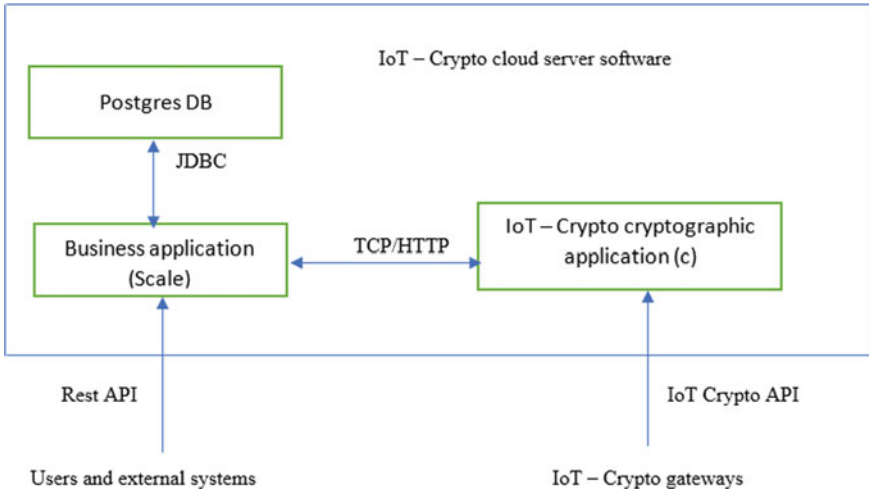


Fig. 3 Crypto service software system

- 2. Since it has previously been established, the heart of IoT-Crypto is built in C and relies on the MbedTLS library for its cryptographic operations. Compatibility and maintenance concerns will arise if the essential cryptographic functions are implemented in other languages. On the other hand, the C language cannot develop complex programmes.

The first is a Scala-based business application linked to a Postgres data warehouse. Random Language and database selections are made, and relational database and API support may have been used in a high-level programming language. The application developed in C using the MbedTLS library is a part of the second server software component.

Internal HTTP API is used for communication between two components of the server software that function as different apps. The business application exposes an HTTPS REST API to enable other systems to connect with and access resources in the IoT network. Cryptographic software provides an API for IoT gateways.

2.4 IoT-Crypto Gateway Software

The IoT-Crypto network relies heavily on the IoT gateway. It can connect to several IoT devices at the same time. Each gateway uses a single cloud server and serves as a middleman between IoT devices and the server. A subnet of IoT devices is formed when all devices are linked to a single gateway. The IoT subnet may use a variety of different communication protocols and wireless technologies at the same time. All of this must be handled by the gateway, which will also do any necessary protocol

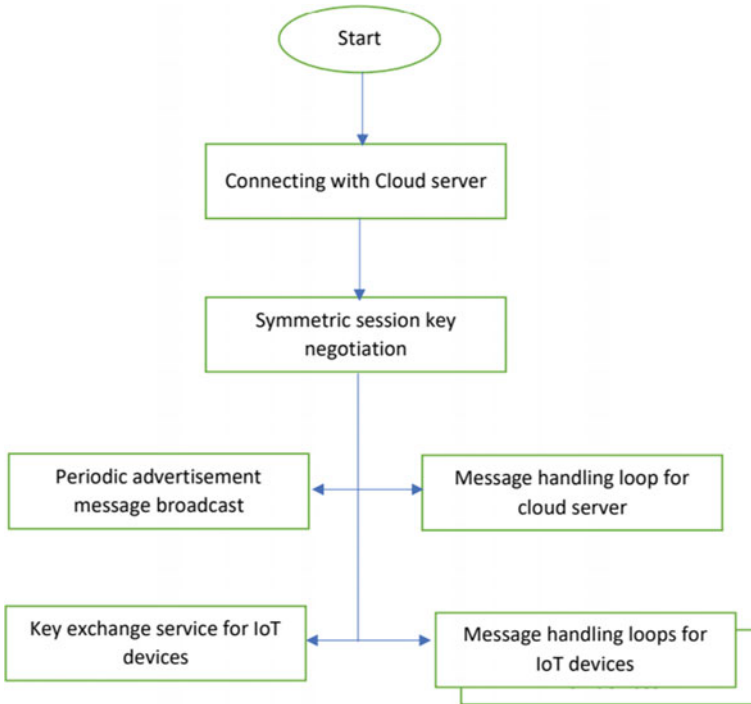


Fig. 4 Crypto gateway tool

translations. Wi-Fi, Bluetooth Low Energy (BLE), and Ethernet are often used at the network access layer of an IoT-Crypto network gateway.

Many tasks are carried out in the background by the gateway's multithreaded software (Fig. 4). The message processing loop begins after the cloud server has been successfully connected to the gateway. It sends data requests, orders to be sent to the devices, and certificate revocation notifications from a server to the client. IoT device advertisements are then broadcasted from the gateway. There are no network addresses or ports to enter manually, and device recognition is done automatically. For IoT devices, double duty is a crucial negotiation service. Before sending any application data, verifying the device and exchanging keys are essential. Every linked IoT device is then launched in a separate message processing loop.

3 Experiments and Tests

The IoT-Crypto solution is put through its paces under settings modelled after those seen in actual IoT installations. Depending on the setup, a Wi-Fi router provides Internet access and communication between devices in the test network. At the same

time, a cloud virtual machine plays the function of an IoT-Crypto cloud server on the virtual machine in the cloud. 3.1's IoT network topology is shown in this test network.

IoT-Crypto solution testing and performance measurements may be carried out with the help of this tool. Reconfiguring the network and testing other wireless communication technologies is also feasible. Each IoT subnet consists of a gateway and several IoT devices are positioned near one another. However, the cloud server may be at a considerable distance from the user's location. It is possible to simulate real-world network circumstances using a cloud virtual machine. Alternately, such circumstances may be intentionally simulated.

3.1 Protocol Overhead Related to the Cryptographic Techniques

IoT network security requires transmission and computing capacity, and handshakes safeguard communication channels. The IoT-Crypto network needs 1639 bytes of data. This cost is reasonable considering a single X.509 certificate might be over 2000 bytes.

Double overhead is data encryption. As seen in Fig. 5, an IoT-Crypto encrypted UDP packet is 29 bytes bigger than a non-encrypted packet and uses DTLS v1.2. DTLS v1.3 reduces this to 11 bytes per transmission. MbedTLS doesn't support this protocol version in IoT-Crypto.

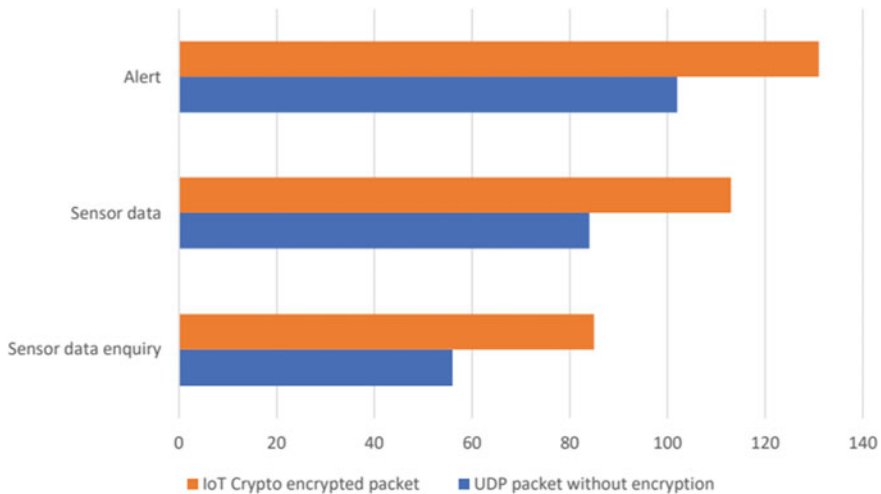


Fig. 5 IoT crypto protocol

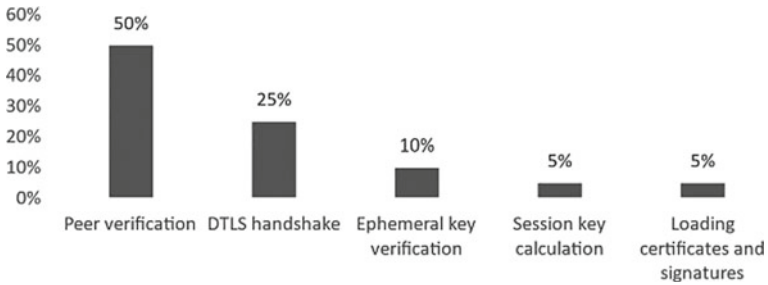


Fig. 6 IoT-Crypto computation overhead divided into categories

Analyzer for cool. The sizes of protocol messages and other data structures are determined using modified IoT-Crypto software. Additional encoding types are included in the upgrade.

According to PKI4IoT [10], DTLS-based solutions should have the same encryption overhead as UDP [11]. IoT-Crypto has less overhead than CoAP or HTTP. In PKI4IoT, CoAP queries are 6 to 28 bytes, and replies are 3 bytes. IoT-Crypto doesn't increase these costs. The PKI4IoT system has limited data on certificate and message sizes and communication overhead. Uncompressed, the certificate is 400 bytes. An IoT-Crypto certificate is 232 bytes in size. Certificates are supplied at the initial handshake effect protocol overhead.

Computers also have overhead. The time and resources needed are compared for the cryptographic procedures with and without encryption. Testing uses a quad-core ARM CPU and 2 GB of RAM. In contrast, this standard is more like a PC than a limited IoT device. Running IoT-Crypto applications with and without cryptographic algorithms did not substantially affect CPU use or power consumption. A-C profilers provide exact readings. Figure 6 displays research data.

3.2 Comparison of Encoding

The CBOR format encodes data that are transferred and stored in the IoT-Crypto network. Like the text-based JSON, it is a binary data structure. The stricter type control of CBOR is an additional advantage over JSON.

For the CoAP protocol, CBOR is a suggested encoding standard. CoAP is an alternative to HTTP that employs JSON encoding, which is more lightweight. This application layer protocol, IoT-Crypto, is developed from the ground up and is not based on any preexisting standard. Constrained systems may benefit from CBOR encoding. This specific application has also shown to be the best option compared to JSON and raw binary format.

CBOR encoded data saved in binary format is comparable to plain C language data structures in terms of size as shown in Fig. 7. More compact data storage and non-optimal C structures make it even more efficient in certain circumstances.

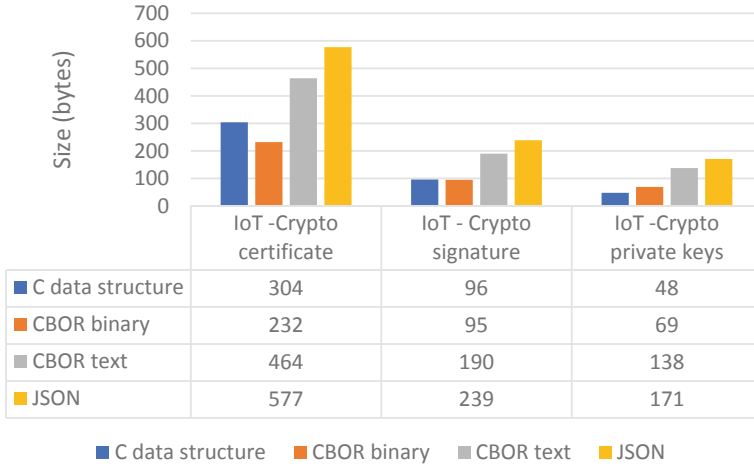


Fig. 7 Sizes of IoT-Crypto data structures using various encoding

3.3 Suitability of the Wireless Standard

IoT networks employ wireless protocols. Although this study did not compare or test, a secure IoT-Crypto network must support the TCP/IP protocol stack to transfer IP packets. Section 4.1’s initial test network uses Wi-Fi and Ethernet. Characteristics and skills are examined.

Small IoT networks aren’t built for Wi-Fi. IoT wireless networks are standardized individually. One of these standards may be examined in IoT-Crypto to verify performance and applicability. The BLE IPSP is tested on the 4.1 networks (IPSP). BLE’s IPSP supports TCP/IP, making it great for IoT devices.

6LowPAN is designed for IEEE 802.15.4 networks. Compression and an adaptive layer split large packages. IPSP removes GATT and utilizes the BLE link control layer to communicate (L2CAP). 6LowPAN over BLE has two appeals. First, it employs IPv6, assigning each device a public IP address. Second, Bluetooth Low Energy (BLE) is excellent for tiny IoT devices.

IoT-Crypto is IPv6-ready. Raspberry Pi is the test network’s IPv6 router and 6LowPAN BLE gateway. This arrangement doesn’t affect IoT-Crypto, and link testing is continued. One meter test achieves 240.001 kbit/s. Interference decreases distance-related throughput as shown in Fig. 8. BLE performance is affected by transmitting interval, wireless interference, and protocol overhead. The measured values are inside the BLE performance assessment range.

Iperf3 software is used to assess the average throughput of 6LowPAN BLE connections in a test network. Each test configuration is run 50 times to get the required data. The outcomes are consistent and reiterative. The averages of the data points can be seen in the graph.

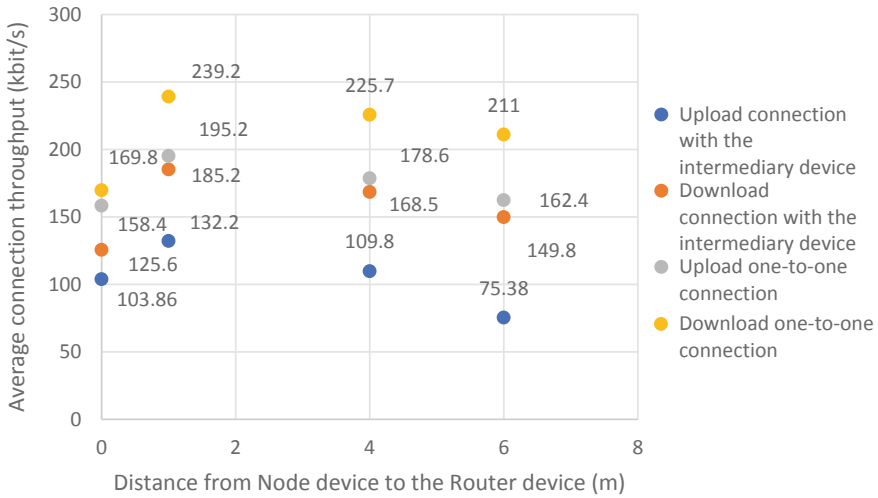


Fig. 8 Graphical representation of distance vs average connection throughput

There is an issue with the LowPAN BLE connection on the Raspberry devices when many devices are connected to the gateway device. The machines are pre-installed with Raspbian OS 4.19, using the Linux kernel 4.19. The Linux kernel's source code has a flaw. In the instance described above, the target address is not resolved appropriately. The kernel mechanism responsible for peer address search has been fixed. The Linux kernel is then patched to include the results of those tests and measurements.

There is a significant difference in measured round-trip times between Internet and Wi-Fi connections. Between the least and maximum findings, there is a considerable disparity. Figure 9 shows the average values of 200 outcomes from each arrangement. The devices are left in their default configuration for the tests to be carried out. The default BLE connection interval ranges from 30 ms (minimum) up to 60 ms (highest) (maximum connection interval). Rather than transmitting data as soon as they are available, data transfer events which occur at predetermined intervals are used. The results align with the expected based on the connection settings made. One-to-one connections may have up to 120 ms delay due to the connection interval. It may take up to 180 ms to connect with an intermediate device. Findings based on the thorough IPv6-based BLE connection investigation are broadly compatible with the results obtained in this study.

6LowPAN BLE transmission delays are observed using the Ping network tool. Each setup includes 200 measurements. The graphic representation shows the average performance.

The ability to encrypt the connection has shown to be unachievable. Plain text is used to transfer data. This problem is not mentioned in the Linux kernel documentation, and the source code does not indicate that this capability is present. The

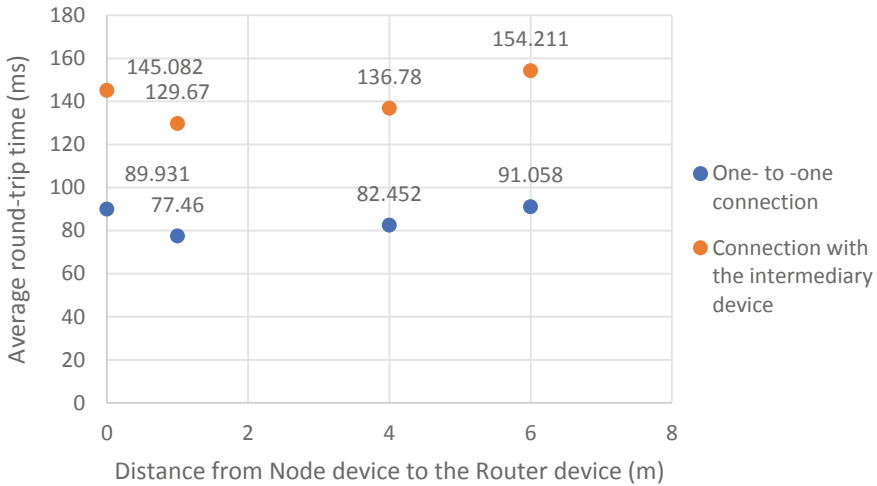


Fig. 9 Graphical representation of distance vs average round-trip time

application layer must implement a data encryption technique (like in the IoT-Crypto system).

4 Conclusion and Future Work

BLE’s connection can be utilized with IoT-Crypto. When sending small amounts of data, throughput and latency are sufficient. The results explain why BLE is more power-efficient than Wi-Fi. Data are transmitted in chunks, causing delays. Smaller throughputs use less power. Using IoT-Crypto, a fully functioning and secure IoT network can be constructed. IoT-Crypto has been developed on the most common architectural aspects, data flows, hardware restrictions, and security concerns.

A brand-new solution to overcome challenges caused by using standard Internet protocols not designed for IoT applications has been devised. This allows for far-reaching optimization. IoT-Crypto employs a novel application layer protocol and lighter X.509 certificate replacements. DTLS uses pre-existing protocols like CBOR, a data serialization standard. This solution’s most innovative feature is a cryptographic trust relationship built on real-world interactions between organizations employing IoT networks. It offers many choices and is simple to set up. Based on X.509 and OpenPGP alternative certificates, the proprietary certificate format has built this trust architecture.

- IoT-Crypto combines cryptography with autoconfiguration. Cryptography is used for device identification, removal, and network structure assessment.

- IoT-Crypto prioritizes security. To secure current communication, channels are built from the ground up utilizing cryptography.
- IoT-Crypto works on many devices. C-written runs on most TCP/IP devices (safe for the cloud server software). The test network successfully tested the anticipated mechanism.

IoT-Crypto calculates session keys using SHA-256, a significant enhancement. A Diffie-Hellmann shared key may be calculated eleven times faster using an elliptic key. Both results are compared using the IoT-Crypto package mbedTLS. IoT-Crypto has improved as a consequence.

Further work will include IoT-Crypto security research, penetration testing, and vulnerability assessments and to boost efficiency, cryptographic methods may be hardware-accelerated. Some microcontrollers may accomplish this. The SHA-256 calculation in Texas Instruments' CC13 \times 2/CC26 \times 2 models is up to forty times faster than an official report.

If the IoT device is interacting with the gateway using UDP/IP, DTLS header compression may minimize power usage. eeDTLS reduces protocol overheads from 77 to 7 bytes. eeDTLS allows clients to input certificate URLs instead of certificates. Using medials may reduce DTLS handshake energy usage.

UI design is also essential (GUI). The cloud server software has REST API. This subject has been omitted from the research since it doesn't affect network security. It may be used to talk with the network, issue commands, and receive sensor data. This API can develop an easy-to-use GUI for network monitoring and management. The third area of concentration will be on IoT-Crypto and other IoT solutions and future standardization and protocol review. IoT-Crypto may be improved and bugs may be fixed.

References

1. Yetis R, Sahingoz OK (2019) Blockchain based secure communication for IoT devices in smart cities. In: 2019 7th international Istanbul smart grids and cities congress and fair (ICSG), pp 134–138. <https://doi.org/10.1109/SGCF.2019.8782285>
2. Fakhri D, Mutijarsa K (2018) Secure IoT communication using blockchain technology. In: 2018 international symposium on electronics and smart devices (ISESD), pp 1–6. <https://doi.org/10.1109/ISESD.2018.8605485>
3. Goworko M, Wyrębowicz J (2021) A secure communication system for constrained IoT devices—experiences and recommendations. *Sensors* 21:6906. <https://doi.org/10.3390/s21206906>
4. Banerjee M, Lee J, Choo KKR (2018) A blockchain future for internet of things security: a position paper. *Dig Commun Netw* 4(3):149–160. ISSN 2352–8648, <https://doi.org/10.1016/j.dcan.2017.10.006>.
5. Attkan A, Ranga V (2022) Cyber-physical security for IoT networks: a comprehensive review on traditional, blockchain and artificial intelligence based key-security. *Complex Intell Syst* (2022). <https://doi.org/10.1007/s40747-022-00667-z>

6. Jabbar R, Kharbeche M, Al-Khalifa K, Krichen M, Barkaoui K (2020) Blockchain for the Internet of Vehicles: a decentralized IoT solution for vehicles communication using ethereum. *Sensors* 20(14):3928. <https://doi.org/10.3390/s20143928>
7. Sicari S, Rizzardi A, Grieco LA, Coen-Porisini A (2015) Security, privacy and trust in internet of things: the road ahead. *Comput Netw* 76:146–164
8. Danzi P, Kalor AE, Stefanovic C, Popovski P (2019) Delay and communication tradeoffs for blockchain systems with lightweight IoT clients. *IEEE Internet Things J* 6(2):2354–2365
9. Hasan RT et al (2021) *J Soft Comput Data Mining* 2(2):27–38
10. Ali J, Ali T, Musa S, Zahrani A (2018) Towards secure IoT communication with smart contracts in a blockchain infrastructure. *Int J Adv Comput Sci Appl (IJACSA)* 9(10):1–9. <https://doi.org/10.14569/IJACSA.2018.091070>
11. Lu Y (2017) Industry 4.0: a survey on technologies, applications and open research issues. *J Ind Inf Integr* 6:1–10
12. Alam T, Benaida M (2018) CICS: cloud-internet communication security framework for the internet of smart devices. *Int J Interact Mobile Technol* 12(6):74–84. <https://doi.org/10.3991/ijim.v12i6.6776>
13. Patro P, Azhagumurugan R, Sathya R, Kumar K, Kumar TR, Babu MVS (2021) A hybrid approach estimates the real-time health state of a bearing by accelerated degradation tests, Machine learning. In: 2021 second international conference on smart technologies in computing, electrical and electronics (ICSTCEE), pp 1–9, <https://doi.org/10.1109/ICSTCEE54422.2021.9708591>.
14. Bhau GV, Deshmukh RG, Chowdhury S, Sesharao Y, Abilmazhinov Y (2021) IoT based solar energy monitoring system. *Mater Today: Proc*
15. Kurnia RI, Girsang AS (2021) Classification of user comment using word2vec and deep learning. *Int J Emerg Technol Adv Eng* 11(5):1–8. https://doi.org/10.46338/ijetae0521_01
16. Dela LA, Cruz LK, Tolentino S (2021) Telemedicine implementation challenges in underserved areas of the Philippines. *Int J Emerg Technol Adv Eng* 11(7):60–70. https://doi.org/10.46338/ijetae0721_08
17. Mustapa RF, Rifin R, Mahadan ME, Zainuddin A (2021) Interactive water level control system simulator based on OMRON CX-programmer and CX-designer. *Int J Emerg Technol Adv Eng* 11(9):91–99. https://doi.org/10.46338/ijetae0921_11
18. Rahman ASA, Masrom S, Rahman RA, Ibrahim R (2021) Rapid software framework for the implementation of machine learning classification models. *Int J Emerg Technol Adv Eng* 11(8):8–18. https://doi.org/10.46338/IJETAE0821_02
19. Khotimah N, Wibowo AP, Andreas B, Girsang AS (2021) A review paper on automatic text summarization in Indonesia language. *Int J Emerg Technol Adv Eng* 11(8):89–96. https://doi.org/10.46338/IJETAE0821_11

CNN and XGBoost Based Hybrid Model in Classification of Fetal Ultrasound Scan Planes Images in Detection of Congenital Heart Defects



S. Satish and S. Sridevi

Abstract This work focuses on hybrid models in the classification of ultrasound scan planes in the detection of congenital heart abnormalities. The key elements of both the Deep Learning model and Machine Learning model are combined in this hybrid model. The Deep Learning model serves as a feature extractor in the proposed model, while the Machine Learning model serves as a binary classifier. As classification of fetal cardiac ultrasound scan plane plays an important role in detection of CHD. In Ultrasound scan planes such as 3 Vessel View (3VV), 4-Chamber View (4CV), and 3 Vessel Tracheal (3VT) are useful in detecting foetal heart abnormalities during the 18 to 24 week gestational period. In this paper, Convolutional Neural Network (CNN) + eXtreme Gradient Boosting (XGBoost) is used in order to classify the foetal ultrasound scan planes. The proposed hybrid model shows that the CNN + XGBoost model achieved a test accuracy of 98.65% using the custom dataset. This shows that hybrid model strategies are better than more standard deep learning and machine learning techniques.

Keywords Convolutional neural network · XGBoost classifier · Ultrasound images

1 Introduction

Congenital heart defects (CHDs) is common birth anomaly, accounting for about 40% of all congenital malformations. This defect is due to deficiency of normal development of heart during early stage of development. These problems are present during the birth that affects the structure and functions of the heart. They can also affect in the functioning of the blood flows through the heart and out to rest of the body [1, 2]. Screening during prenatal for congenital abnormalities, includes CHD,

S. Satish (✉) · S. Sridevi

Department of ECE, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Avadi, Chennai-62, India
e-mail: satishsaiece@gmail.com

S. Sridevi

e-mail: drssridevi@veltech.edu.in

can begin as early as 18 to 22 weeks of gestation period. Screening of CHD aims to measure the rate of heart beat, heart size, heart position, four chambers of heart, atrium, ventricles, atrioventricular junctions, and ventriculoarterial junction [3, 4].

Before the process of detection of CHDs begins, different defects can be examined with the help of Fetal Cardiac Ultrasound Images (FCUI). This FCUI takes different scan planes like the 3 Vessel View (3VV), 4-Chamber View (4CV), and 3 Vessel Tracheal (3VT) view. While detection of CHD deals with computer vision, the classification of scan planes plays an important role in the detection [5, 6]. Several studies involving deep learning approaches have claimed cutting-edge performance in a wide range of tasks. Image classification [7], natural language processing [8], speech recognition [9], and text classification [10] are a few examples. In the existing models employed in the aforementioned tasks, the softmax layer is used at the classification process.

To learn about image classification from the basic, have to study with the conventional based machine learning technique and progress of deep learning techniques. To investigate the basic ideas and techniques of image classification, as well as to high point the latent qualities and importance of deep learning ability over more conventional machine learning technique [14]. However, studies [11–13] have been conducted that looks at an choice for the softmax function for image classification—XGBoost. Accordingly, using XGBoost produces significantly better results than using the traditional softmax layer. There is one disadvantage on this approach that the limitation of binary class classification. Because XGBoost seeks to find best hyper-plane between two different classes in a given dataset, the multi-nomial case appears to be disregard. When XGBoost is used in a multi-nomial classification method, the above case becomes a one-to-one in all situation, with positive classes representing the class with the higher score and the remaining representing the non-positive class. The main benefit of utilising CNN is that it recognises significant features without the need for human intervention, performs dimensionality reduction without sacrificing model quality, and generates high accuracy. Based on the advantages of CNN and XGBoost classifier, a hybrid model is designed and its performance metrics will be analysed with the clinical dataset of fetal cardiac ultrasound images.

2 Related Works

Then researchers have worked on various applications of image processing since its inception in the research fields includes medical imaging, image brightening and rehalibation, sensing of UV, video processing, colour processing, and so on. With real-time applications, the medical field's application of digital image processing is growing by the day. There are two general classification on FCUI innovations like Deep Learning and the traditional Machine Learning methods. Innovations based on Deep learning grows quickly in the field and achieves many notable achievements.

In [15], this work attempts to investigate and nurture the basic understanding of various image classification based approaches and methods, as well as to discuss

various approaches based on various types of possible inputs. In [16], two areas used in image classification are the ANN and SVM. The classification of sub-image results in a responsive class by an ANN. The experimental results show that recognition application and precision rate are both 86%. The model explained is known as ANN SVM because it incorporates numerous ANN and one SVM in the work. In [17], this paper combines the CNN + SVM for image classification of brain tumours with an accuracy of 97.1%. SVM cast-off to improve the accuracy of the proposed model using the features extracted from the CNN model. In [18], we design a CNN model for obtaining feature extraction and then used a SVM model to work with those features and classify them accordingly. The proposed work was then compared with various datasets provides the work's accuracy of 95.82%. Cropping an ultrasound images are done during the preprocessing phases in [19, 20], and [21] in order to identify the scan planes of foetal cardiac ultrasound images. To accomplish effective segmentation, additional steps were required. Zar et al. proposed a transfer learning steps in [22] that included preprocessing before the CNN training process. The VGG-19 pre-trained model was then used for tuning. The proposed model was found to be 94.82% accurate.

Convolutional neural networks had used successfully for handwritten digit recognition, mainly on the MNIST handwritten digit dataset. The CNN model achieves recognition accuracy of 98%. The highest classification accuracy of 99.73% on the MNIST dataset were achieved by analysing with popular method technique of merging number of CNNs into an effective ensemble model [22]. Lauer et al. [23] developed a feature extractor model for the MNIST database based on LeNet5 convolutional neural network architecture. The work demonstrated exceptional recognition accuracy. The research work's impressive performance clearly demonstrates the effectiveness of the CNN based extraction of features. This present work also motivates the performance of deep learning method in the area of medical image analysis. An ductile form based interactive model [25] to segment the CT scanned image of lung and it proves to be more efficient. The proposed algorithm proves that function of energy of system is low than older methods. An form-based method of segmentation had been used and it proves accuracy is better with low function of energy.

3 Proposed Hybrid Model

The work flow of proposed hybrid model is shown below in Fig. 1 and the working of CNN + XGBoost hybrid model given in detail.

In working function of the hybrid model, first the foetal ultrasound images are given to the labelling section, where labelling of the images will take place. Labelling of images had been done based on the scan planes. Four chamber view (4CV) labeled as 0 and three vessel view (3VV) labeled as 1 and grouped under the corresponding folders. Then the labelled images are further splited into training images and testing images. Here the training images and testing images are splitted in the ratio of

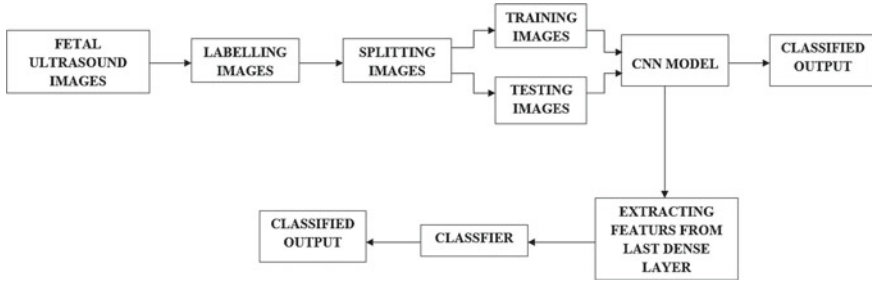


Fig. 1 Work flow of the hybrid model

70:30. Training and testing images are given to the CNN model to train and test the model. The feature is extracted from the final dense layer of the CNN model and extracted feature is given to the XGBoost classifier for the classification of fetal cardiac ultrasound scan plane images.

3.1 CNN + XGBoost Hybrid Model

In the field of computer vision, deep learning is the fast-evolving approach. The term convolutional neural network comes to picture when we talk about computer vision because it is employed in various field. The image based various computer vision applications use CNN. It resembles a rudimentary neural network. CNN has learnable parameters like neural networks, such as weights and biases.

The convolutional layer (CONV) comprises a filter, such as WH3 (W for width and H for height, and 3 for grayscale pictures). In the proposed model it consists of 5 CONV layer. The CONV layer are as a used to slide over the row and column of an input image and it computes the convolution product of the input region and learning weight parameters. As a result, a 2-dimensional activation map containing the filter’s responses to specific regions will be generated.

The Pooling Layer is commonly used as a link connecting the CONV Layer and the fully connected layer. The pooling layer decreases the size of input images based on the CONV filter’s results. It is useful for segregating the dominant features that are of different invariant, allowing that the model to be efficiently trained. In the proposed model it consists of 5 max pooling layer. Max pooling gives the higher value from the filter-covered part of image.

Flattening is to convert all resulting two dimensional array from pooled feature maps into a single linear vector and linearize it before passing it to the Dense layer. In the proposed model it consists of one flatten layer with 1028 neurons.

The dense layer manipulates the ambit of the next layer output, allows the model to easily relate the relationship between the value of data in which the model is working. In the proposed model it consists of 2 dense layer. Dense 1 layer consist of 128 neurons and dense 2 layer consist of 32 neurons as it will be given to the machine

learning classifier. In this model, neuron of the dense layer obtains output from every neuron to the next layer, where the neurons in dense layers perform matrix vector multiplication. The vectors's row of output from the next layer is equal to the vector's column of the dense layer in vector matrix multiplication. The row vector must have the equal number of columns as the vector's column, according to the general rule of matrix-vector multiplication.

Finally, an activation function was employed to make nonlinearities into the computation. The model will learn only linear mappings if this is not provided. The ReLU function is the most commonly used activation function these days (Fig. 2). Rectified Linear Unit (ReLU) is a linear unit. It helps to avoid exponential growth in the computation required to run the neural network. All negative values become null immediately in this case, by reducing the capacity of the model to fit or train from the given data correctly. That is, any non positive input to the ReLU activation function change the value into zero in the graph immediately, which has an effect on the final graph by not mapping the non positive values effectively.

Feature extraction is the process of extracting unprocessed data into features that may be processed while keeping the original dataset context. It produce better outcomes than simply using machine learning to the unprocessed data. Feature extraction helps in the diminshment of unwanted data in a dataset. Atlast, limiting the data makes it convinient to assemble the model with less effort and speeds up the learning rate and general processes in the machine learning.

The gradient boosting machine learning ensemble method is extended in XGBoost. The gradient boosting method sequentially combines the decisions of weak classifiers, resulting in an effective ensemble decision tree learning machine. Additionally, the XGBoost method reduces the gradient boosting method's computational complexity and calculation time. XGBoost is a powerful classifier method that can produce cutting-edge results on a variety of challenges.

4 Result Analysis

To train and test the proposed model, a dataset comprised of 525 foetal cardiac ultrasound images is used, of which 369 images of fetal ultrasound are used for training the hybrid model and 156 images of fetal ultrasound are used for testing the model. Here, performance analysis of the proposed CNN + XGBoost model is compared with the other hybrid models like CNN + Support Vector Classifier (SVC), CNN + Random Forest (RF), CNN + Gaussian Naive Bayes (GNB), CNN + Decision Tree (DT) and CNN + K-Nearest Neighbour (kNN).

In the analysis, various performance metrics have been considered, like accuracy, recall, F1-score, precision, time, and loss shown in Table 1.

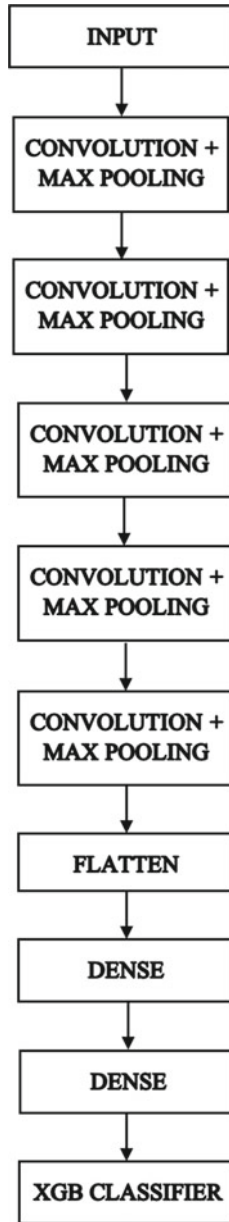


Fig. 2 CNN + XGBoost Hybrid model for image classification

Table 1 Accuracy, Recall, F1-score, Precision, Time, and Loss of hybrid model

Parameter	CNN + XGBoost		CNN + SVC		CNN + RF		CNN + GNB		CNN + DT		CNN + kNN	
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Accuracy	97.68	98.65	82.11	86.53	71.27	76.53	55.01	57.76	71.54	80.76	72.43	74.31
Precision	96.54	99.94	81.22	86.53	69.12	76.53	51.86	55.76	72.08	80.76	73.59	75.43
Recall	98.94	99.94	83.90	86.54	80.61	86.69	57.32	68.89	72.27	80.85	73.57	75.43
F1-Score	97.54	99.92	81.54	86.53	67.41	72.54	42.33	46.67	71.53	80.75	73.58	75.23
Loss	0.0215	0.0124	0.1788	0.1346	0.2872	0.1346	0.4498	0.4423	0.2845	0.1923	0.2628	0.2756
Time	4.046	1.6417	4.1643	1.9144	4.4948	2.0508	4.1392	1.8493	4.1536	1.8539	4.1384	1.8482

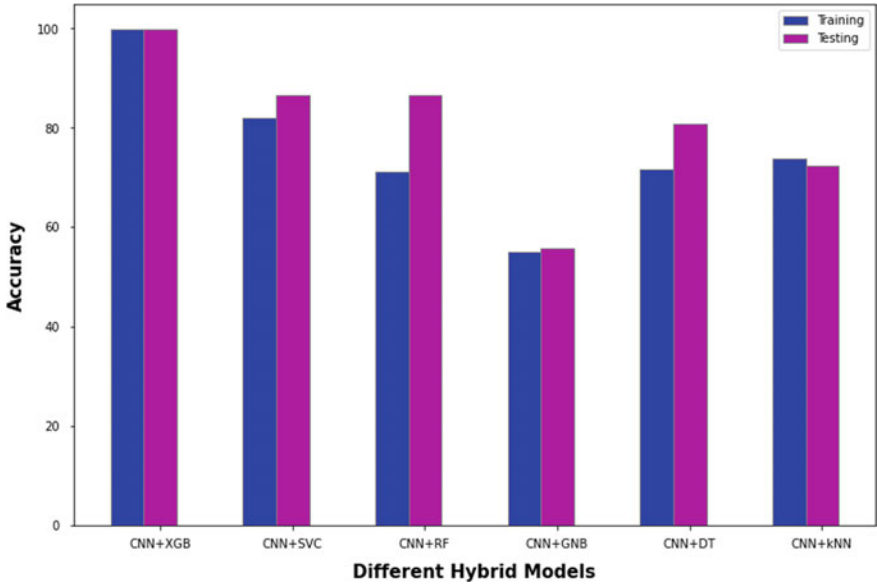


Fig. 3 Performance metrics of CNN + XGBoost hybrid model - training and testing accuracy

4.1 Accuracy

Accuracy is a metric that gives how the model perform well for all classes. It is determined by ratio the valid predictions to the overall predictions. Based on our experimental result, we found that the CNN and XGBoost based hybrid model produces high training and testing accuracy. This shows that the model makes the correct prediction and an equal distribution of classes has been made compared to the other hybrid models as noted from the above Fig. 3.

4.2 Precision

Precision is ratio of properly differentiated non negative samples (True Positive) to sum of positively classified samples (either correctly or incorrectly). It allows to visualise the machine learning models trustability in categorising the model as non negative. Based on our experimental result, we found that the CNN and XGBoost based hybrid model produces high training and testing precision. As can be seen from the above Fig. 4, the suggested model made correct real positive predictions from the total predicted positive observations when compared to the other hybrid models. As higher precision rates are related to low false positive rate.

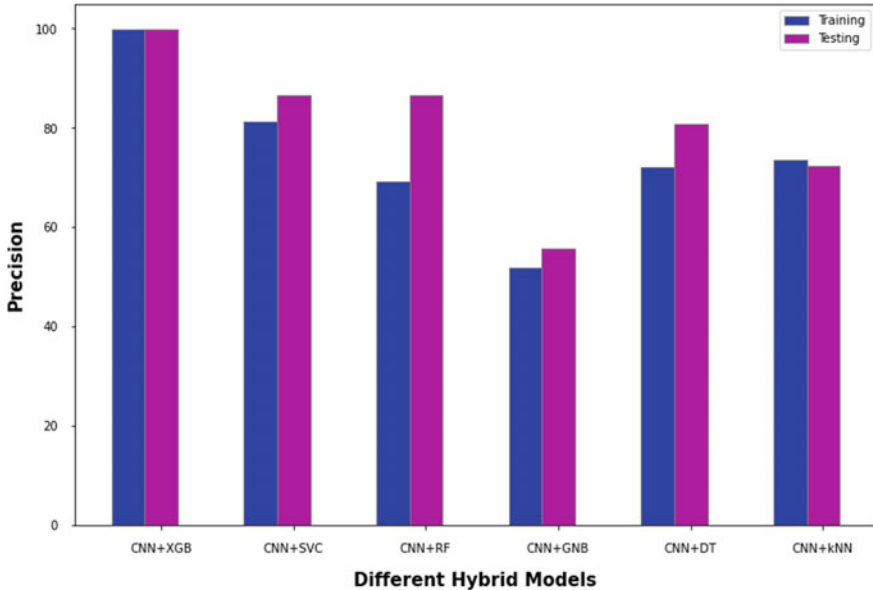


Fig. 4 Performance metrics of CNN + XGBoost hybrid model - training and testing precision

4.3 Recall

Recall is the ratio of non negative samples that are classified correctly to the non negative class samples. It measures its capability to detect samples were positive. The more samples detected, the higher the recall. Based on our experimental result, we found that the CNN and XGBoost based hybrid model produces high training and testing recall. This implies that the model is successful in locating all positive examples in the data, despite the fact that it may incorrectly label some negative cases as positive compared to the other hybrid models, as noted from the above Fig. 5.

4.4 F1-Score

F1-score gives the correctness of a dataset. F1-score is used to appraise binary classification systems that categorise it as positive/negative. Based on our experimental result, we found that the CNN and XGBoost based hybrid model produces a high training and testing F1-score. As shown in Fig. 6, the F1-score of the proposed model is a better estimate of erroneously classified cases than the accuracy metric.

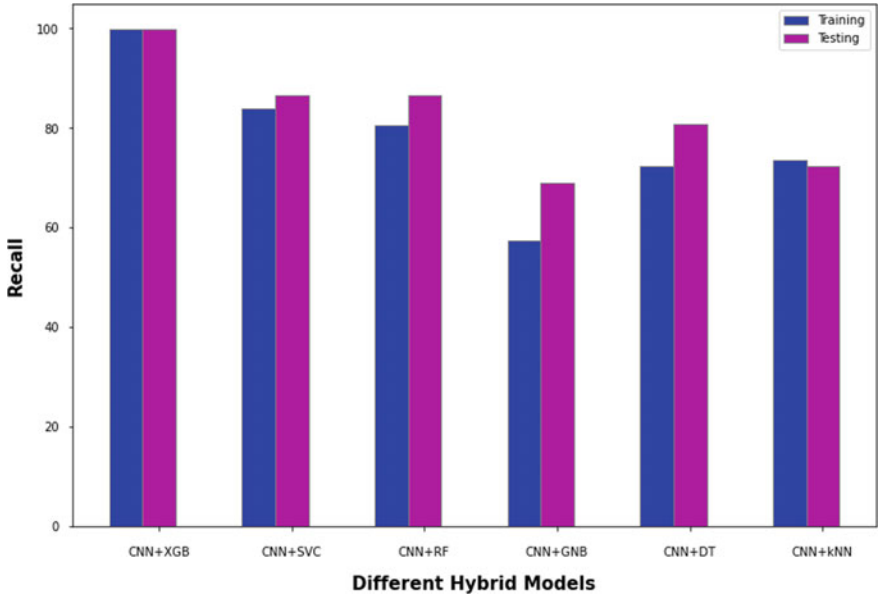


Fig. 5 Performance metrics of CNN + XGBoost hybrid model - training and testing recall

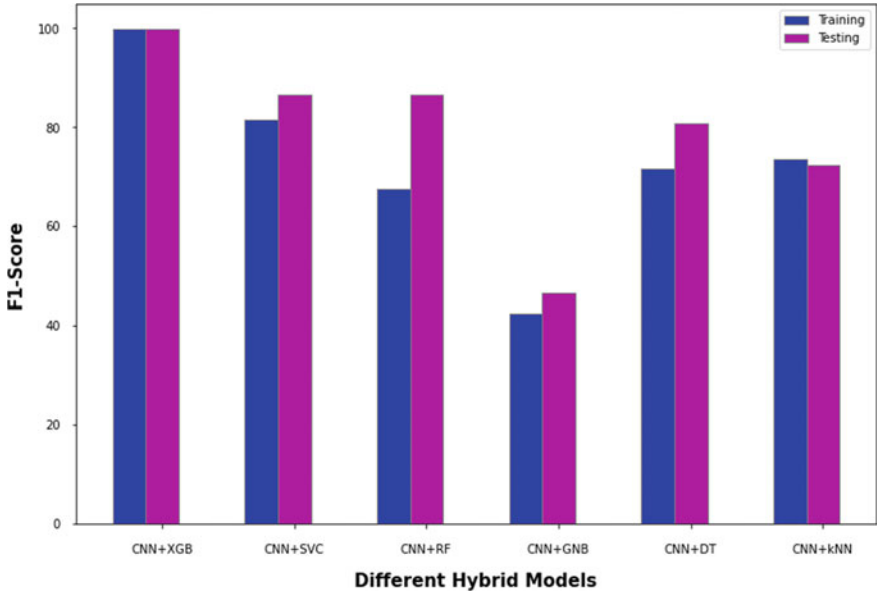


Fig. 6 Performance metrics of CNN + XGBoost hybrid model - training and testing F1-score

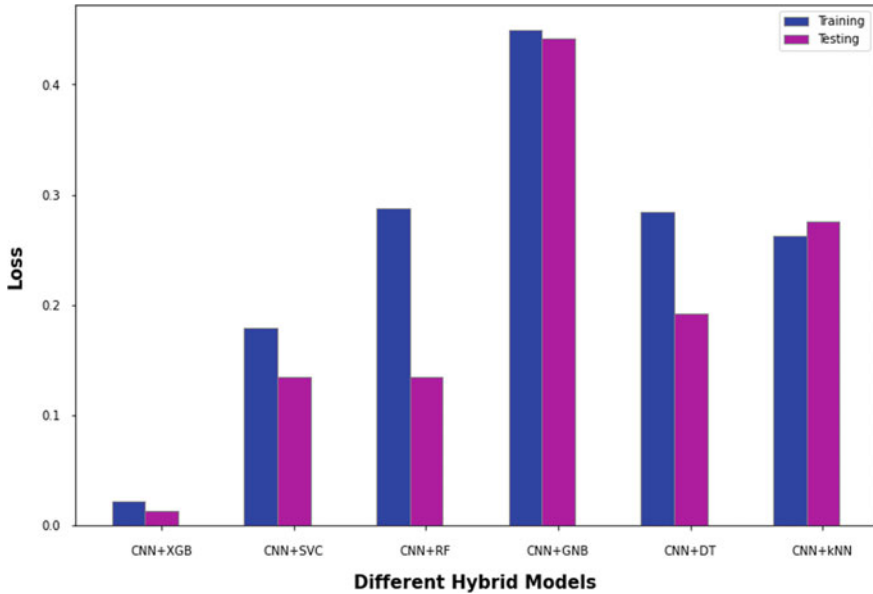


Fig. 7 Performance metrics of CNN + XGBoost hybrid model - training and testing loss

4.5 Loss

A loss is the penalty for making an incorrect prediction. Based on our experimental result, we found that the CNN and XGBoost based hybrid model produces very low training and testing loss. As shown in Fig. 7, the hybrid model operates effectively and has a low loss when compared to the other hybrid models.

4.6 Time

Based on our experimental result, we found that the CNN and XGBoost based hybrid model produces low training and testing time. As shown in Fig. 8, the proposed model takes less time to train and test than other hybrid models.

4.7 ROC Curve

Receiver Operating Characteristics Curve (ROC curve) is metric used to assess the fulfillment of a classifier model. It shows rate of positive instances in affinity to the number of inaccurate non negative. From the above ROC curve for CNN + ML based

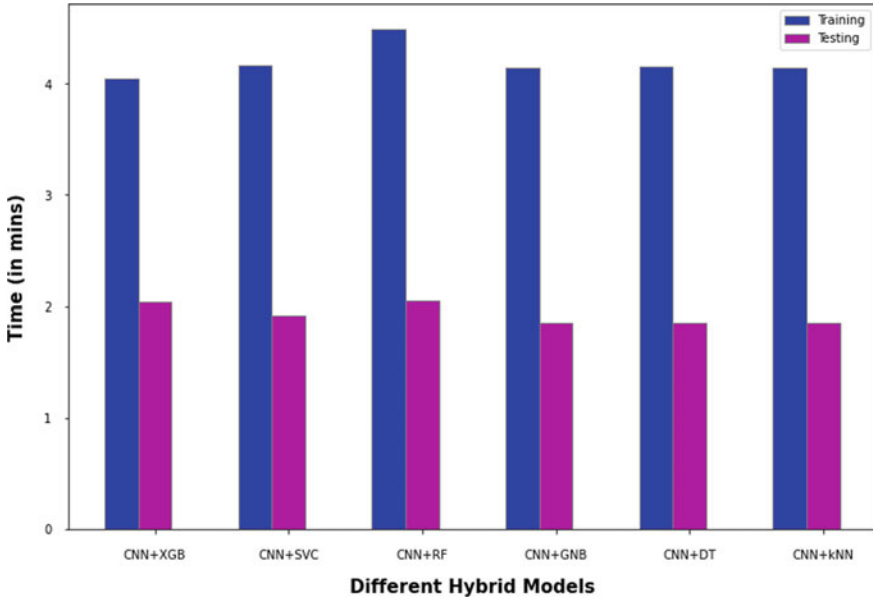


Fig. 8 Performance metrics of CNN + XGBoost hybrid model - training and testing time

hybrid model, it is clearly noticed that the XGBoost classifier (ROC AUC - 0.9801) out performs over the other machine learning classifier. (Random Forest – 0.9248, K-Nearest Neighbour – 0.9769, Gaussian Neighbour Bayes – 0.8446, Decision Tree – 0.9368 and SVM classifier – 0.9783) have been inferred from the above Fig. 9.

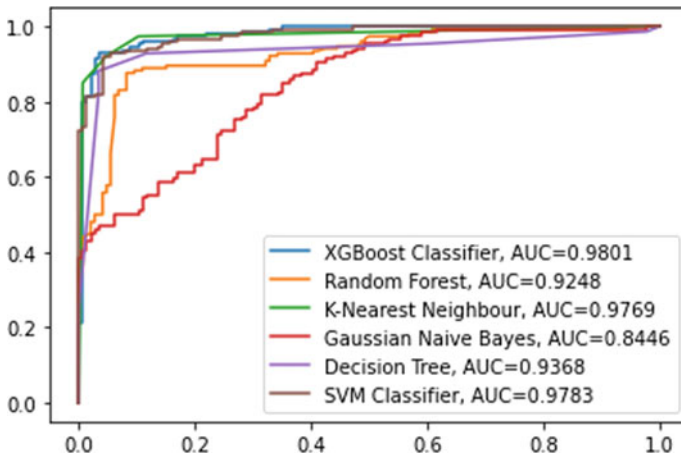


Fig. 9 Performance plot for ROC curve for CNN + ML based hybrid model

Table 2 Sensitivity and specificity of hybrid model

Parameter	CNN + XGBoost	CNN + SVC	CNN + RF	CNN + GNB	CNN + DT	CNN + kNN
Sensitivity	0.0214	0.4589	0.3976	0.1410	0.3845	0.4561
Specificity	0.9754	0.8715	0.8974	0.9743	0.8334	0.7357

4.8 Sensitivity and Specificity

On comparing the sensitivity and specificity of the hybrid CNN + XGBoost classifier to other hybrid models, the findings in Table 2 show that it has good sensitivity and specificity. The sensitivity of each accessible category can predict real positives, whereas its specificity can predict true negatives.

5 Conclusion

In this paper, the hybrid model for CNN + XGBoost classifier is used for classifying foetal ultrasound scan images that involve the features extracted using CNN and predicting the output based on XGBoost classifier. This model merges the advantages of CNN and the XGBoost classifier for classifying the scan planes of foetal ultrasound images. The simulation result showed that the proposed approach achieves a classification accuracy of 98.65% for the given custom dataset. Different performance metrics like loss, time, F1-score, Recall, precision, Sensitivity, Specificity and ROC had been analysed. The simulation results are compared with the other hybrid models and inferred that the proposed hybrid model perform well. Some hyperparameter tuning can be executed to reduce the time taken for classification operation for different applications using the pretrained models.

References

1. Asbagh PA et al (2021) Prevalence of factors associated with congenital heart disease. *Multidiscip Cardio Annal* 12(1):e106026
2. Pavlicek J et al (2019) Associations between congenital heart defects and genetic and morphological anomalies. The importance of prenatal screening. *Biomed Pap Med Fac Univ Palacky Olomouc Czech Repub* 163(1):67–74
3. Bjornard K, Riehle-Colarusso T, Gilboa SM, Correa A (2013) Patterns in the prevalence of congenital heart defects, metropolitan Atlanta, 1978 to 2005. *Birth Defects Res A Clin Mol Teratol* 97:87–94
4. Hunter LE, Seale AN (2018) Educational series in congenital heart disease: prenatal diagnosis of congenital heart disease. *Echo Res Pract* 5(3):R81–R100. <https://doi.org/10.1530/ERP-18-0027>
5. Callen PW (2016) *Ultrasonography in Obstetrics and Gynecology*. Saunders Elsevier, Philadelphia

6. Bernolian N, Kesty C, Widodo BW (2020) Current update on congenital heart disease screening in pregnancy. *Majalah Kedokteran Sriwijaya* 52(2)
7. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. *Adv Neural Inf Process Sys* 1:1097–1105
8. Wen T-H, Gasic M, Mrksic N, Su P-H, Vandyke D, Young S (2015) Semantically conditioned LSTM-based natural language generation for spoken dialogue systems. arXiv preprint [arXiv:1508.01745](https://arxiv.org/abs/1508.01745)
9. Chorowski JK, Bahdanau D, Serdyuk D, Cho K, Bengio Y (2015) Attention-based models for speech recognition. *Adv Neural Inf Process Sys* 577–585
10. Yang Z, Yang D, Dyer C, He X, Smola AJ, Hovy EH (2016) Hierarchical attention networks for document classification In: *HLT-NAACL*, pp 1480–1489
11. Agarap AF (2017) A neural network architecture combining gated recurrent unit (GRU) and support vector machine (SVM) for intrusion detection in network traffic data. arXiv preprint [arXiv:1709.03082](https://arxiv.org/abs/1709.03082)
12. Alalshkumbarak A, Smith LS (2013) A novel approach combining recurrent neural network and support vector machines for time series classification. In: *2013 9th international conference on innovations in information technology (IIT)*. IEEE, pp 42–47
13. Tang Y (2013) Deep learning using linear support vector machines. arXiv preprint [arXiv:1306.0239](https://arxiv.org/abs/1306.0239)
14. Chaganti S, Nanda I, Rao Pandi K (2022) Image Classification using SVM and CNN. [online] ieeexplore.ieee.org.
15. Kamavisdar P et al (2013) A Survey on Image Classification Approaches and Techniques. *Int J Adv Res Comput Commun Eng* 2(1)
16. Thai LH et al (2012) Image classification using support vector machine and artificial neural network. *J Inf Technol Comput Sci* 5:32–38. <https://doi.org/10.5815/ijitcs.2012.05.05>
17. Sejuti ZA, Islam MS (2021) An efficient method to classify Brain tumor using CNN and SVM. In: *2nd international conference on robotics, electrical and signal processing techniques*, pp 644–648
18. Deepak S, Ameer PM (2020) Automated categorization of brain tumor from MRI using CNN features and SVM. *J Ambient Intell Humanized Comput* 1–13
19. Sridevi S, Nirmala S (2015) ANFIS based decision support system for prenatal detection of truncus arteriosus congenital heart defect. *Applied Soft Computing*
20. Sridevi S, Nirmala S (2016) Markov random field segmentation based sonographic identification of prenatal ventricular septal defect. In: *7th international conference on communication, computing and virtualization*
21. Sridevi S, Nirmala S (2014) Fuzzy connectedness based segmentation of fetal heart from clinical ultrasound images In: *Smart innovations systems and technologies*, vol 27. Springer, pp 329–337
22. Swati ZNK et al (2019) Brain tumor classification for MR images using transfer learning and fine-tuning. *Comput Med Imaging Graph* 75:34–46
23. Ciresan DC, Meier U, Masci J, Gambardella ML, Schmidhuber J (2011) Flexible, high-performance convolutional neural networks for image classification. In: *Proceedings of twenty-second international joint conference on artificial intelligence*, pp 1237–1242
24. Krizhevsky A, Sutskever I, Hinton GE (2017) ImageNet classification with deep convolutional neural networks. *Commun ACM* 60(6):84–90
25. Tammina S (2019) Transfer learning using VGG-16 with deep convolutional neural network for classifying images. *IJSRP Volume 9, Issue 10, October 2019 Edition Latest Research Papers/TOC*. [online] [Ijsrp.org. http://www.ijsrp.org/research-journal-1019.php](http://www.ijsrp.org/research-journal-1019.php)
26. Sathish (2020) Adaptive shape based interactive approach to segmentation for nodule in Lung CT scans. *J Soft Comput Paradigm* 2(4):216–225

Review on Web Application Based Infant Health Management



Deekshitha S. Nayak, Sannidhi Rao, Sharan Kumar, Niharika Rao, and Himanshu Bhatt

Abstract The neonatal phase is the most vulnerable time of life and is the important period for child survival. Therefore, improving newborn survival is very important and is a priority. It is also very important to monitor the postnatal health condition of the baby and manage and store the data to examine the health condition and consider the baby as healthy. The health parameters on which infant is monitored are temperature, pulse rate, humidity, and motion capture. To ensure the good health of the newborn, certain health parameters need to be monitored. In this paper Infant health monitoring system consists of a web application-based health management system for a neonatal that keeps track of the neonatal health is reviewed.

Keywords Incubator · Temperature · Neonatal · Health monitoring · Parameters

1 Introduction

During the past twenty years, substantial efforts were made to ensure a healthy birth rate and to reduce morbidity and death among mothers and newborns. In developed countries, lower postpartum and neonatal mortality have been achieved following the advancement of neonatal care [1]. Despite this, many newborns die every day due to a lack of monitoring during the neonatal and postnatal periods. To overcome this, incubators must be monitored in a systematic way to ensure the safety of newborns. To provide a healthy environment to newborns, the two important aspects that need to be monitored are temperature and humidity [2]. Variation in the body temperature of the newborn leads to severe health conditions such as hyperthermia which can lead to abnormally high body temperature in neonates. Furthermore, when the temperature rises, so does the humidity in the air [3]. To avoid this, the temperature, humidity, pulse rate, and other health parameters in the incubator should be regularly checked in order to provide an appropriate environment and preserve the newborn's core

D. S. Nayak (✉) · S. Rao · S. Kumar · N. Rao · H. Bhatt
Department of Electronics and Communication Engineering, Mangalore Institute of
Technology & Engineering, Moodbidri, Mangaluru, India
e-mail: deekshitha.nayak06@gmail.com

temperature [4]. An incubator is a machine that monitors and maintains the greatest possible environment for a newborn baby. It is used in the treatment of premature births and some sick full-term newborns. The health of the baby is well managed. The oxygen supplementation and pressure levels are monitored by the incubator. It also keeps track of the radiation pulse activity, temperature, air humidity, and gas levels in the surrounding area [5]. The system monitors the temperature, humidity and noise inside the incubator. The data is stored and transferred to the cloud storage using sensors and data transfer devices. Medical data can be examined and acted upon using mobile phones and computer systems from wherever they are. The device emits an alarm signal if there is an issue with the medical data. As a result, they can keep the newborn safe from problems.

2 Types of Web Applications

2.1 *Internet of Things*

IoT can be used to monitor baby health and also parameters such as pulse rate, temperature, and humidity. An alert system can also be implemented to notify the parents and doctors [1]. It plays an important role on incubators in medical sector. Wireless technology can be used for remote control and monitoring the incubator which allows the medical staff to monitor the baby and assess the parameters from different locations. Humidity and temperature can be accurately observed. Remote monitoring can be achieved by the help of computers to measure the parameters of the baby [2]. The sensors and data transfer devices are used to store the data and transfer to the cloud [23]. The accurate values are synced for every second and are displayed so that both doctor and parents can have access to the baby's health condition and avoid health problems [24]. If there are any variations in the result an alert message is given to hospital management and patient home. This will be helpful for parents to take care of the infant and also will provide high security for the baby [3].

Figure 1 depicts the block diagram of the proposed system for implementing the monitoring system to monitor the temperature, humidity, pulse rate, and noise inside the incubator. This device contains three sensors that detect the heartbeat of the neonatal which is stored temporarily in the microcontroller, displayed on an LCD, and transmitted to the IoT dashboard using a Wi-Fi module. The parameters can be viewed from the mobile application. If there are any variations or abnormalities, the device gives an alarm signal using the buzzer. In the application of Internet of Things (IoT), the microcontroller used is STM32 F103C8T6. It's an ARM Cortex-M3 core microcontroller from the STM32 family. It is interfaced with sensors, display and Wi-Fi module shown in Fig. 1.

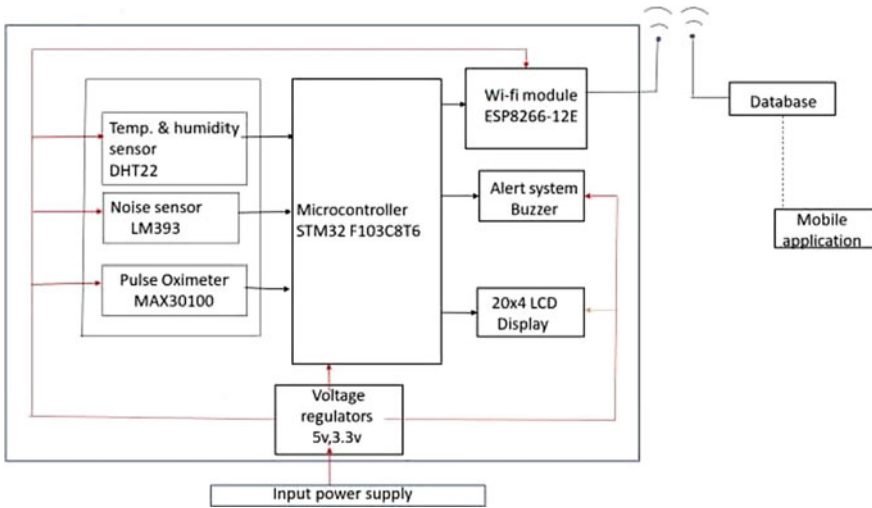
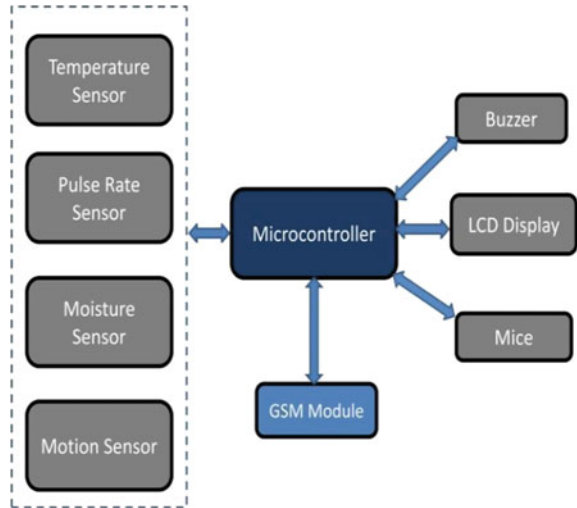


Fig. 1 Block diagram of IoT [1]

2.2 Embedded Systems

Embedded systems can be used for the implementation of real-time monitoring of humidity and temperature. Arduino UNO microcontroller can be used to obtain the parameters, and Dragino LoRa can be used to connect the platform to the sensors. The use of wireless technology is necessary because detecting temperature using a thermometer is not effective [6]. In addition to these parameters, Information on the infant’s heart rate, moisture status, and movement is sent to their parents via an alarm triggering system, which triggers the appropriate control steps. Sensors, LCD screen, GSM interface, and sound buzzer are all part of this system [7]. This detects the baby’s motion and sound and displays it on the monitor so that the doctor or parent may take precautions [5]. The MIC detects the baby’s crying, the PIR motion sensor detects the baby’s movement, and the Pi camera captures the baby’s motion. Based on the results, which are presented on an LCD in the staff room, the temperature monitor system can monitor and detect the temperature in the baby incubator room. A temperature detector called an LM35 can be calibrated directly on the Celsius scale [8]. Data is transmitted from a remote point using the wireless system. The heartbeat sensor counts the heartbeat for a particular time interval using a photodiode. This technique allows sensor equipment to move around and saves money [9].

Fig. 2 Block diagram of Embedded System [7]



The architecture, which includes both hardware and software, is depicted in Fig. 2. The LM35 with accuracy of 86% temperature sensor measures temperatures between 40 and 125 °C and operates at 3 to 5 pulses per second. The GPS Sensor is about 15 m, SEN 1853 - $\pm 5\%$ RH (at 25 °C, 60% RH), and Thermometer - 0.5 to 1 °C. It is made up of a 5 mm photodiode, light-emitting diode, high pass filter, and amplifier. The pulse rate will be detected using optical sensors on the finger and presented on an LCD. Motion sensors will measure acceleration forces which can be static or dynamic. GSM is a digital mobile telephony system that enables a wireless system with no specified range. Moisture sensors are used to determine the moisture condition and the signal obtained is given to the microcontroller. The microcontroller utilised in Embedded Systems is the PIC 18f4520, which is an 8-bit microcontroller with on-chip eight channel 10-bit Analog-to-Digital Converters (ADC). The microcontroller in Fig. 2 detects the enhanced sensor signals.

2.3 Image Processing

Image processing is used to achieve non-contact infant monitoring for proper safety and to track the baby’s activity. When the baby cries, the MIC receives an input signal from the baby and sends it to the Raspberry Pi module, which then activates the pi camera and sends an alert mail to parents and doctors. When the baby cries, the MIC receives an input signal from the baby and sends it to the Raspberry Pi module, which then activates the pi camera and sends an alert mail. It is more baby, less expensive, and less harmful [10]. Baby motion is depicted by a control chart that shows the baby’s abnormal behaviour. When this type of behaviour is detected, a signal is sent to all connected devices in the IoT [11]. The live baby streaming via

smart devices to check on the baby’s condition. It is possible to attain a higher level of time complexity and accuracy.

Figure 3 depicts the proposed system for remotely monitoring a baby’s activities comprises of a sensor, a hardware unit, a cloud server, and a parent’s application. This system is built on a Raspberry Pi with an RPi module that can be wired or wirelessly linked. Three buttons and one LED are connected to GPIO pins on the keypad. The Pi cam is used to detect the baby’s location. The Pi camera is triggered by a noise sensor. The Raspberry Pi Camera Board is a custom-designed Raspberry Pi add-on module. It connects to the Raspberry Pi through a customised CSI interface. In still capture mode, the sensor has a native resolution of 5-megapixels. It can capture video at resolutions up to 1080p at 30 frames per second in video mode. The Pi Camera is intended to keep an eye on the baby and monitor his or her movements. The video of the baby’s current position is taken with the Pi camera module. The Raspberry Pi camera is set up to appropriately capture the baby’s face and body.

The Raspberry Pi with attached camera positioned over the baby bed provides an input in the form of video frames. The video frames are sent to the motion detecting system, which employs two independent approaches, each of which is applied separately, and the results combined to improve accuracy. The output is a binary picture with white pixels representing the baby’s movements. To draw a point on the chart, the control chart construction mechanism computes this binary picture. The control chart is reviewed for anomalous behaviour for the entire interval (currently 5 frames per interval). When the majority of points cross the upper or lower barrier in a single interval, it is considered abnormal, and an alarm is generated. The baby’s movement is detected using image processing, and an alert is delivered to the parents via mail.

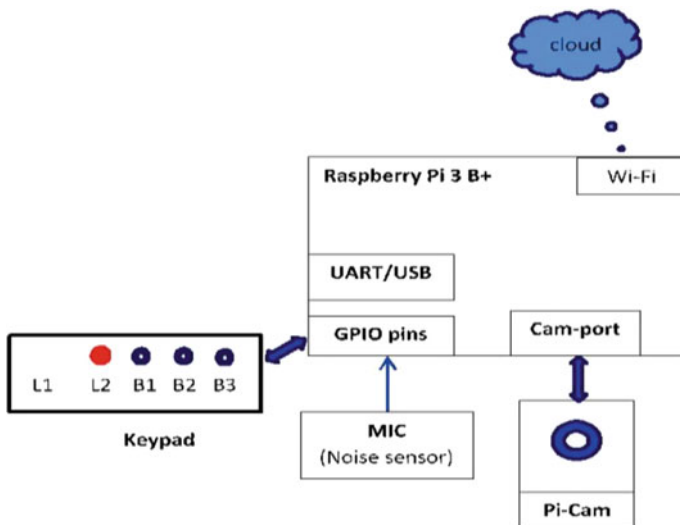


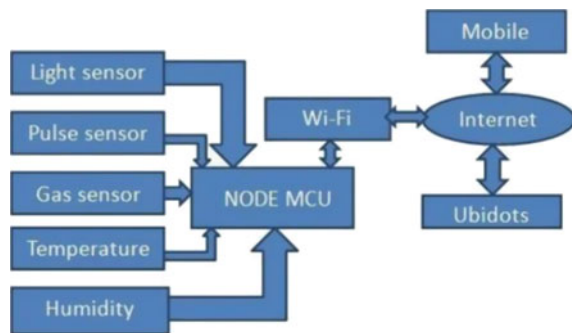
Fig. 3 Block diagram of Image Processing [10]

2.4 Smart Incubator

One of the necessary specialties in medicine is Pre-term infant care. An incubator is a biomedical device used for providing warmth and humidity to the baby. It is also used to regulate the input oxygen quantity [12]. The baby feels comfortable if the temperature of the room is kept warm. An incubator, therefore, requires a temperature monitoring system [8]. The infants might face hyperthermia, dehydration, and several other health issues with the increase or decrease of the body temperature. The air humidity also increases with temperature [3]. A huge number of death of infants annually clearly dictates the need for a thermal environment that is well-regulated for the survival of infants. The currently existing incubators are very expensive for developing countries [13]. Portable and cost-effective incubators can be manufactured using Phase Change Materials since they are suitable for the storage of thermal latent heat systems [14]. Incubators are used all over the world to reduce heat loss from the body of infants, thereby improving their survival [15]. The newborn babies can be monitored in real-time in the incubator using an embedded device. The data of the baby can also be accessed by doctors using computers through the internet [4]. With the advancements in incubator technologies, the death rates of infants are gradually decreasing by maintaining the temperature, blood pressure, heartbeat, and other parameters [16].

Figure 4 depicts an implementation with help of the Node MCU controller which consists of different types of sensors. Medical data can be stored in cloud storage and a Wi-Fi network. Ubidots is cloud storage that stores a large number of information and record using the internet of things. Cloud computing is defined as modifying, configuring, and gaining online access to the application. It provides more reliable, available, and updated services.

Fig. 4 Block diagram of Incubator [4]



2.5 *E-Health*

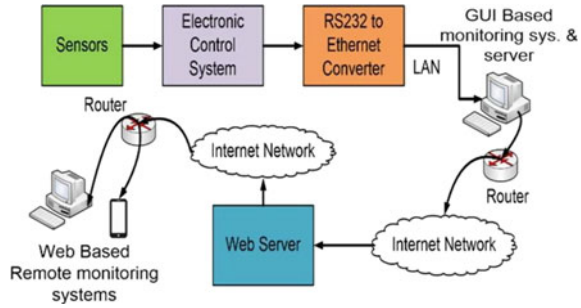
Information management is an important feature of a web-based infant monitoring system. The new baby information record can be created and updated by the doctors or administrators. The administrators can operate all data in the records [17]. The correct implementation of hospital management systems will contribute to an increase in the efficiency of healthcare. Digitalizing the data prevents the problems like difficulty in communication between departments and services [18]. The system consists of smart devices for real-time analysis of different parameters of the incubator. A set of modules are developed to help diagnose the doctors through telemonitoring of the baby [19]. The collection and analysis of data can be utilized to improve the survival of the baby. The information can also be used to identify any medical problem. A continuous monitoring system is a good solution for the observation of the baby [20]. The system can also include scheduling of vaccination for the baby and nutrition requirements [21]. The information regarding tests and diagnosis can also be provided.

2.6 *Temperature and Real-Time Monitoring*

Monitoring temperature is very important for the health care of newborn babies. The baby is comfortable when the temperature in the incubator is warm. The temperature should be monitored in real-time and made accessible to the doctor [8]. The humidity also increases with the temperature rise. The drift from the required temperature range will lead to issues regarding the health of the baby. These issues can also lead to the death of the infant. Therefore, it becomes crucial to monitor the humidity and temperature [22]. The temperature monitoring system can be based on a microcontroller. Many of the monitoring systems are offline but monitoring the baby in real-time is very efficient. The temperature is taken using a temperature sensor, and the information is communicated to the doctor [9].

Figure 5 depicts the system architecture of the health management system; it consists of different sensors for parameter observation. The LM35 sensor is used as the body temperature, SEN 1853 as air temperature sensor, Humidity sensor is SY-HS-220, DS-100A.5 as pulse oximeter, and HW01 as Heartbeat sensor. The body temperature sensor is accurate, linear and also has capability of self-heating. This system can transmit real-time and setpoint values using LAN. This provides the doctor access to remote monitoring at any given time. The Arduino Mega 2560 microcontroller is at the core of the control system, as sensors such as air and body temperature sensors, as well as a humidity sensor. The current system also includes a touch screen LCD for displaying and setting air and body temperature and humidity, as well as displaying set points for air and body temperature and humidity, which helps the doctor and nurse monitor the air and body temperatures and humidity of

Fig. 5 Block diagram of Real-time monitoring



immature infants locally. A buzzer serves as an alarm mechanism in the current system. It also includes a heater and heating control system, as well as an RS 232 to Ethernet converter and a custom-designed GUI / Web Page for remote monitoring. The below Table 1 lists the relevant literature works.

The present methods enable remote health monitoring of newborn with less accuracy and tolerance, according to the literature survey. The values of newborn health parameters differ from those of adults, such as heart rate, which is greater in newborns, more precise sensors must be employed. Newborns skin is also delicate; thus, the system design must be compact and lightweight so that it may be put in an optimum location. To improve the web application, big data analytics can be used to improve strategic assets in the health care industry by offering better services to patients, increasing patient satisfaction, and boosting customer relationships. A low-cost Bluetooth Low Energy (BLE) pseudo lite can be used to accurately find newborns participating in indoor activities and alert parents in a timely manner. It has a placement precision of more than 2 m and can significantly limit the risk of damage to newborns and young children [23, 24].

Table 1 Summary of relevant literature works

Paper	Web application	Inference
[1–3]	Internet of Things (IoT)	Remote monitoring is accomplished in these articles using smart phones or PCs, which give a low-cost way for remote monitoring with a very tiny tolerance when compared to the incubator’s built-in sensors. These papers do not cover the monitoring of other critical health markers
[5–9]	Embedded systems	To provide a secure, safe and healthy environment to the infant certain parameters need to be monitored very accurately. In these papers, limited monitoring techniques have been implemented

(continued)

Table 1 (continued)

Paper	Web application	Inference
[10, 11]	Image processing	This paper proposed a non-contact baby monitoring system that attempts to keep the infant from slipping out of bed in absence of a parent. However, the whole setup requires proper maintenance and many installations of packages to train the datasets. This system just acts as an alert system to detect the motion of the baby. It only detects if the baby is asleep or awake. It does not completely prevent the baby from falling
[3, 12, 13]	Smart incubator	In these papers, a customized incubator is being set up to provide a safe environment for the baby. For a fully smart incubator, the incubator needs to be made fully functional which could keep track of health parameters as well as could manage them. The current paper reflects the only implementation of the incubator but there are requirements of the smart incubators that could provide health information more accurately
[17–20]	E-health	These applications only provided registration of patients and maintenance of health records. To have a systematic e-health application, there should be tracking of other vital parameters, and accessing those parameters should be easy
[22]	Temperature and real-time monitoring	In these papers, the system cannot perform multi-tasking since it uses a real-time monitoring system. Also, it has complex algorithms and program crashes can be frequently experienced

3 Conclusion

In this paper, web application-based health management for neonatal is reviewed. Based on the above papers reviewed, it was observed that the system had low tolerance and does not provide a secure and safe environment to the infant. The literature papers reflect only the implementation of the incubator, but there are requirements of a smart incubator that could provide and keep track of the health parameters more accurately. There is a necessity for new techniques and methods that could provide accurate health information and also managing and accessing the health information would be easy and also be economical and affordable.

References

1. Nair A, Karthikeyan A et al (2021) IOT based neonat incubator. *Int J Eng Technol* 8(6):3182–3185
2. Wahab MA, Md Nor D (2021) Safety and health monitoring system for baby incubator using IoT. *Evol Electr Electron Eng* 2(2):256–264
3. Shabeeb G, Al-Askery AJ, Nahi ZM (2019) Remote monitoring of a premature infants incubator. *Indonesian J Electr Eng Comput Sci* 17(3):1232–1238
4. Sivamani D, Sagayaraj R, Ganesh RJ, Ali AN (2018) Smart Incubator using Internet of Things. *Int J Mod Trends Sci Technol* 4(9):23–27
5. Symon AF, Hassan N, Rashid H, Ahmed IU, Reza SMT (2017) Design and development of a smart baby monitoring system based on Raspberry Pi and Pi camera. In: 4th international conference on advances in electrical engineering, ICAEE 2017, vol 20, pp 117–122
6. Hashim F, Mohamad R, Kassim M, Suliman SI, Anas NM, Bakar AZA (2019) Implementation of embedded real-time monitoring temperature and humidity system. *Indonesian J Electr Eng Comput Sci* 16(1):184–190
7. Patil SP, Manisha RM (2014) Intelligent baby monitoring system. *ITSI Trans Electr Electron Eng* 2(1):11–16
8. Latif A, Arfianto AZ, Poetro JE, Phong TN, Helmy ET (2021) Temperature monitoring system for baby incubator based on visual basic. *J Rob Control (JRC)* 2(1):47–50
9. Parihar VR, Tonge AY, Ganorkar PD (2017) Heartbeat and temperature monitoring system for remote patients using arduino. *Int J Adv Eng Res Sci* 4(5):55–58
10. Dubey YK, Damke S (2019) Baby monitoring system using image processing and IoT. *Int J Eng Adv Technol* 8(6):4961–4964
11. Hussain T, Muhammad K, Khan S, Ullah A, Lee MY, Baik SW (2019) Intelligent baby behavior monitoring using embedded vision in IoT for smart healthcare centers. *J Artif Intell Syst* 1(1):110–124
12. Ihebuzo G, Ndubuka N (2021) Design and fabrication of a triplet baby incubator with renewable power source. *Afr J Med Phys Biomed Eng Sci* 45–57
13. Tran K et al (2014) Designing a low-cost multifunctional infant incubator. *J Lab Autom* 19(3):332–337
14. Yadav S (2018) Application of combined materials for baby incubator. *Procedia Manuf* 20:24–34
15. Megha K, et al (2018) Intelligent baby incubator. In: IEEE international conference on electronics, communication and aerospace technology, pp 1036–1041
16. Subramanian M, Sheela T, Srividya K, Arulselvam D (2019) Security and health monitoring system of the baby in an incubator. *Int J Eng Adv Technol* 8(6):3582–3585
17. Xiaosong L, Gang S, Yong C (2015) Design and development of the web-based health and chronic disease assessment management system. In: Proceedings - 2015 7th international conference on information technology in medicine and education, ITME 2015, pp 72–75
18. Calado MP, Ramos A, Fabiano D (20019) e health in hospital information management. In: 2019 IST-Africa week conference (IST-Africa), pp 1–7
19. Mukherjee S, Dolui K, Datta SK (2018) Patient health management system using e-health monitoring architecture. *IEEE*, pp 1–5
20. Najib SM, Hassan NB, et al (2018) Intelligent neonatal monitoring system based on android application using multi sensors. *IEEE*, pp 131–135
21. Bhavani S, et al (2017) Providing a friendly e-health care environment to rural women during pregnancy and child growth. In: IEEE international conference on technological innovations in ICT for agriculture and rural development, pp 215–217

22. Agresara NY, Vyas DD, Bhensdadiya BS (2016) System for remote monitoring and control of baby incubator and warmer. *Int J Futuristic Trends Eng Technol* 3:13–18
23. Sharma RR (2021) Design of distribution transformer health management system using IoT sensors. *J Soft Comput Paradigm* 3(3):192–204
24. Smys S, Joe CV (2019) Big data business analytics as a strategic asset for health care industry. *J ISMAC* 1(02):92–100

IEEE 802.11g Wireless Protocol Standard: Performance Analysis



G. U. Shreelatha and M. K. Kavyashree

Abstract Wireless networking technology is used to access networks or connect with other user devices in a wireless medium. Wireless Local Area Network IEEE 802.11g is one of the standards of wireless technology. IEEE 802.11g is a Wi-Fi 3 standard, which was designed in 2003. This has the properties of both 802.11a and 802.11b. This paper analyses the performance of this protocol standard with single access point. The effects of varying data rates are observed on the delay, media access delay, queue size, and maximum throughput are analyzed. Simulations were performed using OPNET. Along with these, the analysis of access points using HTTP is done to analyze object and page response time and further voice data analyses were performed to perceive jitter, packet end-to-end delay (time taken by the packet from source to destination), traffic sent, and traffic received between the nodes.

Keywords Hyper text transfer protocol · Wireless local area network · File transfer services · Wireless protocol

1 Introduction

Multiple access control (MAC) protocol and physical layer are implemented at the MAC layer in IEEE 802.11 standard. Infrastructure and ad hoc networks are supported by this standard. The performance of a wireless network depends on constraints which includes data rates buffer sizes threshold and throughput.

G. U. Shreelatha (✉) · M. K. Kavyashree
Electronics and Communication, JSS Science and Technology, University, Mysore, Karnataka, India
e-mail: shreelathagu1998@gmail.com

M. K. Kavyashree
e-mail: kavyashreemk@sjce.ac.in

1.1 WLAN- Physical Architecture

There are two types of architecture for WLAN:

1. Infrastructure less architecture
2. Infrastructure architecture

1. Infrastructure less architecture:

This architecture is the simplest of WLAN configurations which is an autonomous WLAN and it is also known as an ad-hoc network. This contains a set of computers grouped with the help of a LAN client adapter. No access point is required in this kind of configuration. The Fig. 1 below shows this architecture.

2. Infrastructure architecture:

In this type of analysis, there is an access point present. These access points are connected with a distribution system like ethernet. These access points determine the cell and also define a confined radius over which the network has access. When a device moves out of one cell to another the connection with the new access point is made. This whole process is called handover. The architecture is shown in the Fig. 2 below.

2 References

In [2], the authors analyze the service availability, network performance, and user connection.

In [11], the authors have used DCF+ method to determine the performance of the network. This model provides a correct behavior of IEEE 802.11.

In [6], the authors have used opnet simulator to analyze the performance of IEEE 802.11g WLAN's. Here they have studied the performance based on the number of workstations. The work is carried out on mobile as well as stationary stations to get a better understanding.

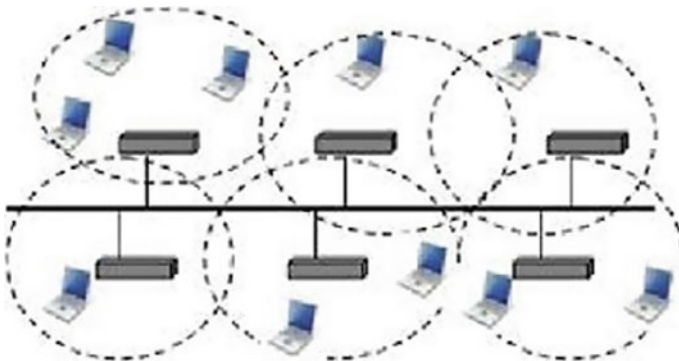


Fig. 1 Infrastructure less wireless network



Fig. 2 Infrastructure wireless network

3 Selection of Access Point

3.1 Delay Analysis

A delay can be described as the time that exists between the time from when the packet was generated at the source to the time at which the packet reached the destination. So, in simple words, it's the time that a packet takes to move in the network. The unit for this is seconds. This delay can be termed a packet end-to-end delay. This delay can also be caused due to latency. Latency refers to time taken by the packet from the required point to the destination. So, if this increases it implies that delay also increases.

The efficiency of network is determined by how fast the packet reaches the destination on time. Thus, if delay is minimum then the packet reaches the destination on time. So minimum delay means an efficient time flow.

3.2 Media Access Delay

This delay is described by the time from when the data reaches the MAC layer until it is successfully sent out to the wireless medium. This is important because in a real-time network if the data is not received within the required delay, then the data becomes useless. This delay should be as minimum as possible to provide an efficient real-time flow.

3.3 Queue Size

To study the performance parameters of a computer system queue size is the best. This will help provide explanations as to some of the issues like bottlenecks or compromised computer performance caused due to waiting for too long for the data to arrive.

3.4 Throughput

This is the most important parameter which is used to measure the amount of data that is transferred from source to destination within the given time frame. This can also be termed as the capacity of the network in bits per second or data per second. The throughput of a network helps in understanding the speed of the network.

Throughput gives the data transmitted per second. So if the throughput is high it can be concluded that the speed of the network is high/.

4 Software Requirement

OPNET simulator is one of the best simulators which provides a higher power and versatility. Due to its excellent features OPNET is chosen as the simulator. OPNET SIMULATOR

Opnet stands for Optimized network Engineering Tools which is a network simulator tool that is used to simulate the behavior and performance of any kind of network. Compared to other simulator tools Opnet has better power and versatility. This tool can be used to simulate a large area network with various protocols. This tool was built for military usage, but it has grown indefinitely for commercial simulation tools for networks. Though, it is an expensive software free licenses are available for the educational process. This tool defines network topology, nodes, and links that form the network. The processes that take place in a particular node and transmission link can also be user-defined. Opnet has a vast library set that provides simulation, and analysis of the network to compare the effect in different scenarios and the IDE of OPNET is user-friendly which makes the user use it with no issue.

5 Network Design

In this paper, a network is designed in such a way that there is one access point along with 15 nodes attached to it. The parameters considered here are throughput and delay.

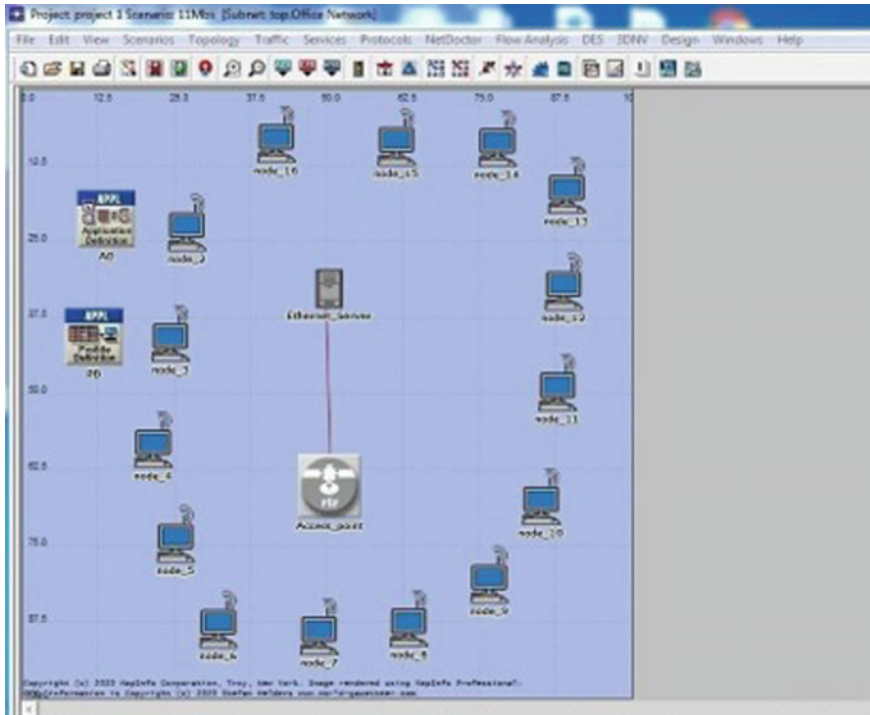


Fig. 3 Simulation model structure

The Fig. 3 above shows the network design.

6 Simulation Results

6.1 Delay Study

For the 'delay' parameter a study of 4 scenarios is considered and also when the data rate is increased from 1 to 54 Mbps, the delay decreases. The result shown below proves that as the data-rate increases the delay decreases. The Fig. 4 shows the delay comparison.

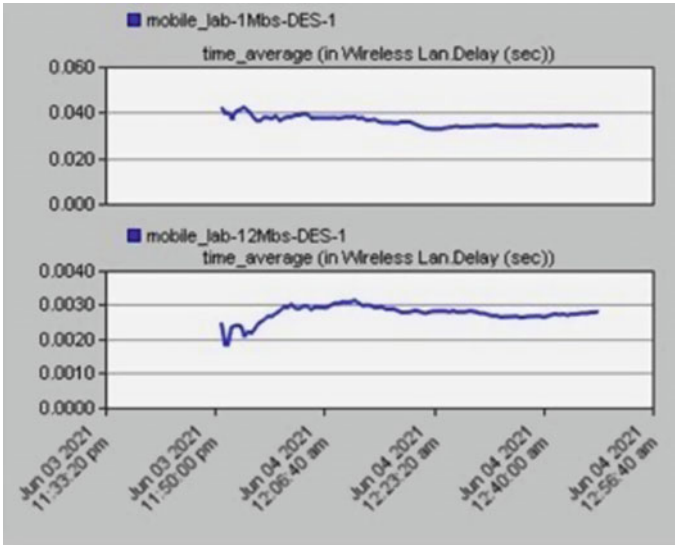


Fig. 4 Delay comparison between 1 and 12 Mbps

6.2 Media Access Delay Study

As the data rate raises from 1 to 12 Mbps, the media access delay is decreased. Based on the graphical result shown below it is proved that the media access delay decreases with the increase in data-rate.

6.3 Queue Size Study

As the data rate raises from 1 to 12 Mbps, the queue size decreases. Theoretically, as the data-rate increases, the speed of data transmission from source to destination increases and the queue decreases. The graph below Fig. 6 proves the same in a practical scenario.

6.4 Throughput Study

When the data rate is increased from 1 to 12 Mbps the throughput increases. So as the data-rate increases the number of bits received also Figs. 5 and 6.

Increases thereby increasing the throughput of the system. The graph for the same is shown in the below Fig. 7.

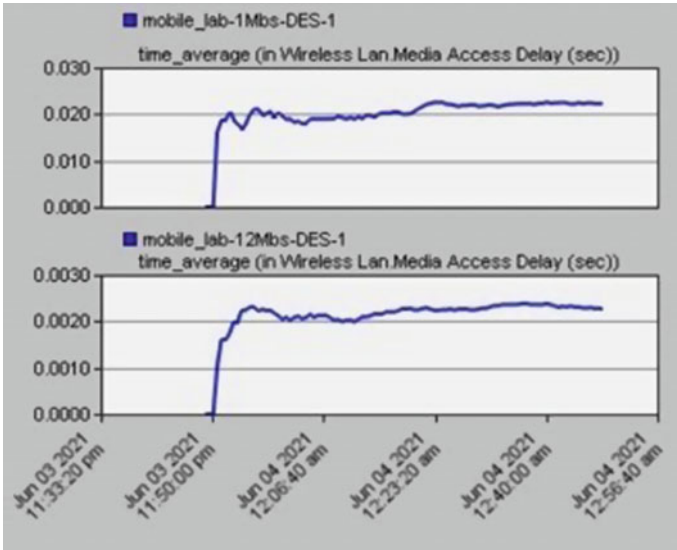


Fig. 5 Media access delay

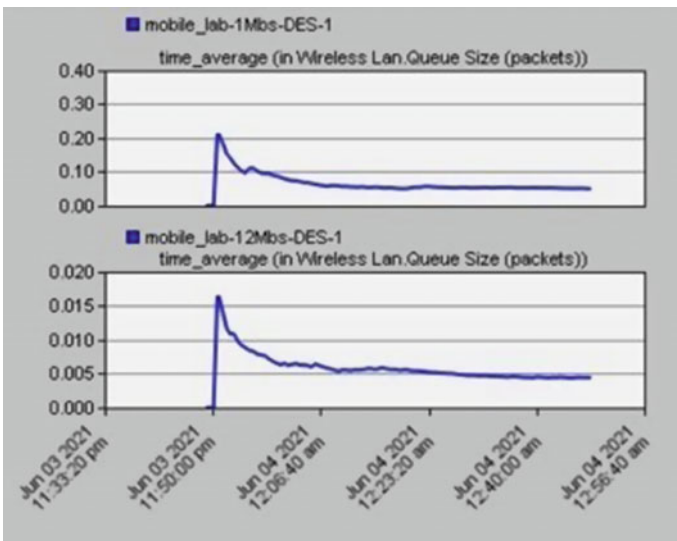


Fig. 6 Queue size comparison between 1 and 12 Mbps

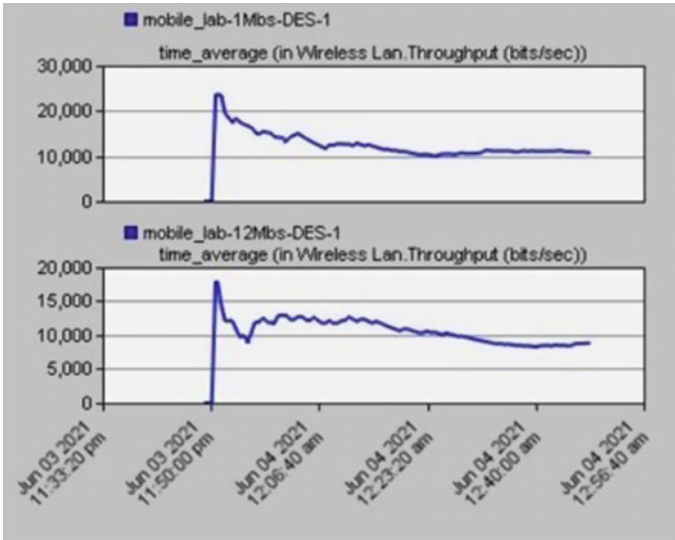


Fig. 7 Throughput comparison between 1 and 12 Mbps

7 Analysis of Access Point Using HTTP

Object Response time: provides the time that is taken to respond for every individual inlined object from the HTML page.

Page Response time: provides the time required to regain the entire page with all the contained inline objects as shown in Fig. 8.

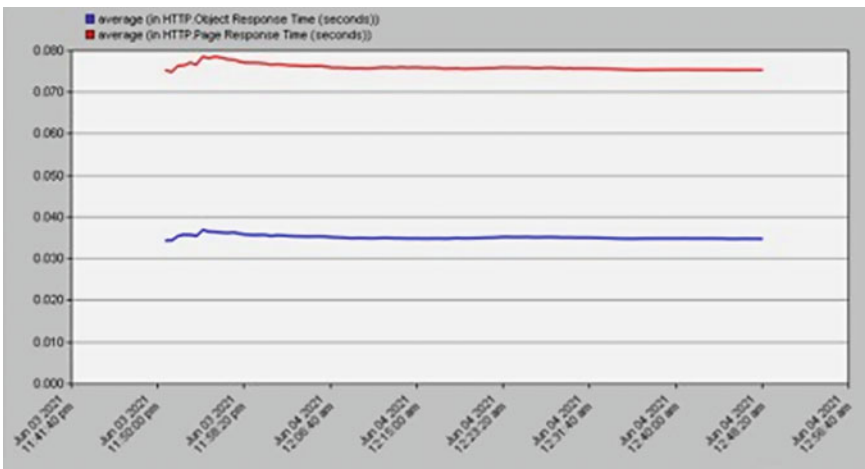


Fig. 8 Object and page response time

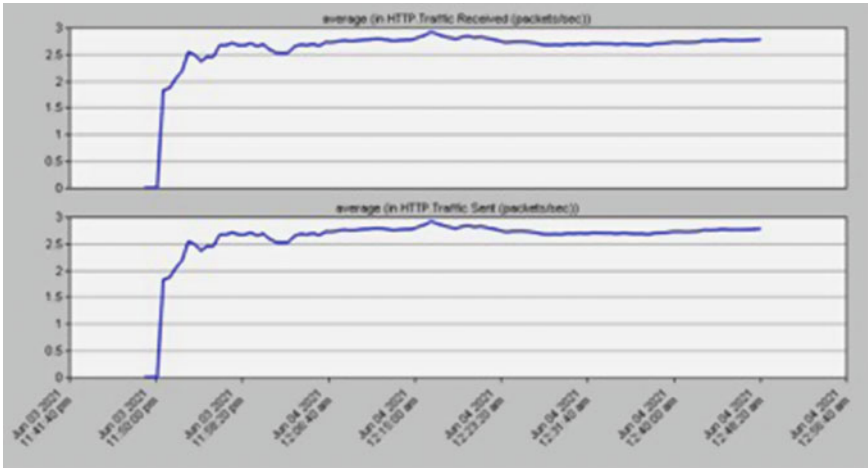


Fig. 9 Traffic is sent and received in the network

Figure 9 shows the traffic is sent and received in the network.

8 Analysis of Voice Application Based on Different Parameters

1. Jitter study: When two packets start at the destination point say at times t_1 and t_2 and reach the destination at a time t_3, t_4 , then the jitter is expressed as $\text{jitter} = (t_4 - t_3) - (t_2 - t_1)$. When the time difference at the destination node is less than that at the source node a negative jitter is formed.

From the above Fig. 10 we can see that different nodes experience different jitter. This is due to a packet getting queued or delayed anywhere in the network, where there was no delay or queuing for other packets.

2. Packet end-to-end delay:

Network delay is the delay caused when the sender node gives the packet at RTP and the receiver gets it at RTP. The encoding and decoding delay is computed from the encoder scheme and the latter is considered as same as that of the encoding delay.

From the above Fig. 11, it is observed that the Packet end-to-end delay is different for different nodes. This is due to variation in delay jitter experienced at different nodes in the network.

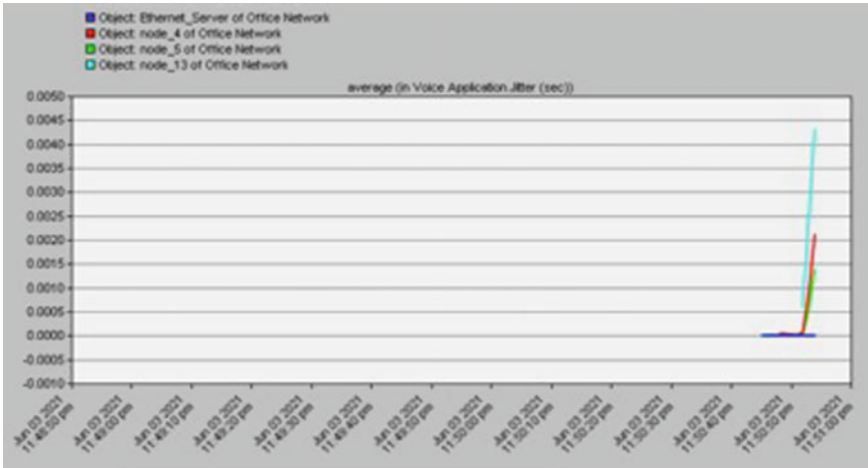


Fig. 10 Jitter in voice application

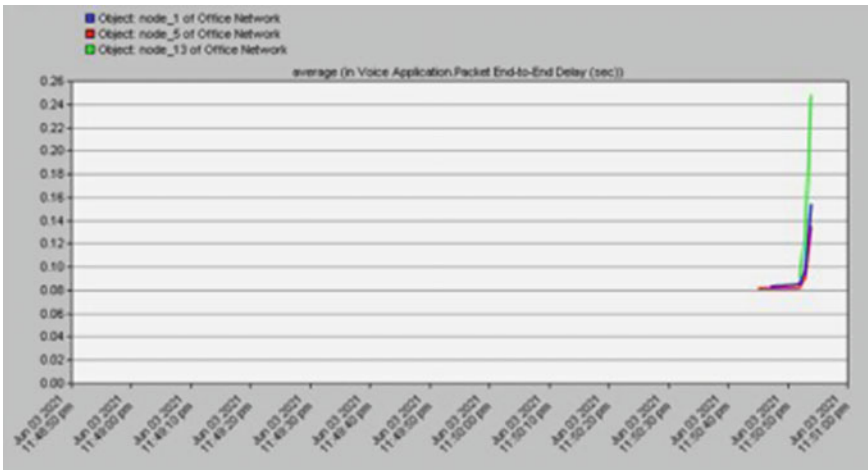


Fig. 11 Packet end-to-end-delay in Voice application

- 3. MOS Value: This gives the quality of voice and video sessions. From the above plot, we can see that the MOS value for voice applications varies for different nodes due to variations in jitter, latency, and packet loss are shown in Figs. 12, 13 and 14.

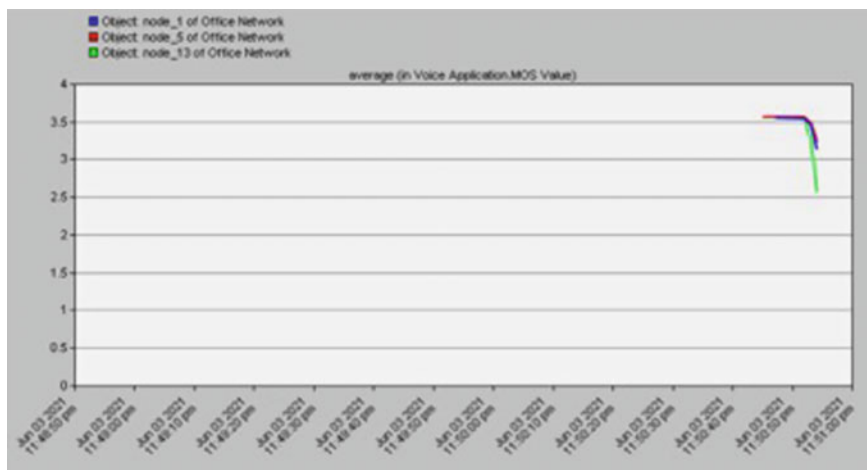


Fig. 12 MOS value of voice application

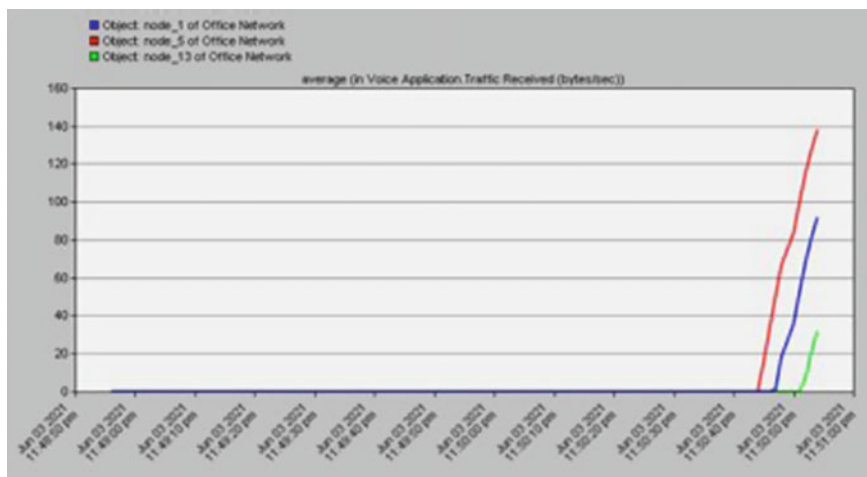


Fig. 13 Traffic received in voice application (bytes/sec)

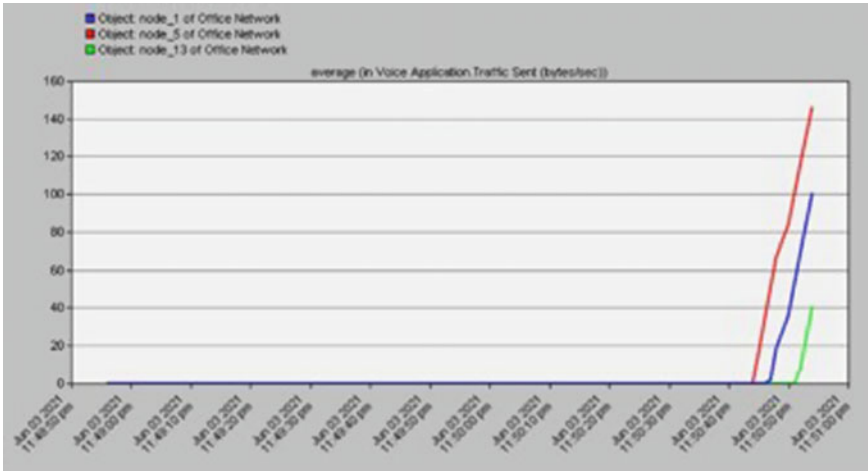


Fig. 14 Traffic sent in voice application (bytes/sec)

9 Conclusion

In most of the research, the performance was analysed based on the number of workstations in the network. However, here in this paper the discussion is made on a wireless network based on various parameters. The paper here shows the design of the wireless network. The major challenges in such types of the network include delay, media access delay, queue size, and throughput. From simulations performed, the observations clearly show that when data rate is changed it effects other parameters also. From the results one can observe that when the data rate is raised the delay, Media access delay, and Queue size decrease whereas the throughput increases. This indicates that the data is delivered accurately and efficiently. Some of the parameters like MOS value, jitter, and packet end-to-end delay of Voice application are observed and found that they are interdependent.

References

1. Hamdan YB (2021) Construction of statistical SVM based recognition model for handwritten character recognition. *J Inf Technol* 3(02):92–107
2. Raj JS, Vijesh Joe C (2021) Wi-Fi network profiling and QoS assessment for real time video streaming. *IRO J Sustain Wirel Syst* 3(1):21–30
3. Karuppusamy P (2020) Effective test suite optimization for improving the coverage standards using hybrid wrapper filter memetic algorithm. *J Soft Comput Paradigm* 2(2):83–91
4. Dasgupta S, Roy PJ, Sharma N, Misra DD (2020) Application of IPv4, IPv6 and dual stack interface over 802.11ac, 802.11n and 802.11g wireless standards. In: Third international conference on advances in electronics, computers and communications (ICAIECC) (2020)

5. Zuo PL, Peng T, Wu H, You K, Jing H, GuoW, Wang W (2020) Suppression of 802.11 transmission in 2.4 GHz ISM band: method and experimental verification. *IEEE Xplore*
6. Karki M, Shrestha A, Singh VL (2017) Performance comparison of IEEE 802.11g WLANs with respect to increasing number of workstation using OPNET modeler. In: International conference on computing and communication technologies for smart nation
7. Wahyuda DV, Achmadi YD, Sari RF (2017) Comparison of different WLAN standard on propagation performance in V2V named data networking. In: IEEE Asia Pacific conference on wireless and mobile
8. Wang T, Refai HH, Wang Q (2004) Performance analysis of the WLAN based on IEEE 802.11a/b/g standards in the presence of an interfering cordless phone. In: Henty BE, Rappaport TS (eds) *IEEE 1st wireless and optical wireless network conference*
9. Chatzimisios P, Boucouvalas AC, Vitsas V (2003) Influence of channel BER on IEEE 802.11 DCF. *Electron Lett* 39(23):1687–1689
10. Wu H, Cheng S, Peng Y, Long K, Ma J (2002) IEEE 802.11 distributed coordination function (DCF): analysis and enhancement. In: *IEEE international conference on communications (ICC)*, vol 1, pp 605–609
11. Wu H, Peng Y, Long K, Cheng S, Ma J (2002) Performance of reliable transport protocol over IEEE 802.11 wireless LAN: analysis and enhancement. In: *Proceedings of INFOCOM*, vol 2, pp 599–607
12. Bianchi G (2000) Performance analysis of the IEEE 802.11 distributed coordination function. *IEEE J Sel Areas Telecommun Wirel Ser* 18:535–547

Optimal Cluster Head Selection Using Vortex Search Algorithm with Deep Learning-Based Multipath Routing in MANET



S. Venkatasubramanian

Abstract The sensor network is Mobile Ad hoc Network (MANET), which is susceptible to node or connection failure. There may be a network disconnection if the link fails, so data cannot reach the Base Station or a sink node. In a hostile environment, a connection failure reduces the network's quality. The Vortex Search algorithm established an effective Cluster Head (CH) selection mechanism and found the multipath routing using a weighted deep learning model to alleviate this issue and tolerate network loss. Dynamic Multipath Routing Protocol (DMPRP) with energy measure is used in the suggested approach for an effective CH selection procedure. To upsurge the presentation of routing without increasing the amount of packet overhead, a multipath routing protocol based on projected probability sites and path diversion is presented. Deep learning procedures are used to identify the most likely locations of nodes, which are combined by using a weighting function. Finally, the route maintenance team is called in to monitor data packet delivery and report any link failures.

Keywords Mobile ad hoc network · Deep learning · Cluster head selection · Vortex search algorithm · Multipath routing

1 Introduction

The MANET has become a popular choice for an extensive range of applications thanks to the advancements in wireless communication [1]. MANET [2] is a framework that uses nodes or strategies that are mobile to create a powerful network that does not trust any substructure [3]. Due to their unique properties, mobile adhoc networks face several issues in routing since nodes take on both the role of hosts and routers in the network. As a result, MANET's architecture is heavily influenced by the routing algorithm [4]. Routing is the challenge of determining paths to deliver data packets or items from sources to destinations while achieving the QoS requirements

S. Venkatasubramanian (✉)
Saranathan College of Engineering, Trichy, India
e-mail: veeyes@saranathan.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
P. P. Joby et al. (eds.), *IoT Based Control Networks and Intelligent Systems*,
Lecture Notes in Networks and Systems 528,
https://doi.org/10.1007/978-981-19-5845-8_7

under numerous limitations on systems, and it is of critical relevance in networked systems [5]. As a result of its inherent difficulty and the enormous benefits that may be gained by efficient routing, routing has recently become a major research topic. Network topology, node processing capacity, traffic pattern, and other factors influence routing strategy performance metrics such as throughput, packet delivery rate, and link usage ratio [6]. A networked system is commonly described as a weighted graph, where the weights are connected to node and connection parameters, with but not limited to latency, capacity, dependability, and energy, to evaluate and improve routing algorithms. Because of its simplicity and efficiency, shortest route routing (SPR) is an often-used routing method. When numerous constraints are met, it is difficult to discover the least expensive path through the graph [7].

Both in wired and wireless networks, many heuristic routing policies have been put forth to deliver less-than-ideal results. There are few link weight adjustments required to get desirable performance in practical IP networks when using the SPR protocol [8]. By minimizing the greatest generalized node betweenness centrality, routing methods are developed for large-scale complex networks [9]. The use of multipath techniques, such as equal-cost multipath [10] and optimization of link weights (OLW) [11], in multihop networks, where packets are delivered to their destinations over several channels, increases throughput and end-to-end reliability. It is possible to attain the same level of performance with a generalized destination-based multipath routing (GDMR) technique [12]. As a result, issues like energy balance and multipath cooperation arise when using multipath routing. Many networks lack the necessary infrastructure for measuring traffic and the statistical characteristics of communication channels, making it impossible to provide an accurate picture of traffic demand [13].

Deep reinforcement learning (DRL) has been presented to signify a vast state-action space and perform automatic feature extraction in a high-D space, inspired by the recent success and comprehensive research on hybrid neural networks [14–17]. Nodes' projected locations are taken into account when adjusting the routing. No forecasting algorithm can be 100% accurate. There is continuously some ambiguity in prediction. When the chance of uncertainty is significant, the routing protocol built on the prediction protocol fails to work. As part of this study, a location prediction model is built using node location and node relative mobility. The model provides results in terms of probable locations. Weighted Fusion, a strategy for combining predictions from several models, forms the foundation of deep learning. Multipath routing protocols are fine-tuned to route packets based on the projected location. Only in cases where the position prediction uncertainty probability is substantial, does the multi-path forward decision depart from the end-to-end multipath. Multi-path routing's packet overhead is reduced as a result of this. The use of many parameters for hybrid modelling to account for historical node movements, changes in the relative location of nodes with neighbours, and changes in node location over time intervals is an important aspect of the approach. This multi-parameter modelling improves the accuracy of the forecast for various mobility models. To validate the proposed model's performance, Vortex Search Algorithm uses the DMPPR to identify the best CH nodes.

The following is how the respite of the paper is put together: An overview of MANET and routing is provided in Sect. 1. Section 2 focuses on existing strategies for CH selection and multipath routing. Section 3 gives a mathematical equation account of the projected methodology. Experimentation is illustrated in Sect. 4 to test the projected method's efficiency. Section 5 provides a final statement.

2 Literature Review

Using the table-driven and on-demand routing protocols, Majd et al. [18] examine the MANET's energy consumption, routing overhead, throughput, and delay to determine how well it performs.

Another investigation by Malar et al. [19] has offered bio-inspired approaches to enhance MANET routing. They've employed Ant colony optimization to create an energy-efficient MANET routing scheme and improved the way the network makes use of its available power. K-means algorithm and artificial neural network are used to solve the cluster formation and head selection problems, based on the current state of knowledge.

Firefly algorithm by Darwish et al. [20] is presented in the paper to identify the best routing path in wireless ad hoc networks. As a result of Li et al. [21] article, information can be sent with little delay and hops in the VANET. For ad hoc networks, the author Ghasemnezhad et al. [22] suggest utilizing fuzzy logic to determine the quickest route. For the internet of things, Thangaramya et al. [23] propose a CNN trained fuzzy network that provides the shortest route for WSN. [24] The author introduces the PSO-based geographical routing method, which he claims can outperform geographical routing in terms of network lifetime.

To choose CHs and non-cluster heads (NHS), Prakash, V. [25] developed a new DMGRP based on the M-PSO method. The Genetic Algorithm was used to discover the optimally shortest path after calculating the probabilities and selecting the best CH. GA uses a network-based objective function to select the best path. A hybrid model with weight fusion and DMGRP routing is employed in the proposed method, but the CH is determined using VSA and multipath routing is taken into account. However, the shortest path in [25] is determined using GA.

Uses a hybrid model of Polynomial Regression, AR modelling, Mobile Pattern corresponding, and Link stability prediction to estimate the probable locations of nodes [26]. A multi path routing protocol based on estimated probability locations with path diversion at necessary places along path is proposed for improved routing performance without larger packet overhead. But, the proposed model uses only two deep learning techniques for multipath routing and considered the CH selection using VSA.

A wireless power transmission using SDWSN has been proposed by anand [27]. We've devised a way for locating the transmitters with the least amount of energy consumption by putting the transmitters inside the node itself. The location of the energy transmitters is determined by a trade-off between the fair distribution of

energy and the highest amount of energy charged in the network. By defining a utility function, this approach aims to improve fairness and the total amount of energy charged. The transmitters must use less energy while the sensor nodes must preserve their charge. According to the simulation results, the proposed technique consumes the least amount of energy while still delivering the greatest amount of utility in terms of power, number of tasks, energy transmitters, and fairness.

2.1 Objectives Identified and Proposed

- To reduce power consumption and optimize network lifetime for optimal cluster head selection, and effective multipath with cluster optimization model was developed.
- Using the Vortex Search method, the optimal cluster head selection has been made based on energy consumption.
- A weighted deep learning model was applied to Vortex Search selection results of cluster heads to determine the optimized shortest path.

3 Proposed Methodology

To communicate with a destination node in MANET, source nodes broadcast or receive control posts like route appeal and route reply to other nodes that are connected to it. Until a path to the destination node is found, this activity continues. There is a lot of overhead in the conventional routing strategy when networks have a large number of nodes or are very dense. Rather than relying on individual nodes to commence routing, clusters of nodes are used to reduce the additional burden. Finally, routes have been established and data will be transferred through the cluster node instead of individual nodes. When the mobile nodes are grouped into clusters, a cluster leader is chosen based on how well the cluster performs. The cluster head is responsible for determining which nodes are part of the cluster. Cluster head design can degrade performance in a dynamic network due to frequent cluster head selection [28]. The network model and energy consumption parameters are specified before the selection procedure.

3.1 Network Ideal and Energy Ideal

The methodology outlined in this area aims to increase wireless sensor network energy efficiency by organizing Cluster Heads properly and making them more scalable for use in a variety of network sizes. The clustering procedure, an algorithm for selecting CH nodes, and a routing algorithm to uphold power utilization and extend

the network's lifespan are the primary points of attention here. Our suggested DMRP includes the following network assumptions:

- Every single one of the sensor nodes is the same in shape and size.
- All nodes remain stationary [25], when organized in the field.
- There is only one base station.
- Source routing is used by all nodes, which all have the data to transmit. This protocol's routing features include route discovery and management.
- As part of the route discovery process, the source node will send out route request and reply (RREP) messages.
- A route reply (RREP) message can only be sent back to the sender by the destination node.
- Using this method, the distance between sender and recipient can be cut in half.
- Provide several paths to the destination, subsequent in load balancing and improved network presentation.
- In the event of a routing protocol failure, an alternative route is provided by the various routing protocols.

3.2 Energy Model

Sensor nodes' energy consumption is mostly based on the transmission, amplification, and reception of radio sensors' electronic energy. The distance between the reception node and the transmitter node is an extremely critical characteristic that should not be overlooked. If the propagation distance is smaller than the threshold distance t_0 , the system uses a free-space or multipath disappearing communiqué channel. Otherwise, the amount of energy consumed is directly proportional to d^4 . To send an l-bit packet via a network, First Equation determines the maximum amount of energy that can be used:

$$E_{con}(l, d) = \begin{cases} lX E_{elec} + lX E_{fs} X d^2, & d < t_0 \\ lX E_{elec} + lX E_{mp} X d^4, & d \geq t_0 \end{cases} \quad (1)$$

There are two types of energy model amplifier coefficients for transmitting one-bit sensor node E_{con} data: free-space and multipath fading (abbreviated E_{mp}). The distance t_0 calculated by the Second Equation is the threshold distance.

$$t_0 = \sqrt{\frac{E_{fs}}{E_{mp}}} \quad (2)$$

While E_{fs} and E_{mp} are both parameters. When $d < t_0$, the E_{fs} is the amplifier stricture used for the free-space energy ideal usage by the nodes when $d < t_0$ and $> t_0$, A multipath fading energy model amplifier parameter, E_{mp} , has been used by

the sensor nodes. Sensor nodes cannot transmit more than t_0 , which ensures that all nodes in the clusters are within the planned transmission variety.

3.3 Cluster Formation

For efficient use of resources, cluster-based architectures employ cluster members (CM), CHs, and other non-cluster heads (NCHs). This CH grouping of nodes is managed by a structured head node, which receives data from the cluster members and transfers it to the base station for further processing and storage therein. The weight of a CH may be linked to the node's ability to handle additional tasks, such as being near BS or having many neighbours. Neighbour node (neighbour ID, cluster head IP), Cluster member (ID), and Node residual energy can all be taken into account when calculating (NCH-IP, cluster gateway-IP). With this information, we were able to establish which nodes were close to our base station location. Using a network's multi-object function, we may calculate the k-optimal clusters in the mathematical representation.

The query is known as Expectation–Maximization after using the k-means approach.

Assigning data points to the nearest cluster is the job of the E-step.

$$\begin{aligned} \frac{\partial y}{\partial x} &= \sum_{i=1}^m \sum_{k=1}^k \|x^i - \mu_k\|^2 \\ = w_{ik} &= \begin{cases} 1 & \text{if } k = \operatorname{argmin}_j \|x^i - \mu_j\| \\ 0 & \text{otherwise} \end{cases} \end{aligned} \quad (3)$$

The M-step shall work out every centroid of the cluster.

$$\begin{aligned} \frac{\partial y}{\partial x} &= 2 \sum_{i=1}^m w_{ik} (x^i - \mu_k) = 0 \\ \mu_k &= \frac{\sum_{i=1}^m w_{ik} x^i}{\sum_{i=1}^m w_{ik}} \end{aligned} \quad (4)$$

That can be mathematically solved as described in the following section.

It has a multi-purpose function:

$$J = \sum_{i=1}^m \sum_{k=1}^k w_{ik} |x^i - \mu_k|^2 \quad (5)$$

where $w_{ik} = 1$ fits to cluster k , for data point x^i ; and $w_{ik} = 0$. μ_k is also the centroid of a cluster of x^i . It's a two-part issue. Regarding w_{ik} , we first decrease J and treat

μ_k set. After that, we reduce J with $_k$ and correct w_{ik} . We next split J from $_k$ and rebuild the centroid after the previous step's cluster assignments (M-step). In other words, square the distance between x_i and the nearest cluster's centroid to apply x_i as a data point.

$$\frac{1}{m_k} \sum_{i=1}^{m_k} ||x^i - \mu_c^k||^2 \quad (6)$$

3.4 CH Selection

Selection criteria for cluster heads include the following in particular: "Neighbour node" (neighbour node IP), "Cluster-head" (CH) "(cluster member IP)", and "Neighbour cluster" (NCH-IP, cluster gateway IP). The energy consumption ratio (ECR) for each of the CH is measured using DMRP's cluster selection:

$$ECR(m) = \frac{E_0}{E_0 - E_r} \quad (7)$$

DMRP computes their residual energy broadcast after scheming the ECR:

$$RETR(m) = \frac{E_r}{ECR \times d_{toBS}} \quad (8)$$

where d_{toBS} is the distance of cluster from BS. If we add ECR, the RETR will be

$$RETR(m) = \frac{E_r}{E_0/(E_0 - E_r) \times d_{toBS}} \quad (9)$$

To calculate DMRP's current round value, ECR and RETR are combined into one value for a specified duration of the current round. Riemann sum is employed by total power degradation over time because the rate of energy ingesting power $p = E_r / T$. As an approximation, this ECR can estimate the overtime period of an instance ($E_{instance}$).

$$E_{instance} - ECR = \sum_{r=0}^{R_{instance}} p(t_i) \Delta t_i \quad (10)$$

where $E_{instance} - ECR$ is exclusively E.C.R. The energy consumption combined over a time

$$E_{instance} - ECR = \int_{r=0}^{R_{instance}} p(t) dt \quad (11)$$

When DMGRP is used to calculate RETR, BS chooses a collection of relevant nodes that consumes less energy and has an optimal residual amount. Choosing a cluster as a CH node for the central region's node, where the distance is estimated using VSA, is also influenced by the distance factor.

3.4.1 Distance Calculation Using Vortex Search Algorithm (VSA)

An algorithm is known as VSA [29], a metaheuristic optimization algorithm activated by the vertical flow of stirred fluids. Like other single-solution algorithms, this method is, it uses streamlined generation steps. With the help of values, any generation of VSA populations can be transformed into a modern single solution. Furthermore, a crucial part of displaying a single-solution is the efficiency with which each iteration pass's update and seek are performed in the search area. This stability is achieved within the suggested VSA by the use of a vortex-like search pattern. Some of the nested circles simulate some of the tactics of vortex sampling.

Producing the Initial Solution

The preliminary process initials 'center/BS' μ_0 and 'radius' r_0 . In this phase, the early 'center'(μ_0) can be designed using Eq. (12).

$$\mu_0 = \frac{\text{upperlimit} + \text{lowerlimit}}{2} \quad (12)$$

where upper and lower limits are the bound restraints of the BS, which can be distinct in a vector of $d \times 1$ dimensional space of a network. In addition, σ_0 is the early *radius* r_0 generated with Eq. (13).

$$\sigma_0 = \frac{\max(\text{upperlimit}) - \min(\text{lowerlimit})}{2} \quad (13)$$

Generating the Candidate Solutions: The development of populations $C_t(s)$ in any number of iterations is the goal of the technique of developing candidate solutions. Using a Gaussian distribution, $C_0(s) = \{s_1, s_2, \dots, s_m\}$, $m = 1, 2, 3, \dots, n$ signifies the solution and n denotes the total sum of candidate solutions, which is used to form the VSA. The total number of potential solutions is given by the number n . The multivariate Gaussian distribution equation is exposed in Equation (Eq). (14)

$$p(x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp \left\{ -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right\} \quad (14)$$

While d signifies the dimension, x is the $d \times 1$ vector of the random variable, μ is the $d \times 1$ vector of the sample mean (i.e., center), and Σ is the covariance matrix, which shows the covariance matrix in Eq. (14). Then, Eq. (15) shows that a spherical

distribution is produced when the values' diagonal elements (i.e., variances) are equal and the off-diagonal elements (i.e. covariance) are equal to zero. Equal variances with zero covariance are used to calculate the value.

$$\Sigma = \sigma^2 \cdot [I]_{d \times d} \quad (15)$$

Distribution variance is depicted as σ^2 , and identity matrix is given as $d \times d$ as well as mentioned in Eq. (13), σ_0 is the initial radius (r_0).

Extra of the Current Solution: For the selection process, the current solution is replaced. There are several solutions to this problem, but only one is chosen and memorized from $C_0(s)$. This answer is used to replace the existing circle center. Before the selection process, the candidate solutions must be checked to ensure that they are located within the search spaces.

$$s_k^i = \begin{cases} \text{rand.}(\text{upperlimit}^i - \text{lowerlimit}^i) + \text{lowerlimit}^i, & s_k^i < \text{lowerlimit}^i \\ \text{rand.}(\text{upperlimit}^i - \text{lowerlimit}^i) + \text{lowerlimit}^i, & s_k^i > \text{upperlimit}^i \end{cases} \quad (16)$$

In this case, the numbers k and I are distributed equally, and the term "random" denotes the distribution of the numbers. To find the next best solution, VSA switches to utilizing s as a new center and Eq. (14) to shrink the vortex. It is, therefore, possible to generate a brand-new set of solutions, $C_1(s)$. If the chosen answer is better than the best solution, it can be regarded as the new best solution and memorized if it is superior.

The Radius Decrement Process: To reduce the radius at each iteration, the VSA uses the inverse incomplete gamma function. There are numerous applications of Eq. (17)'s incomplete gamma function in probability theory, particularly the chi-square distribution.

$$\gamma(x, a) = \int_0^x e^{-t} t^{a-1} dt > 0 \quad (17)$$

where $a > 0$ is the shape parameter while $x \geq 0$ is a random variable. Alike to the incomplete gamma function, its balancing $\Gamma(x, a)$ is usually also presented (Eq. (18)). In Eq. (18), $\Gamma(a)$ is a (1).

$$\Gamma(x, a) = \int_0^\infty e^{-t} t^{a-1} dt > 0 \quad (18)$$

Pseudocode of VSA algorithm is shown in the below Table 1. By using the distance between BS to clusters from VSA, the best CH is selected and then multipath routing is occurred to transfer the data.

Table 1 Pseudocode of VSA algorithm

Initializing step	
❖	Algorithm parameters: Input the population size, the lower and upper
❖	Fitness of best solution
❖	Center of the circle (μ_0), Eq. (12)
❖	The radius of the circle (σ_0), Eq. (13)
Repeat	
❖	Create candidate solutions within the circle by Eq. (14)
❖	If Exceeded, then shift values into the boundaries by Eq. (16)
❖	Select the best solution to replace the current center
❖	Decrease the standard deviation (radius) for the next iteration by Eq.
End Output The best solution found so far S_{best}	

3.5 Multipath Routing

The parameters for predicting node position and neighbourhood change are collected and sent to the two classifiers in the proposed model, and the results from the respective classifier are fused using weighted fusion. A probable location presence map is constructed based on the expected position of nodes. This map provides the nodes' probable locations at a given moment. When a connection is unavailable, the proposed multipath routing analyses this information to determine the link's life span and replaces the links with the next reliable path. The hybrid model takes into account changes in node location, node velocity, node relative site, and node neighbourhood across time. A probabilistic map of location prediction is constructed using a hybrid model that combines the findings of two different models in a weighted fusion. It's possible to fine-tune the multi-path routing method to determine the propagation at each of the routing stops using the prediction map.

There are two outputs from the hybrid model such as the likelihood of a node's location and the change in its immediate surroundings. The following inputs yield both of these outcomes.

1. Location of Node over time
2. Node comparative location with its neighbourhood
3. Velocity and direction over time

This model has two separate prediction models, which are fused using the Weighted Fusion approach.

3.5.1 Deep Neural Network (DNN)

DNN has input and output layers, as well as several further layers that are not visible. DNNs can solve linear and non-linear problems by using the appropriate activation function to compute the probability of each output layer. DNNs are commonly used

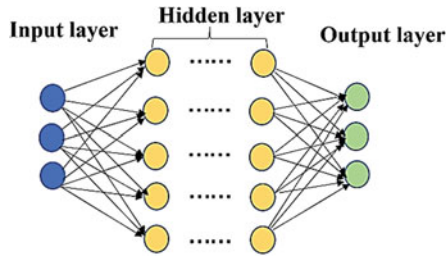


Fig. 1 Deep neural network structure

in image and speech recognition, as well as other applications. Full-connected neural networks are what DNNs essentially is. Multi-layer perceptron is another name for a deep neural network (MLP). The hidden layer transforms input feature vectors, which ultimately reach the output layer, where they are finally classified. There were only two categories in the linear classification model and its classification abilities were very limited, hence it was mostly utilized for linear classification. It is possible to employ various continuous functions to avoid this difficulty, such as the tanh function or the sigmoid function, which are discrete transfer functions. The number of neurons and hidden layers can be added to create DNNs.

A fully linked deep neural network is depicted in Fig. 1. The first and last layers are both input and output layers; the middle layer is a concealed layer that does not appear to be part of the first or second layers. To train a deep neural network, the following steps must be completed: initialization, forward propagation, layer transmission between layers, error propagation, and finally backpropagation. To minimize error in a stochastic gradient-descent algorithm, each parameter is deducted and its direction of descent is determined.

3.5.2 Deep Belief Network

Geoffrey Hinton proposed the DBN in 2006 [30, 31]. It's a blueprint for future generations. We can increase the chance of the neural network as a whole generating training data by adjusting the weights between its neurons. It is possible to stack many Restricted Boltzmann machines to build a Deep Belief Network by using their layers as the hidden layer of prior layers, and the training of Restricted Boltzmann machines is straightforward because they can utilize their output to teach the next Restricted Boltzmann machine. Figure 2 depicts the general layout. Multiple layers of nonlinear variable connections constitute this generative model. RBM is the deepest layer of the belief network, while many Bayesian belief networks are found closer to the visible layer. In addition, this training strategy allows DBN to obtain a deep feature representation from unlabeled data, detect features, categorize data, and produce new data.

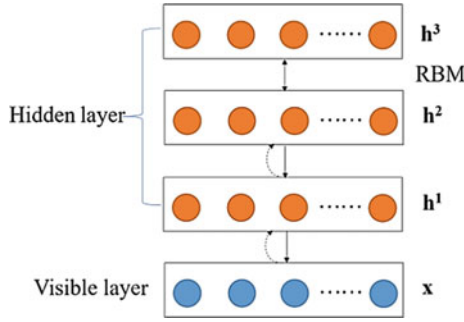


Fig. 2 Deep belief network

Each layer of DBN’s neurons is further separated into visible and buried ones. Hidden elements are utilized to extract features from the visible elements. As a result, feature detectors are names given to the hidden elements. An undirected connection between layers one and two of the DBN creates a form of associative memory. For example, each neuron in RBM’s bottom layer represents a single dimension in the data vector, and all levels of DBN are connected layer-by-layer to form a data vector. Train DBN, it must be done step-by-step. Using the data vector of the previous layer, the hidden layer is inferred, and this hidden layer is utilized as the data vector for the following layer.

Weighted fusion is used to assess the node’s likelihood of location and a location map is constructed for each node based on the findings of these two models. Using the location map, you can see which grids the node is most likely to migrate into in the upcoming period. As a starting point, each model used to estimate the location is given two weights $\{w_1, w_2\}$ and w_2 with identical values of 0.25. At the end of that time interval, the most correct model gets a boost in weight, while less accurate models get a fall in weight. This process is repeated for each place predicted by each model.

$$W_{xt} = \begin{cases} W_{xt-1} + \beta W_{xt-1}, & \text{if } P_t = \text{Predictat}'t' \\ W_{xt-1} - \beta W_{xt-1}, & \text{if } P_t \neq \text{Predictat}'t' \end{cases} \quad (19)$$

The probability of a node’s anticipated location at time ‘t’ can be calculated using the location map.

Node forwarding decisions are based on a location map and the following three rules.

- Forwarded to node P in the event of a 100% chance of node location at P.
- Node location at P is forwarded to P if the likelihood of node position at P is less than 1 but additional projected locations are all within a reasonable distance of P.
- The projected locations are clustered according to neighbourhood and packets are routed to each cluster area if the chance of a node being located at P is less than 1, but the subsequent locations are not in the same neighbourhoods.

The position and speed of the node are conveyed in a new control message. While moving from one grid area to the next, this message UPDATE is transmitted together with the current node and its location.

4 Results and Discussion

It was determined that the proposed technique, in return for a stable and short path, was effective using the network simulator-II tool. Table 2 shows the simulation settings.

In the proposed solution, a comparison is made between the predicted position and the actual location. The largest and lowest error ranges are shown in the table below. Because each grid is so small, the disparity in predictions between neighboring grids is smaller. Table 3 summarises the findings.

Node speed of 20 m results in a low error of 5 m, a medium error of 10 m, and a high error of 35 m. The low, average, and high error values are all reduced when the node’s speed is reduced. As an example, if the speed of the node is 10 m, the proposed model consumes 20 m for the maximum value, 8 m for the average value, and just 5 m for the lowest value. Mobility models were examined with a variety of MANET performance metrics in these studies.

1. Packet Delivery Ratio (PDR)
2. Delay
3. Network overhead
4. Energy Consumption

Table 2 Description of the network

Number of nodes	100–500
Simulation area	2500*400 m ²
Tool	NS-II
Energy	100 J
Simulation length	100 s

Table 3 Location error for various speed

Speed of node (m/sec)	Highest value of error (m)	Lowest value of error (m)	Average value of error (m)
5	15	5	7
10	20	5	8
15	30	5	10
20	35	5	10

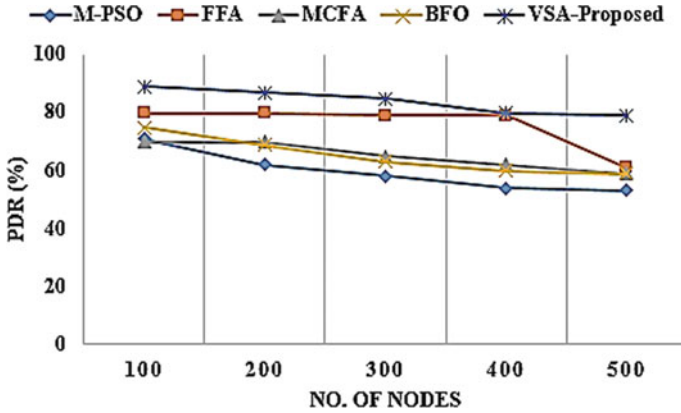


Fig. 3 Graphical illustration of projected ideal in terms of PDR

It is shown in Fig. 3 that the suggested model outperforms existing strategies such as modified Particle Swarm Optimization (M-PSO) [25], mapping cuttlefish swarm optimization (MCFA), and fruit fly optimization (FFA), and bacteria foraging optimization (BFO).

When the nodes are fewer, the PDR of all techniques is high, but it slightly degrades when the nodes are high. For instance, the M-PSO achieved 71%, FFA achieved 80%, MCFA achieved 70%, BFO achieved 75%, and proposed VSA achieved 89% for the nodes 100. The same techniques achieved 58%, 79%, 65%, 63%, and 85% of PDR, when the nodes are 300. While comparing with all techniques, M-PSO achieved low performance, i.e., 62% for 200 nodes, 54% for 400 nodes, and 53% for 500 nodes. But the proposed model achieved 87% for 200 nodes, 80% for 400 nodes, and 79% for 500 nodes. The reason for the poor performance of M-PSO is it easily falls into local optimum and uses GA for multipath routing, but the proposed model uses a weighted deep learning model for multipath routing and achieved high performance. Figure 4 shows the comparative analysis of the Transmission delay of packets.

The proposed model has 0.35 ms for delayed transmission of packets when it has 500 nodes, but the M-PSO has 0.91 ms, MCFA and BFO have 0.55 ms, and FFA has 0.7 ms. When the nodes are fewer, the delay time for the proposed model is only 0.2 ms, BFO has 0.3 ms, MCFA has 0.4 ms, FFA has 0.65 ms and M-PSO has 0.85 ms. While compared with all techniques, MCFA has average performance than other existing techniques, i.e. 0.38 ms for 200 nodes, 0.48 ms for 300 nodes, and 0.49 ms for 400 nodes. But, the proposed model achieved better performance than MCFA, i.e., 0.21 ms for 200 nodes, 0.25 ms for 300 nodes, and 0.3 ms for 400 nodes. Figure 5 shows the graphical representation of the proposed model in terms of network overhead.

The low network overhead means the best model for transmission of a large amount of data over MANET, which saves more energy. In the analysis for 200 nodes, the M-PSO has 0.8, FFA has 0.71, MCFA has 0.63, BFO has 0.55 and the proposed

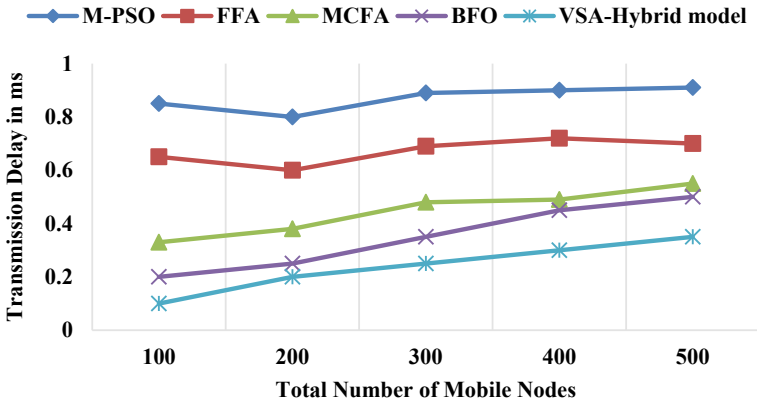


Fig. 4 Delay in transmission

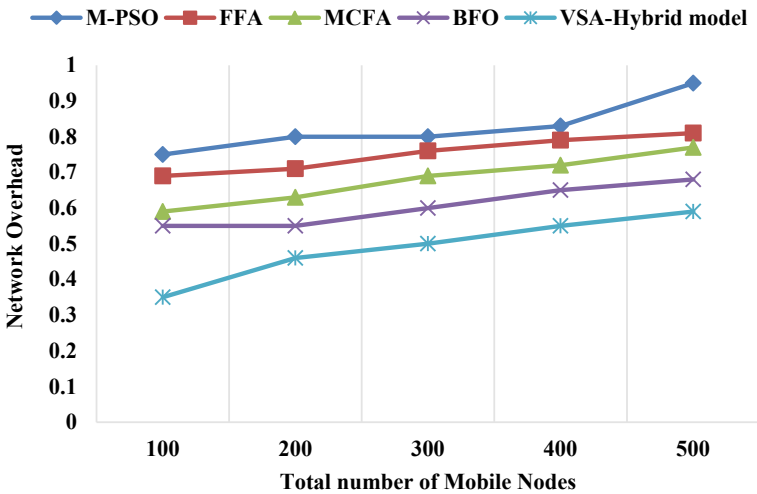


Fig. 5 Comparative analysis over various techniques in terms of network overhead

model has 0.46 network overhead. When the nodes are increased, the network overhead also increased, even for the proposed model. For instance, the proposed model has 0.5 for 300 nodes, 0.55 for 400 nodes, and 0.59 of network overhead for 500 nodes. But the existing models have nearly 0.6 to 0.8 of network overhead for 300 nodes and nearly 0.81 to 0.95 of network overhead for 500 nodes. Figure 6 shows the comparative analysis of all techniques in terms of energy consumption.

When the node is 100, the M-PSO consumed 50 J, FFA consumed 40 J, MCFA consumed 35 J, BFO consumed 25 J, and the proposed model consumed 15 J. However, the energy consumption is high for the proposed model, when a large amount of data is transmitted over the MANET network. For instance, the same

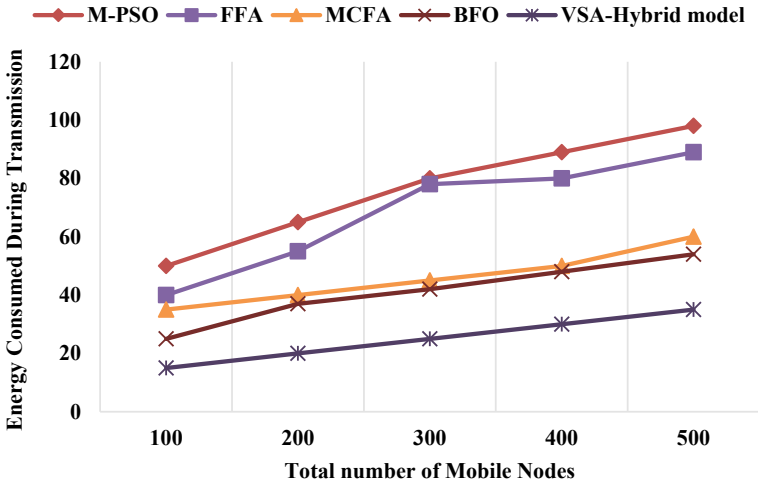


Fig. 6 Energy consumed during transmission

techniques consumed 98, 89, 60, 54, and 35 J, when the node is 500. The reason for less consumption of energy in the proposed model is that the various routes are considered for data transmission by using the weighted deep-learning technique, but the existing model uses either an optimization model or AODV protocol for multipath routing to transmit the data. From this analysis, it is proved that the proposed model achieved better performance not only in less energy consumption but also in network delay, overhead, and PDR.

5 Conclusion

DMPRPs based on VSAs were proposed in this research to pick the best cluster heads. An efficient energy-consumption ratio (ECR) can be achieved by grouping network field topology’s nodes into various clusters, which are then used to select the cluster head, hence improving load balancing and performance. We also need to find the most efficient route from NCH to CH depending on the CH we’ve chosen. Multipath routing is then performed based on the most likely location of each node. Node positions can be predicted using a multi-parameter, Spatio-temporal modeling technique. By analyzing the spatiotemporal findings, multipath routing can be improved in terms of both efficacy and dependability. It was observed that through simulations, the delivery ratio was advanced and the delay was lower than using existing methods. For scheduling non-real-time traffic, multiple paths can be selected based on how much delay each path can tolerate in the future.

References

1. Saeed NH, Abbod MF, Al-Raweshidy HS (2012) MANET routing protocols taxonomy. In: 2012 international conference on future communication networks. IEEE, pp 123–128
2. Sadhana S, Sivaraman E, Daniel D (2021) Enhanced energy-efficient routing for wireless sensor network using extended power-efficient gathering in sensor information systems (E-PEGASIS) protocol. *Smart Syst Innov Comput*, 159–171. https://doi.org/10.1007/978-981-16-2877-1_16
3. Bai Y, Mai Y, Wang N (2017) Performance comparison and evaluation of the proactive and reactive routing protocols for MANETs. In: 2017 wireless telecommunications symposium (WTS). IEEE, pp 1–5
4. Nayak PM, Sinha P (2015) Analysis of random waypoint and random walk mobility model for truthful routing protocols for MANET using NetSim simulator. In: 2015 3rd international conference on artificial intelligence, modelling and simulation (AIMS). IEEE, pp 427–432
5. Sivaraman E (2010) Dynamic cluster broadcasting for Mobile Ad Hoc Networks. In: 2010 international conference on communication and computational intelligence (INCOCCI), pp 123–127
6. Gao J, Zhao L, Shen X (2018) Network utility maximization based on an incentive mechanism for truthful reporting of local information. *IEEE Trans Veh Technol* 67(8):7523–7537
7. Sedrati M, Benyahia A (2018) Multipath routing to improve quality of service for video streaming over mobile ad hoc networks. *Wirel Pers Commun* 99(2):999–1013
8. Magnani DB, Carvalho IA, Noronha TF (2016) Robust optimization for OSPF routing. *IFAC-PapersOnLine* 49(12):461–466
9. Guan ZH, Chen L, Qian TH (2011) Routing in scale-free networks based on expanding betweenness centrality. *Phys A Stat Mech Appl* 390(6):1131–1138
10. Thaler D, Hopps C (2000) Multipath Issues in Unicast and Multicast Next-Hop Selection, document IETF RFC 2991
11. Holmberg K, Yuan D (2004) Optimization of Internet protocol network design and routing. *Networks* 43(1):39–53
12. Zhang J, Xi K, Chao HJ (2015) Load balancing in IP networks using generalized destination-based multipath routing. *IEEE/ACM Trans Netw* 23(6):1959–1969
13. Roughan M, Thorup M, Zhang Y (2003) Traffic engineering with estimated traffic matrices. In: Proceedings of 3rd ACM SIGCOMM conference internet measurement, New York, NY, USA, pp 248–258
14. Schmidhuber J (2015) Deep learning in neural networks: an overview. *Neural Netw* 61:85–117
15. Guan Z-H, Hu B, Shen X (2019) Introduction to hybrid intelligent networks: modeling, communication, and control. Springer, Berlin
16. Hu B, Guan Z-H, Chen G, Lewis FL (2019) Multistability of delayed hybrid impulsive neural networks with application to associative memories. *IEEE Trans Neural Netw Learn Syst* 30(5):1537–1551
17. Hu B, Guan Z-H, Qian T-H, Chen G (2018) Dynamic analysis of hybrid impulsive delayed neural networks with uncertainties. *IEEE Trans Neural Netw Learn Syst* 29(9):4370–4384
18. Majd, NE, Ho N, Nguyen T, Stolmeier J (2019) Evaluation of parameters affecting the performance of routing protocols in mobile ad hoc networks (MANETs) with a focus on energy efficiency. In: Future of information and communication conference. Springer, Cham, pp 1210–1219
19. Malar ACJ, Kowsigan M, Krishnamoorthy N (2020) Multi constraints applied energy-efficient routing technique based on ant colony optimization used for disaster resilient location detection in the mobile ad-hoc network. *J Ambient Intell Human Comput* 12:4007–4017
20. Darwish SM, Elmasry A, Ibrahim SH (2017) Optimal shortest path in mobile ad-hoc network based on fruit fly optimization algorithm. In: International conference on advanced machine learning technologies and applications. Springer, Cham, pp 91–101
21. Li F, Song X, Chen H, Li X, Wang Y (2018) Hierarchical routing for vehicular ad hoc networks via reinforcement learning. *IEEE Trans Veh Technol* 68(2):1852–1865

22. Ghasemnezhad S, Ghaffari A (2018) Fuzzy logic based reliable and real-time routing protocol for mobile ad hoc networks. *Wirel Pers Commun* 98(1):593–611
23. Thangaramya K, Kulothungan K, Logambigai R, Selvi M, Ganapathy S, Kannan A (2019) Energy-aware cluster and neuro-fuzzy based routing algorithm for wireless sensor networks in IoT. *Comput Netw* 151:211–223
24. Nallusamy C, Sabari A (2019) Particle swarm based resource optimized geographic routing for improved network lifetime in MANET. *Mobile Netw Appl* 24(2):375–385
25. Prakash V, Pandey S (2021) Best cluster head selection and route optimization for cluster-based sensor network using (M-pso) and Ga algorithms
26. Farheen NS, Jain A (2020) Improved routing in MANET with optimized multipath routing fine-tuned with hybrid modeling. *J King Saud Univ-Comput Inf Sci* 3:2443–2450
27. Anand C (2020) Scheduled optimal SDWSN using wireless transfer of power. *IROJ Sustainable Wirel Syst* 2(1):23–32
28. Srungaram K, Krishna Prasad MHM (2012) Enhanced cluster based routing protocol for MANETS. In: Meghanathan N, Chaki N, Nagamalai D (eds) *Advances in computer science and information technology. Networks and communications. CCSIT 2012. Lecture notes of the institute for computer sciences, social informatics and telecommunications engineering*, vol 84. Springer, Heidelberg. https://doi.org/10.1007/978-3-642-27299-8_36
29. Doğan B, Ölmez T (2015) A new metaheuristic for numerical function optimization: vortex search algorithm. *Inf Sci* 293:125–145
30. Hinton GE, Osindero S, Teh YW (2006) A fast learning algorithm for deep belief nets. *Neural Comput* 18(7):1527–1554
31. Bengio Y (2009) Learning deep architectures for AI. *Found Trends® Mach Learn* 2(1):1–127

Secure Wireless Smart Car Door Unlocking System



Navneet Vinod Melarkode, Varun Niraj Agarwal, Avaneesh Kanshi,
and Anisha M. Lal

Abstract Security is the most essential factor in software as well as internet environment. In the current era, software and hardware security is the most important thing to secure our digital and physical assets. In many cases, tangible entities like smart homes, smart car etc. having some IoT base support enhance the supervision. This paper proposes a secure alternative for the car unlocking system by dynamically generating disposable encryption keys that cater to the future demands of smart cars powered by Internet of Things while employing ultra-modern communications techniques like MAC-then-encrypt. Encryption keys are used to facilitate locking and unlocking functionality of car via the internet to ensure security. The paper compares the proposed methodology with existing algorithms like Digital Signature Authentication to verify its effectiveness in real world scenarios.

Keywords Connected cars · Internet of vehicles · Pseudorandom number · Hashing · Rolling code

1 Introduction

The recent growth of smart cars in Internet of Vehicles (IoV) is expected to increase from 40% in 2020 to 70% in 2025. Rapid development of infrastructure, systems, artificial intelligence learning models would cause a major shift of the consumers from the conventional type of vehicles to a more modern type. This is may also bring increased security issues, one of the major concerning issues being car door unlocking system. IoV vehicles are equipped with certain secret key that is transmitted over the internet to the owner's car that is also connected to the internet. There are two types of security aspects when it comes to a car key, i.e., software and hardware security.

N. V. Melarkode (✉) · V. N. Agarwal · A. Kanshi · A. M. Lal
School of Computer Science and Engineering, Vellore Institute of Technology, Vellore 632014,
Tamil Nadu, India
e-mail: naveetmelarkode@gmail.com

A. M. Lal
e-mail: anishamlal@vit.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
P. P. Joby et al. (eds.), *IoT Based Control Networks and Intelligent Systems*,
Lecture Notes in Networks and Systems 528,
https://doi.org/10.1007/978-981-19-5845-8_8

Software security is the area which deals with the securing the data that is being transmitted over the IoT network. Hardware security deals with the inbuilt security system of hardware to prevent any intruder from entering and making the hardware vulnerable. The paper focuses on the software security of the car unlocking system. When the secret key is matched with the one stored inside the car, the car is unlocked or locked as per the needs. This method may seem to be very time inexpensive and easy for the manufacturers to implement into the system. But it has a drawback, if an attacker is able to get inside the network, they can sense any piece of information that is being sent over the internet. If their intentions are for committing a car theft, they can easily read the key and replicate it later on for unlocking the car. The current methodology of car door unlocking system compromises its security aspect. Several approaches like symmetric and asymmetric encryption and decryption have been implemented but later on found out that it can easily be cracked with hardware-induced attacks. This paper proposes a novel approach to avoid the aforementioned problem by implementing an end-to-end system that generates disposable keys to verify the authenticity of the client. To send the keys over the internet via an unsecure public channel compromises the security. The proposed model ensures that any message sent through unverified channels are encrypted and hashed to avoid breach of trust and privacy.

1.1 Organization of the Paper

- Section 2 explains about the existing methodologies that have been proposed by other authors.
- Section 3 explains about the proposed methodology and how a unique framework was created by implementing ideas from existing security algorithms.
- Section 4 explains about the metrics that are considered while comparing the proposed model with the industry standards. It also talks about the comparison of different sub processes that are used in the algorithm.

2 Related Work

Pranab Ray et al. [1] incorporate the rolling key algorithm to overcome the flaws in its predecessor methodologies. Previous studies employed RF security which would have to pass through unsecure channels. The signals that pass through these channels are essential for locking and unlocking the system which brings about a blaring disadvantage in the security standpoint. To prevent and overcome the relay threats caused by unsecure channels, the rolling key algorithm is implemented which enables a 2-way handshake. The rolling key algorithm provides an unused key at every instance of data transfer between the sender and receiver.

R. Valanarasu [2] present a secure architecture for hospital environments with the help of an Internet of Things backend. They recognized limitations in the current framework such as inflexible modes of networking, data security etc. and strive to alleviate these threats with the help of the proposed method. The method presents a major upgrade from the existing methods by using a regulation and policy layer to overlook all the trust components such as safety, privacy and dependability.

Colin Urquhart et al. [3] analyze the underlying cyber security of a renowned car brand. It has been noted that as the technology of automobiles advance, many instances of cars being connected via 3G and 4G mobile networks are reported. While these services increase the ease of use and aid the consumer, they come with vulnerable system defects that lead to increase in the attacking area of the vehicle. The paper also carries out experiments to assist the study in the field of in-car attacks as well as standard benchmark attacks such as key fob rolling code and various other privacy attacks.

L. Jamjoom et al. [4] focus on developing a wireless car lock controller built on a mobile device. The research implements the said system incorporating Internet of Things concepts. The proposed methodology revolves around granting requests through a server-based utility over the internet. To accommodate the authorization, a code is sent to the mobile device. The mobile device must have the companion application pre-installed to complete the communication of the code. The code is transmitted via Bluetooth. It is then sent to a front-end controller which employs recognition techniques and relays a flag back to the mobile device. Future research in this field can aim to improve the monitoring of other parameters that include but are not limited to fuel consumption, Carbon Dioxide emission rate etc.

Sophia Auer et al. [5] use the growing popularity of in-car sharing and the rising number of applications of Blockchain technology as motivation for devising a new methodology for shared mobility that involves key aspects from both its constituents, Blockchain and Internet of Things technologies. The paper presents an architecture for encapsulating these technologies to assist car-leasing and car-sharing. The proposed method goes a step ahead by eliminating the requirement for keys to gain vehicle access. The authors identify that while sustainable, blockchain technology alone cannot expand this field in the future. Future researches can aim to evaluate the authenticity of the Internet of Things devices and the scalability of the proposed model. At present, the model fares well for simulated data.

Jeffrey Cashion et al. [6] presents a unique way of authentication of the client to the server using rolling code technology. It prevents the intruders from making copies of the code and using it to perform sidejacking. With the help of this technique, the client easily proves the server of its legitimacy. This in return prevents hijacking of the system to its best. The efficiency of the proposed code puts out optional payload integrity and confidentiality via a multi-level security model.

Kyle Greene et al. [7] performs set of infiltration experiments which revealed the vulnerabilities in the radio frequency communication of cars and garages in remote keyless systems. A timestamp-based solution was presented to enhance existing rolling code algorithm. The output of the rolling code algorithm is appended with certain command and the timestamp which is further encrypted using AES algorithm.

The algorithm proved to be highly secure with low complexity and able to be power efficient.

L. Jamjoom et al. [4] develops a smartphone based wireless controlled car lock. The idea is to allow a group of authorised people to share a lot of cars. Whenever an authorised person requires a car, a request is firstly submitted via internet to a server-based utility. On the availability of a car the request is granted and vice versa.

Sophia Auer [5] presents a high-level architecture for a blockchain-IoT-based platform for promoting shared mobility combining car-sharing and car-leasing. The proposed platform requires secure information sharing among multiple stakeholders (such as user, lessee, and service provider), leading to the decision to choose blockchain for its facilitation. IoT data generated by vehicles is of significant relevance to all involved stakeholders helping to streamline processes and features. This proved to provide security, privacy, authenticity, reliability and scalability without making the system more complex.

Hamdan Y. B. and Sathesh [15] highlight about 2 kinds of threats, namely information privacy attack and context aware privacy attack. Privacy is a major concern while communicating over the internet and hence needs proper security measures to be taken to ensure privacy of client-server channel.

It can be observed that the recurring pitfalls in previous keyless entry methods are data leakage, replay attacks and session hijacking. Data leakage occurs due to plain text being sent over unsecure communication channels which can be easily intercepted. Replay attacks are extremely common in keyless entry systems. They work on the assumption that the same signal can be used repeatedly to achieve the same result. Session hijacking is similar to a man-in-the-middle attack which interrupts the communication channel and tries to exploit the vulnerabilities of the system. Previous methods highlight the need for scalable systems as the Industry of Vehicles continue to grow at an exponential rate.

2.1 Pseudorandom Numbers

[8] A pseudorandom number sequence is a number sequence which appears to be statistically random, but are completely generated with the help of a deterministic function. These sequences are not truly random as they are heavily determined by the initial value set by the user. This initial value is also called as the seed. These generators are key in practice owing to their speed in generating these sequences [9]. While pseudorandom number generators are calculated using a deterministic function, it is necessary for the sequence to show approximate characteristics of a true random distribution. To do so, Python's NumPy library has been employed. By pre-generating the seed, the pseudorandom sequence can be generated in the desired way.

2.2 *Rolling Code*

[10] Rolling codes essentially transmit a counter that is incremented by each button press in a cryptographically authenticated way. [11] Rolling codes have been incorporated into many keyless entry doors unlocking systems due to its extreme versatility and lightweight nature. From [12], it can be seen that introducing a pseudo-random number generator into the rolling code can help extend the protection to prevent template attacks as well as alleviate the risks posed by brute force attacks. This is possible because the two random number generators on either side of the pipeline generate the same sequence of numbers which otherwise seem random for any intruder trying to intercept this stream.

2.3 *Hashing*

Hashing functions were proposed as an alternative to fulfil aspects of information security due to various reasons. [13] Firstly, it is computationally easy to calculate the hash of any given message when compared to asymmetric and symmetric key algorithms. Furthermore, two different messages are highly unlikely to be associated with the same hash because of the sheer number of possible hash values. Hash functions being employed in the proposed model are collision resistant to a high degree. Subsequently, messages cannot be altered without changing the hash. Ultimately, generating a message from a given hash is very unfeasible thus nearly eliminating brute force attacks [14].

The proposed method employs SHA hashing function over the MD5 hashing function due to very less and infrequent potential collisions. MD5 generates a 128-bit output while SHA256 generates a 256-bit output. While the SHA256 hashing may be slower than MD5, the speed is not slow enough for us to overlook the security advantages it provides.

3 *Methodology*

3.1 *Objective*

The paper presents a novel approach by combining the power of the tried and tested state-of-the-art algorithms which are designed to work in a simplex communication environment and ultra-modern hashing technologies.

This enables us to encrypt any data in an irreversible manner, making the communication over the internet extremely secure. Our proposed method employs a rolling filter as a key generation method for a modified MAC-then-encrypt authenticated encryption algorithm. Instead of using a bi-directional encryption algorithm in the

final stage of MAC-then-encrypt, the proposed model uses another (different from key hash) hashing algorithm which makes the encrypted message doubly hashed and impossible to crack. Double hashing and a unidirectional approach can be integrated due to the fact that the message is predefined in both of the systems and only the source of the messages needs to be authorised. A unidirectional double hashing methodology can be implemented by chaining the outputs of subprocesses to form the resultant message string.

The proposed methodology brings elements of security and robustness together. By sending uncrackable messages over the internet, the proposed system ensures authentication in both the parties. This has not been implemented in predecessor systems which adds to the novelty of the entire concept.

3.2 Modules

The proposed method can be divided into 2 modules. The first one being client/key side and the second one, the car side. The client will be interchangeably used as key since the client is the key in the proposed framework. The client will try to lock or unlock the car from their remote device connected through the internet. Connected cars are always connected to the internet. The car will be listening for a hash digest over the internet. As soon as it receives the digest, the car will then validate and verify the authenticity to proceed to take a decision on the car controls.

Certain assumptions made in the following 2-way handshake system include the initialization of a seed before the client and car are actually made public. This seed will be completely random, and it is necessary for the same seed to be set in both the car as well as the client. This is essential because the entire architecture of Pseudorandom number generated rolling codes is dependent on this. Furthermore, in case the client side rolling code queue does go out of sync with the car rolling code queue, a manual reset must be performed with the initialization of a new random seed.

3.2.1 Key Side

The key will generate a rolling code queue with the seed that is initialized with. The rolling code queue uses the pseudorandom number generator with the deterministic function to generate a stream of numbers. These numbers are then enqueued sequentially. Upon pressing of the car control button such as lock or unlock; the client then polls the queue and hashes the polled pseudorandom number. Depending on the car control button pressed, the corresponding code will then be appended to the newly generated hash digest. After appending the code, the entire message is now hashed again in order to protect the status of the car from being compromised in the case of a man in the middle attack. The architecture followed here is an indirect implementation of the MAC-then-Encrypt scheme where a MAC is produced based on

the plaintext, and the plaintext and MAC are encrypted again to produce a ciphertext based on both. While the ciphertext is sent, the result hash digest is sent to the car in the proposed system.

The key side has to perform hashing twice for every button click, and given the computational prowess of modern-day machines, the given model feels viable. In the off chance where the key isn't connected to the car side, the counter of the rolling code will keep getting incremented until a point where the car queue won't be able to look ahead and get back in sync. In the case where a hacker impersonates a client and sends repeated signals to the car, a Denial-of-Service prevention mechanism is in place depending on the number of false signals being sent in a given amount of time.

From Fig. 1, It can be seen that the doubly hashed message sent over the internet is secure. It is to be noted that the client side does not perform any validation or verification of the person pressing the button. This is analogous to a traditional key, where nobody can actually verify if it is actually the key owner opening the lock. Apart from the 2 hashes and the appending of car control signal, the client side does not perform any other computations. The same flowchart is given in the form of words in Algorithm 1.

1. Initialize queue rq
2. Populate rq
3. if button_press is true:
 - a) $rq.top \rightarrow temp$
 - b) $rq.pop$
 - c) $hash(temp) \rightarrow temp$
 - d) $temp + \text{"LOCK"/"UNLOCK"} \rightarrow temp$
 - e) $hash(temp) \rightarrow temp$
 - f) Transmit

3.2.2 Car Side

The car side algorithm is fairly trivial as compared to the key side. Just like the key, the car too will be fitted with the same randomly generated seed. The car generates the same set of pseudorandom numbers and will listen for any message that is sent over the internet.

Upon receiving the message, the car checks whether the queue is empty. Queue being empty is an edge case for when the car and key go out of sync. If the queue is empty, the car and key would then have to be manually reset with a brand-new seed. Until then, the internet functionalities would be restricted in the car to prevent future

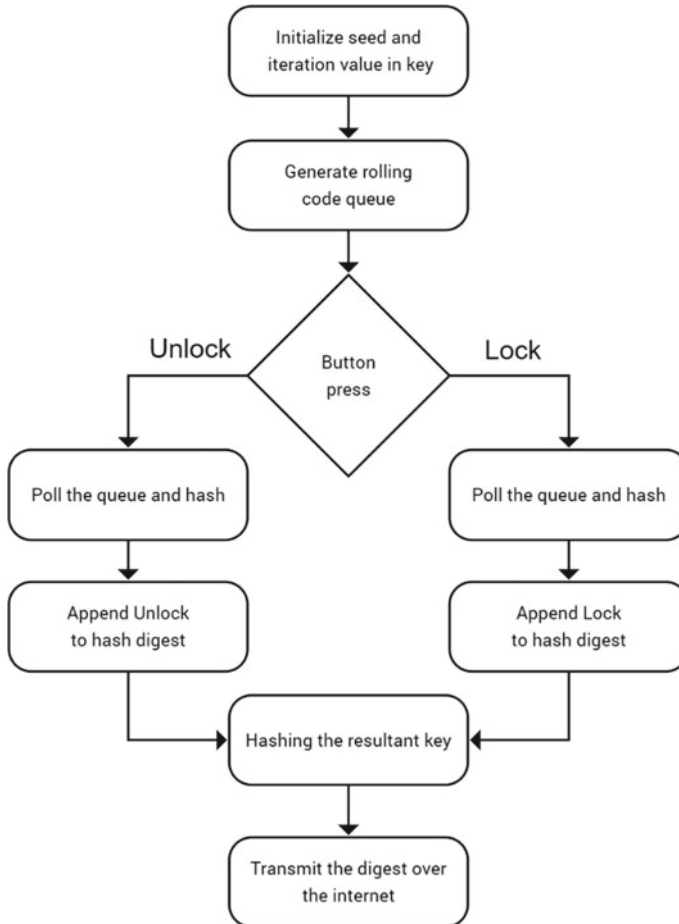


Fig. 1. Flowchart for client side

attacks as the car is now vulnerable. If the queue is not empty, the car then polls the queue and subjects it to the exact same hashing system to generate a hash digest.

As mentioned earlier, hashes cannot be decrypted, but it can only be compared with other hash digests to check their validity. Consequently, each number from the queue would have to be hashed and then appended with each control signal/code of the car (Lock or unlock) and then hashed again.

The resulting hash digest will be compared with the message sent over the internet. If they are a match, the car will execute the corresponding control, and if they are not, the car moves ahead until the entire queue is tested. The ideal rolling code queue size is 256, which would mean it would take exactly 256 out of sync clicks from the key side to restrict the car's internet activity and make it available only for physical locking/unlocking.

From Fig. 2 below, it can be seen that the car actively listens for a message and upon receiving of the message, it validates the hash digest by subjecting the rolling code queue to the same hash functions in the same order. For each number, the car will have to perform 1 common hash with the addition of 2 hashes after appending the respective car control signal (Lock or unlock) to the intermediate hash digest. The same can be seen with the following algorithm.

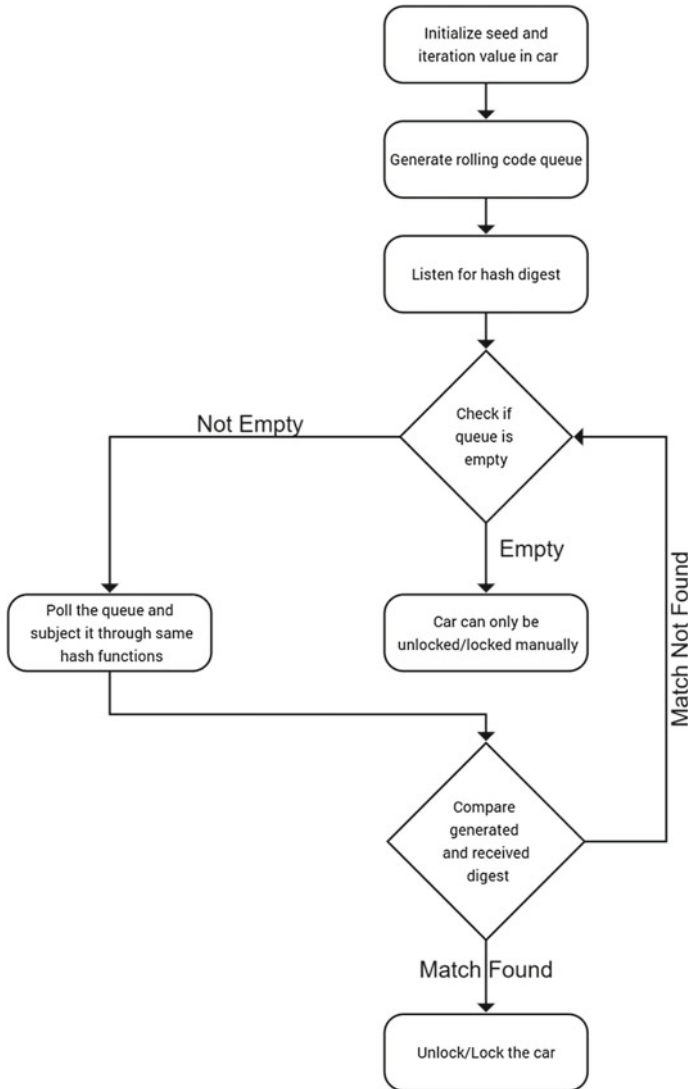


Fig. 2. Flow chart for Car side

1. Initialize queue rq
2. Populate rq
3. Initialize TCP socket
4. Receive 256 bits of data $\rightarrow d$
5. for each num in rq :
 - a) $hash(num) \rightarrow temp$
 - b) if $temp == d$:
 1. Lock or Unlock the car
6. if no match found:
 - a) Restrict car's activities over internet

4 Performance Evaluation

4.1 Type

In the proposed model, the encryption function is a hash digest and the decryption function is a hash comparison. Hence making the program extremely secure towards brute force attacks.

4.2 Function Analysis

SHA256 is used for hashing, rolling code for disposable key generation and finally a MAC-then-encrypt architecture for the network authentication.

This sequential model involves using multiple tried and tested models which theoretically yields a non-brute forceable encryption standard at the cost of computational power while maintaining the integrity of the data.

4.3 Key Size

The proposed model uses a random number generator which yields a key of size 10^8 . This can be increased depending on the computational power available. However, since the final output is hashed, the output is always a fixed size of 256 bits.

4.4 Rounds

The proposed model hashes in 2 rounds and ensures maximum encryption.

4.5 Time complexity

The time complexity for each encryption/digest is,

$$\sum = T(n) = 2 * ({}^{29}C_1 + {}^4 C_2) N_0 + 2 * ({}^{10140}C_1 + {}^2 C_3 + {}^{64} C_4 + {}^2 C_5) N_1 = \theta(N) \quad (1)$$

From (1), it can be clearly observed that the encryption function is linear in time and hence can provide great benefits while not compromising on quality of the hash digest. This encryption function is running a total of k times. Hence, the total time complexity for the embedded system in the car is $\theta(l * N)$. However, once the queue is generated, the time complexity reduces to $\theta(N)$ and the time complexity to check becomes $\theta(l * k * 256) = O(l * k)$. Where k represents the queue length and l represents the number of functions. The complexity of the sender-side and client-side encryption is $\theta(n)$.

4.6 Space Complexity

For the car side,

$$\sum = S(n) = l * k * (256) = l * k \quad (2)$$

In (2), k are the cycles and l are the number of functions. Increasing the number of cycles will increase the space complexity, however, the space complexity is linear. This enhances the scalability of the proposed method. On the car side the space complexity is mere $\theta(1)$.

4.7 Common attacks

- 1) Brute force - Brute force attacks are practically impossible since the hash digest is hashed again with a disposable key.
- 2) Replay attacks - Replay attacks are not possible because the rolling code produces a new key which cannot be reused every cycle.
- 3) Rolling code overflow - Rolling code overflow is a common issue in systems employing rolling code. The proposed model aims to minimize it with the help

of a failsafe which blocks failed IP addresses however this is not a full proof method to solve this issue.

- 4) DoS attacks - DoS attacks can easily be addressed by employing a simple software firewall/IPS but it can lead to not being able to open the door over the internet. Firewalls provide a filter-based as well as a rule-based packet forwarding mechanism which blocks anomalous requests upon multiple failed attempts.

4.8 Comparison with DSA Algorithm Standard

Digital Signature Algorithm (DSA) serves as the benchmark for the proposed method as it is the industry standard for authentication systems. DSA uses a private key to sign messages which can be verified in the client side using a public key. DSA provides users with integrity – clients can verify the contents of the message and non-repudiation – the server cannot claim that they have not signed the message. This is the reason DSA has been chosen as the evaluation algorithm.

- 1) Architecture – Both DSA and the proposed architecture employs a MAC-then-encrypt architecture for its foundation. However, the message is visible in digital signature algorithm and it is extremely prone to replay attacks.
- 2) Key – The key size in rolling code is variable and the keys are disposable and changes constantly.
- 3) Rounds – Both DSA and the proposed algorithm employs the same number of rounds.
- 4) Time complexity – DSA outperforms the proposed methodology in this comparison metric. DSA is $\theta(k \cdot l)$ times faster however since the values of k and l are relatively small, it can be considered that DSA is constant time faster than our proposed work.
- 5) Space complexity – Both the algorithms are similar in terms of space complexity.

4.9 Usage of SHA vs MD5 in Our Proposed Method

From Table 1, it is evident that SHA outperforms the MD5 hashing technique in the department of security. While SHA is more time consuming, it is not slow enough to overlook the security benefits it provides.

Table 1. Comparison of SHA and MD5

SHA	MD5
Highly Secure as the final output is 256/512 bits	Exponentially less secure as the final output is only 128 bits
Half as fast as MD5	Double the speed of SHA
No known attacks	Many reported attacks known
Fixed input size	Any input size works

Table 2. Comparison of DSA and the proposed method

Proposed method	DSA
Based on MAC-then-encrypt architecture	Based on MAC-then-encrypt architecture
Slower compared to DSA	Faster than proposed method
Highly secure	Less secure
Replay attacks do not work	Susceptible to common replay attacks
New disposable key every cycle	Fixed key

4.9.1 DSA vs Proposed Method

From Table 2, it can be observed that the proposed method performs better than DSA in the security aspect. On the contrary, it lacks speed in comparison to DSA.

5 Conclusion

The proposed algorithm which aims to secure the IoT based car lock system incorporates the basic notion of MAC-then-encrypt architecture layered with cryptographic mathematical function and pseudorandom number generator function. This results in increasing the randomness of the serial keys generated while double securing the signal being sent over the internet. This provides the smart car with improved layer of security for locking mechanism and a secure solution for the booming industry of IoVs. Even though the algorithm requires multiple computations to be performed, it doesn't compromise on its efficiency, making the proposed methodology highly efficient. The algorithm succeeds in strongly preventing any intruder launching a replay or session hijacking attack due to the encryption key's dynamic nature. The integrity was tested against several attacks and proved to remain uncrackable. Despite providing good security, the time and space complexity can be further optimized to elevate the performance. On a large scale, it may prove to be fatal due to large amount of storage space being required to store the dynamic changing keys hence further research is required for management of encryption keys.

References

1. Ray P, Sultana HP, Ghosh S (2019) Removing RF vulnerabilities from IoT devices. *Procedia Comput Sci* 165:421–427
2. Valanarasu MR (2019) Smart and secure IoT and AI integration framework for hospital environment. *J ISMAC* 1(03):172–179
3. Urquhart C, Bellekens X, Tachtatzis C, Atkinson R, Hindy H, Seeam A (2019) Cyber-security internals of a skoda octavia vRS: a hands on approach. *IEEE Access* 7:146057–146069
4. Jamjoom L, Alshmarani A, Qaisar SM, Akbar M (2018) A wireless controlled digital car lock for smart transportation. In: 2018 15th learning and technology conference (L&T). IEEE, pp 46–51
5. Auer S, Nagler S, Mazumdar S, Mukkamala RR (2022) Towards blockchain-IoT based shared mobility: car-sharing and leasing as a case study. *J Netw Comput Appl* 200:103316
6. Cashion J, Bassiouni M (2011) Robust and low-cost solution for preventing sidejacking attacks in wireless networks using a rolling code. In: *Proceedings of the 7th ACM symposium on QoS and security for wireless and mobile networks*, pp 21–26
7. Greene K, Rodgers D, Dykhuizen H, McNeil K, Niyaz Q, Al Shamaileh K Timestamp-based defense mechanism against replay attack in remote keyless entry systems. In: 2020 IEEE international conference on consumer electronics (ICCE). IEEE, pp 1–4
8. Lagarias JC (1993) Pseudorandom numbers. *Stat Sci* 8(1):31–39
9. James F (1990) A review of pseudorandom number generators. *Comput Phys Commun* 60(3):329–344
10. Oswald DF (2016) Wireless attacks on automotive remote keyless entry systems. In: *Proceedings of the 6th international workshop on trustworthy embedded devices*, pp 43–44
11. Garcia FD, Oswald D, Kasper T, Pavlidès P (2016) Lock it and still lose it—on the (in) security of automotive remote keyless entry systems. In: *25th USENIX security symposium (USENIX Security 16)*
12. Moradi A, Kasper T A new remote keyless entry system resistant to power analysis attacks. In: 2009 7th international conference on information, communications and signal processing (ICICS). IEEE, pp 1–6
13. Tang J, Tian Y (2016) A systematic review on minwise hashing algorithms. *Ann Data Sci* 3(4):445–468
14. Gupta S, Goyal N, Aggarwal K (2014) A review of comparative study of md5 and ssh security algorithm. *Int J Comput Appl* 104(14):1–4
15. Hamdan YB (2021) Smart home environment future challenges and issues-a survey. *J Electron.* 3(01):239–246

A Comprehending Deep Learning Approach for Disease Classification



Ankita Nainwal, Bhaskar Pant, and Garima Sharma

Abstract For medical image processing analysis, deep learning is one of the most popular research subjects. It is subset of machine learning comprising of one or more neural network layers to simulate human behavior of learning and predicting. The purpose of this work is to investigate the application of deep learning models in image processing for disease analysis and medical innovations. The work showcased a generic deep learning model based on convolutional neural networks to classify diseases upon image analysis. To demonstrate the extensive medical use case of proposed model, the results demonstrated classifying pneumonia x-ray images alongside normal chest x-ray images, wrist r-ray pictures able to distinguish between normal and fractured wrists, etc.

Keywords Convolutional neural network · Pneumonia detection · Covid-19 · X-ray images · Image processing

1 Introduction

Deep learning is one of the most prominent image processing algorithms in recent years. Deep learning is a machine learning technique that focuses on learning from data. It can learn from historical data and assist in the prediction and development of intelligent systems [1]. Deep learning has a wide range of applications in today's environment. Image classification, face recognition, human action recognition, audio processing, text analysis, natural language processing, autonomous systems, robotics, medical diagnostics, computational biology, physical sciences, finance, economics, market analysis, and others are just a few examples of application areas [2, 3].

Artificially Intelligent methods have been developed as a result of recent technological advancements, and they have a wide range of applications in numerous fields. The fundamental goal of evolution in today's world is to create a system that can

A. Nainwal (✉) · B. Pant · G. Sharma
Graphic Era Deemed to be University, Dehradun 248001, Uttarakhand, India
e-mail: Ankitanainwal.cse@geu.ac.in; ankitanainwal1424@gmail.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
P. P. Joby et al. (eds.), *IoT Based Control Networks and Intelligent Systems*,
Lecture Notes in Networks and Systems 528,
https://doi.org/10.1007/978-981-19-5845-8_9

113

think and act like a person, and deep learning falls into this category. Convolutional Neural Networks (CNN) and Deep Belief Networks are two techniques utilized in the deep learning process (DBN). CNN is a neural network that is mostly used to process two-dimensional input such as photos and movies. CNN [4] is used to separate the images into smaller sections and process them. DBN [5] is a probabilistic or mathematical model that learns through hierarchical abstraction.

Due to their widespread popularity, machine learning technologies are becoming increasingly popular in medical applications. The analysis and processing of medical pictures is one of the applications of machine learning and deep learning in such applications. Because it takes a lot of time, effort, and money to teach a person to interpret and process medical images for prediction and testing, deep learning methods prove to be more effective, efficient, and speedier. Computed Tomography (CT scans), X-ray pictures, Radiography, endoscopy, Positron Emission Tomography (PET), and other medical images for image processing can be categorised into many categories. Such medical photos are used as input, which is then processed and evaluated using deep learning. Multiple layers and processes are used to extract and process features from the photos [6]. Various studies in the field of medical image processing in the deep learning model employ supervised and unsupervised learning methods often. CNN and recurrent neural networks are two commonly used supervised learning approaches (RNN). Unsupervised learning methods, on the other hand, include auto encoders (AE) and Deep Confidence Networks (DBN) [7].

2 Related Work

A study proposed in [8] uses CNN for analyzing MRI Images for predicting Alzheimer's disease. Identifying Mild cognitive impairment (MCI) in patients helps in the early detection of Alzheimer's disease and therefore CNN is used to predict MCI in Alzheimer's disease however the method performed tested and used for MRI images. Another study [9] examines images of endoscopy which uses the Yolov3 algorithm for the detection of a polyp. The proposed method helped in automating the process of detection of polyp as gastric polyp may lead to stomach cancer however the method is only limited for the detection of gastric polyps. This method helps in making the medical examination much faster and more accurate. Brain tumor detection can be very complicated therefore [10] presents a method for detection of tumors in the brain by initially preprocessing methods followed by the convolutional neural network. The preprocessing methods included global thresholding, adaptive thresholding, sobel filter, high pass filter, median blur, histogram equalization, dilation, and erosion. Image processing and normalization of the model are done wherein the dataset is cleaned by resizing, removing noises, etc., further the model is trained using CNN. However the model introduced is only used for classifying the images with brain tumors and without brain tumors. Also the preprocessing of images done could corrupt the images thereby hindering results.

A deep learning method is used to detect heart disease [11]. The model is processed and normalised using MRI scans, and the dataset is cleaned by shrinking and removing noises, among other things. The model is then trained using CNN. The method helps in detecting global hyperkinesia which helps in the early detection and prevention of heart disease have been only used for the prediction of heart disease based on MRI tests in laboratory. Research for dental problem detection and classification is presented in [12]. The proposed study helps in the diagnosis of 14 different dental problems by using panoramic X-Ray images. The presented study uses CNN models for the detection of teeth and the classification of dental problems. Another use of deep learning in medical science is the detection of bone fractures. The typical examination of an X-Ray image by doctors is time-consuming, and the risk of error is high. A method [13] is proposed to identify fractures because the traditional examination of an X-Ray image by doctors is time-consuming and error-prone. The method uses CNN to classify healthy and fractured bones. A study [14] uses COVID-19 thoracic x-rays and the histogram-oriented gradients (HOG) feature extraction methodology to create an accurate classification method for performing a reliable detection of COVID-19 viral patterns. The presented model can classify images with an accuracy of 85 percent for binary class prediction, but the proposed method of CNN + HOG can classify with an accuracy of 93 percent.

3 Methodology

The proposed study uses a deep learning method for the classification of images. The method involves the classification of healthy lungs with an unhealthy ones thereby determining whether the patient has pneumonia or not. The method uses the CNN model to classify the lungs images. The model is then again trained for classification of wrist images to identify the fractured wrist with normal one. Figure 1 depicts the methodology followed for the presented study. The initial step includes acquisition of data followed by data preprocessing and after processing the data model building is done wherein an application is built and classification of images are done and in the final step evaluation of the presented model is done.

In the proposed study we train a deep neural network for image classification using transfer learning. We used CV studio as it is very easy, fast and open source image annotation tool.



Fig. 1 The methodology for the proposed framework

1. Data Acquisition and Preprocessing:

- **Datasets:** For the processing of the medical images, the datasets were taken from [15, 16]. For pneumonia detection, the datasets consist of data of normal lungs and that of a pneumonia patient. The images were first categorized into test and train sets. The system was then trained using training images from datasets and later a different set of images from test sets were taken for both the normal lungs and infected lungs for classification. Similarly for the training of wrist x-ray images, the datasets was obtained from google datasets research which contains the x-rays images of fractured wrist and normal wrist and the same process was used to train and test the model.
- **Data Preprocessing:** Image processing is carried out to increase the quality of photographs for better analysis. In this phase, the datasets are preprocessed by changing the image form, converting to tensor, and normalising the image channels. Normalization is an important stage in the preprocessing process. This is the process of rescaling pixel values to fit within a specific range. One of the reasons for doing so is to aid with the problem of gradient propagation. Many times, the amount of data we have is insufficient to adequately accomplish the classification task. In such circumstances, data augmentation is used. In image-based deep learning problems, augmentation is frequently employed to improve the volume and variation of training data.

2. Data Exploration and Visualization: The datasets are preprocessed in this phase by modifying the picture form, converting to tensor, and normalising the image channels. In addition, we also perform data augmentation on the training dataset. The images shown in Figs. 2 and 3 depicts some of the transformed data for both of the studies performed, ie; wrist fracture detection and pneumonia detection.

3. Model Building and Model Evaluation: A deep learning neural network is trained for the classification of images making use of transfer learning. We have

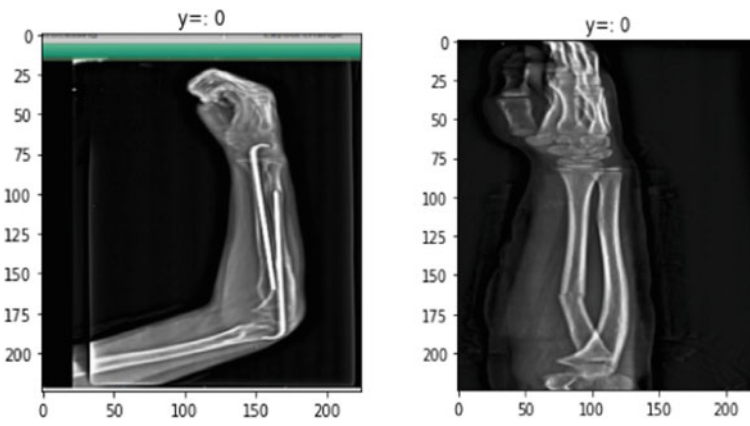


Fig. 2 Wrist X-ray datasets [15]

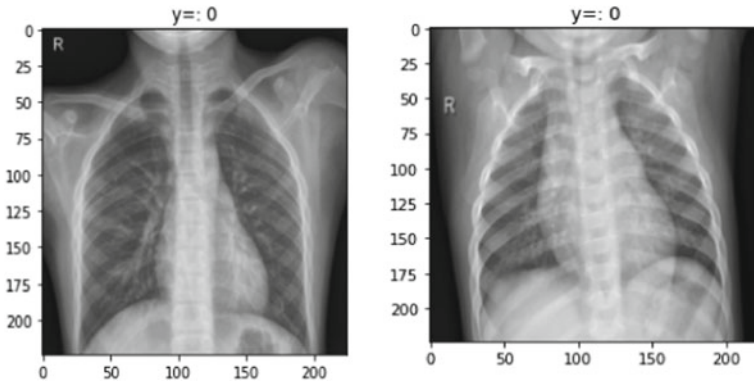


Fig. 3 Chest X-ray datasets [16]

used the Convolutional Network as a feature generator, only training the output layer. The main idea behind using CNN model is because CNN has a very high accuracy in image classification and recognition. Also as compared to other machine learning algorithms like KNN, SVM and Logistic Regression, CNN is more efficient and needs less time. Deep learning method uses transfer learning therefore it is more efficient and less error prone.

Convolutional Neural Network (CNN):

One of the most popular architectures in deep learning is Convolutional Neural Network as it can automatically retrieve features and the process of manual selection and extraction could be reduced for classification and categorization. The CNN is composed of the input layer and output layer and various hidden layers between the input and output layers [17]. The in-between layers could be further divided as the convolutional layers, pooling layers, and fully connected layers. The main idea behind when to use the deep learning model is that it works in fields where the human experts are unavailable or the human experts are not able to make decisions or come to conclusions. Deep learning methods can also be used in fields where humans require great expertise and cost to make decisions [18]. Further using CNN architecture in deep learning helped in extracting the required features easily without manual intervention [19].

Hyper parameters: The study was done using various hyper parameters some of them are:

- Epoch: Indicates the number of passes of the entire training dataset, we have taken the number of epochs to 10. When training a neural network using sample data, overfitting is a major risk. When the number of epochs used to train a neural network model exceeds the required amount, the training model learns patterns that are extremely specific to the sample data. This makes it impossible for the model to perform effectively on a new dataset. On the training set, this model is very accurate, but not on the test set. To prevent overfitting and improve the

Table 1 Model training and evaluation

Epoch	Learning rate	Validation cost	Validation accuracy
1	0.0028000000000000002	0.7308348655700684	1.0
2	0.0046	0.6124793767929078	1.0
3	0.0064000000000000001	0.26566229611635206	1.0
4	0.0081999999999999999	0.320123690366745	0.9333333333333333
5	0.0100000000000000002	0.3066509410738945	1.0
6	0.0081999999999999999	0.19219330921769143	0.8
7	0.0064000000000000001	0.16543330997228622	0.9333333333333333
8	0.0046	0.2259145550429821	0.9333333333333333
9	0.0028000000000000002	0.13125711306929588	0.189045762270689
10	0.001	0.9333333333333333	0.9333333333333333

neural network's generalization ability, the model should be trained for an optimal number of epochs.

- **Learning rate:** In the training of neural networks, the learning rate is used. The learning rate is a hyper parameter that has a modest positive value, usually between 0.0 and 1.0.

Table 1 depicts various hyper parameters used for training and evaluating the presented model.

The novelty of the presented approach is that the model is able to classify any x-ray images for the disease classification. The method is not only tested for pneumonia detection but also for the detection of any fractures in the wrist x-rays. The method is applicable for any image processing for disease classification after training the model with the desired inputs or training images.

4 Results and Discussion

The study uses a deep learning method for image classification. For classification of lungs images for pneumonia we used transfer learning with Convolutional Neural Network for Classification with Computer Vision Learning Studio (CV Studio). CNN architecture is used to classify whether the image is of a healthy lung or an infected one. The study presented has an accuracy of 93% for determining healthy lungs with the infected covid lungs.

Figure 4 depicts the analysis of the covid-19 pneumonia detection in the lungs. The red line indicates loss per iteration and the blue line depicts the accuracy of the model.

Similarly the Fig. 5 portrays the analysis of wrist x-ray images to determining the cost train cost and validation accuracy where the blue line exhibits the accuracy per iteration. As the iteration increases the changes in the loss and accuracy is depicted.

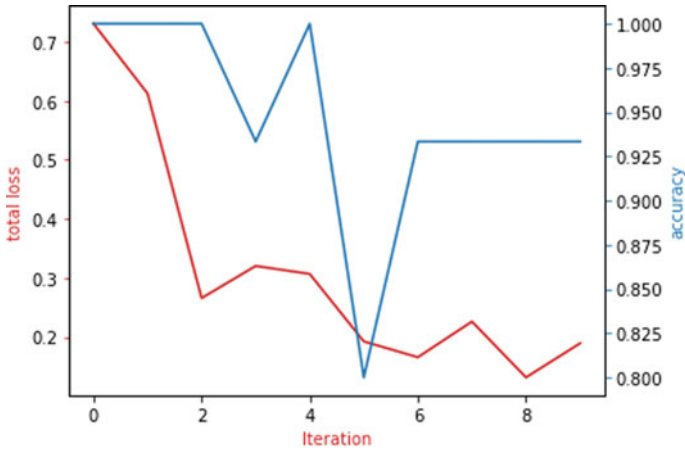


Fig. 4 Analysis of the study for covid-19 pneumonia detection

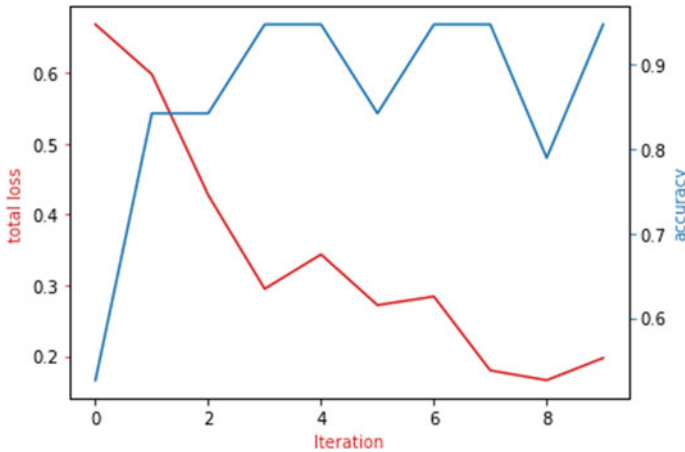


Fig. 5 Analysis of the study for wrist x-ray classification determining cost train cost and validation accuracy

The model again is trained for wrist x-ray images and is able to classify the normal wrist x-rays with the fractured one wherein the same values of loss and accuracy is shown per iteration.

For classification of lungs images for pneumonia we used transfer learning with Convolutional Neural Network for Classification with Computer Vision Learning Studio (CV Studio).

Figure 6 shows the snapshot of the application made wherein the user is able to classify the images just by uploading the images. The deep learning model classifies the x-ray images and is able to predict the outcome whether the lungs are or infected.

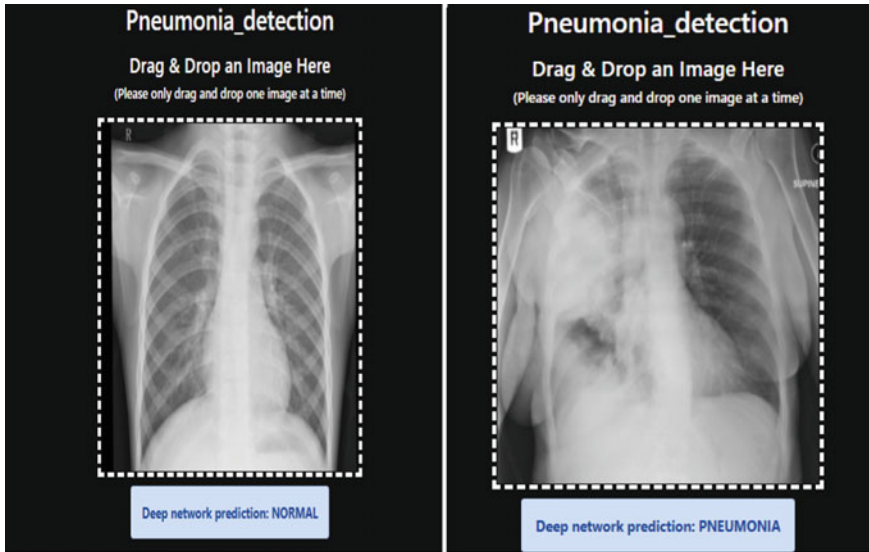


Fig. 6 Snapshot of the system predicting covid-19 infected lungs with the healthy one

5 Conclusion

Deep learning-based image analysis can be leveraged to build Image processing algorithms for various domains. This study helps in understanding how deep learning is used in various medical applications. The paper reviews various applications of image processing using deep learning in medical images and provides a generic model for disease classification based upon image analysis. The methods reviewed was mostly done for a single disease classification however the proposed method have been used for both pneumonia detection in lungs as well as for the classification of fractured bone images. The study has tremendously worked better than the previous methods used like SVM, KNN, etc. and also provided better results as compared to various studies presented. The study uses CNN architecture to classify lungs images of patients for pneumonia detection and was able to classify the normal lungs images from the infected ones with accuracy of 93%. The model further checked upon for classifying the wrist x-ray images and was successfully able to classify the normal wrist with the fractured wrist using the same architecture with accuracy 93.74%. The presented model could be further used for various disease classifications using the images.

References

1. Sarker IH (2021) Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions. *SN Comput Sci* 2:420. <https://doi.org/10.1007/s42979-021-00815-1>
2. Wu Q, Liu Y, Li Q, Jin S, Li F (2017) The application of deep learning in computer vision. In: IEEE 2017 Chinese Automation Congress (CAC), Jinan, China, 20–22 October 2017, pp 6522–6527. <https://doi.org/10.1109/CAC.2017.8243952>
3. Hatcher WG, Yu W (2018) A survey of deep learning: platforms, applications and emerging research trends. *IEEE Access* 6:24411–24432
4. Arel I, Rose DC, Karnowski TP (2010) Deep machine learning - a new frontier in artificial intelligence research [research frontier]. *IEEE Comput Intell Mag* 5(4):13–18. <https://doi.org/10.1109/MCI.2010.938364>
5. Rizk Y, Hajj N, Mitri N, Awad M (2019) Deep belief networks and cortical algorithms: a comparative study for supervised classification. *Appl Comput Inf* 15(2):81–93
6. Puttagunta M, Ravi S (2021) Medical image analysis based on deep learning approach. *Multimed Tools Appl* 80:24365–24398. <https://doi.org/10.1007/s11042-021-10707-4>
7. Liang C, Xin S (2020) Research status and prospects of deep learning in medical images. In: 2020 international conference on communications, information system and computer engineering (CISCE), pp 380–382
8. Lin W, Tong T, Gao Q, Guo D, Xiaofeng D, Yang Y, Guo G, Xiao M, Min D, Xiaobo Q (2018) Convolutional neural networks-based MRI image analysis for the alzheimer's disease prediction from mild cognitive impairment. *Front Neurosci* 12:777. <https://doi.org/10.3389/fnins.2018.00777>
9. Laddha M, Jindal S, Wojciechowski J (2019) Gastric polyp detection using deep convolutional neural network. In: Proceedings of the 2019 4th international conference on biomedical imaging, signal processing 2019. Association for Computing Machinery, New York, pp 55–59. <https://doi.org/10.1145/3366174.3366185>
10. Methil AS (2021) Brain tumor detection using deep learning and image processing. In: 2021 international conference on artificial intelligence and smart systems (ICAIS), pp 100–108. <https://doi.org/10.1109/ICAIS50930.2021.9395823>
11. Sharma A, Kumar R, Jaiswal V (2021) Classification of heart disease from MRI images using convolutional neural network. In: 2021 6th international conference on signal processing, computing and control (ISPCC), pp 358–363. <https://doi.org/10.1109/ISPCC53510.2021.9609408>
12. Muresan MP, Barbură AR, Nedevschi S (2020) Teeth detection and dental problem classification in panoramic X-ray images using deep learning and image processing techniques. In: 2020 IEEE 16th international conference on intelligent computer communication and processing, pp 457–463. <https://doi.org/10.1109/ICCP51029.2020.9266244>
13. Yadav DP, Rathor S (2020) Bone fracture detection and classification using deep learning approach. In: 2020 international conference on power electronics & IoT applications in renewable energy and its control, pp. 282–285. <https://doi.org/10.1109/PARC49193.2020.236611>
14. Chen J-Z (2021) Design of accurate classification of COVID-19 disease in X-ray images using deep learning approach. *J. ISMAC* 3(02):132–148
15. Malik H, Jabbar J, Mehmood H (2020) Wrist fracture - X-rays. *Mendeley Data V1*. <https://doi.org/10.17632/xbdsnzr8ct.1>
16. Kermany D, Zhang K, Goldbaum M (2018) Labeled optical coherence tomography (OCT) and chest X-ray images for classification. *Mendeley Data V2*. <https://doi.org/10.17632/rscbjbr9sj.2>
17. Fayyaz S, Ayaz Y (2019) CNN and traditional classifiers performance for sign language recognition. In: Proceedings of the 3rd international conference on machine learning and soft computing 2019. Association for Computing Machinery, New York, pp 192–196. <https://doi.org/10.1145/3310986.3311011>

18. Alzubaidi L, Zhang J, Humaidi AJ et al (2021) Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *J Big Data* 8:53. <https://doi.org/10.1186/s40537-021-00444-8>
19. Verma P, Tripathi V, Pant B (2021) Comparison of different optimizers implemented on the deep learning architectures for COVID-19 classification. *Mater Today Proc* 46(20):11098–11102

Restaurant Automation Through IoT and NLP Techniques



Partheesh Ranjan Singh, Tejas Kumar Nazre Amarnath,
Mohammed Khurram, Swathi Tripathi, and Deepti Chandrasekharan

Abstract Internet-Of-Things and NLP have been a trending topic for a long time. The use of these two domains has been instrumentally transforming the way restaurants are working. Restaurants have different tasks including staff allocation, food quality monitoring and menu automation module. These tasks can be performed with precision and accuracy with the help of IOT and NLP techniques. Greater flexibility in menus, an increase in restaurant productivity and capacity for extensive business auditing are the primary benefits associated with the restaurant management system. Food quality monitoring is also an important factor related to the health of the customers. In this paper, we present an architecture for menu and inventory management automation and also for food quality monitoring. The system can be accessed from android mobile application and web application.

Keywords Internet of Things · Natural Language Processing · Food quality · Restaurant · Automation etc.

1 Introduction

Restaurants are one of the business areas which employ a large number of people and serve a much larger number of customers whose number keeps increasing day-by-day. The operational tasks in the restaurant involve a lot of manpower. This involves a high risk of human error. Human error if large in magnitude, makes the business suffer a large cost in terms of profit and customer satisfaction. The customers are very particular with the quality of the food which is being served to them. The quality of the food served in the restaurant is the major contributor to the profit and reputation of the restaurant. The restaurant industry is moving towards reforms with automation as the main tool to achieve greater profits and greater customer satisfaction. The automation of a restaurant helps to reduce the effort for repetitive tasks and helps the restaurant owners to save time, money and resources. There are

P. R. Singh · T. K. Nazre Amarnath · M. Khurram · S. Tripathi · D. Chandrasekharan (✉)
Department of Computer Science Engineering, PES University, Bangalore, India
e-mail: deeptic@pes.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
P. P. Joby et al. (eds.), *IoT Based Control Networks and Intelligent Systems*,
Lecture Notes in Networks and Systems 528,
https://doi.org/10.1007/978-981-19-5845-8_10

123

various tasks in the restaurant which can be automated. Some of these are Menu Automation, Allocation of jobs to staff, Billing and Auditing etc. These tasks can be automated with the help of technologies like Internet of Things and Natural Language Processing. Internet of Things helps to connect different devices together which allows automation to become more feasible. NLP helps to make the automation more “human-like”, i.e., enabling proper communication with the system. The proposed work in this paper implements two modules, menu automation and food quality monitoring using Internet of Things and Natural Language Processing.

2 Literature Survey

RFID systems have been used in various applications like digital toll plazas. This has been explained in the work conducted by authors in [1], where a RFID is implicated in image of the car number. The RFID reader present in the toll plaza reads the RFID of the car and allows the car to pass through. Another example where RFID can be used is demonstrated in the work by N. S. Kumar et al. [2], where the authors have developed an alert system for notification regarding a filled dustbin. A master and slave architecture is followed by the system developed [3] by Harapanahalli et.al., where the Raspberry Pi acts as the Master and Arduino acts as a slave. The Arduino executes the command which it gets from the Raspberry Pi. The ARDUINO handles operations related to the RFID tags and readers. The entire GUI system is implemented on RASPBERRY PI which also handles the operations related to payment and other backend operations. The development of a smart kitchen cabinet as presented by the authors describes the usage of sensors to measure the weight of the items in the cabinet. The inventory is managed based on the weight of the items in the cabinet. The RFID antennas and tags are used for the identification of grocery items. The usage of an Android application has also been included in the work by authors [4]. The sensors are used to monitor the grocery items. The data from the sensors is sent to the cloud to be monitored by the Android application. The prototype developed is low-cost only in the range of 35 USD. In [5, 6], the work described uses an Android application that is built to inform the user about the items purchased. The products are added to the shopping list when the level of an item falls below a given threshold. The users are notified through the application about the products and a shopping list is generated. The database on a cloud platform is used to store the product levels and the position of the item. In this project, the kitchen wardrobe is monitored continuously through live video monitoring. Most of the related work in this area utilizes facial recognition for the authentication of the customer. For the experiments conducted on facial recognition using PCA and Euclidean distance, the ORL database was used. Pre-processing techniques were applied on the ORL database to increase the recognition rate. Different number of training and testing images were examined to investigate the performance. An increase in the number of training images led to an improvement in the recognition rates due to the rich information in the feature space. The action of resizing the images using 0.3 scale, resulted in the size of 34×28 pixels

for each image, thereby improving the recognition rate. An appropriate scale was carefully selected while resizing the images to avoid losing the face features. Small sized image reduced both feature dimension and computation complexity[7]. Multi-Layer Perceptron (MLP) with feed forward learning algorithms were chosen for the proposed system because of their simplicity and capability in supervised pattern matching. This approach has been successfully applied to many pattern classification problems [8]. A more recent approach to face detection with Gabor wavelets & feed forward neural network was presented in [9]. The authors in [10] described a method that used Gabor wavelet transform and feed forward neural network for both finding feature points and extracting feature vectors. The experimental results have shown that the proposed method achieves better results compared to the graph matching and eigenfaces methods, which are known to be the currently most successful algorithms. In [11], the work demonstrates how voice based chat bot can be used in restaurant industry to improve efficiency. The customer is authenticated with text based login. The authentication of the customer into the restaurant system is completed with the help of facial recognition. A CNN module implemented by face recognition biometrics is used to encrypt the face images.

The food recommendation system is also designed taking into consideration the popular algorithms of machine learning. Here, the authors have proposed a food recommender system that uses matrix factorization. For better prediction accuracy, rating information and user tags are fused. The essence of the recommender is formed by collaborative filtering. KNN algorithm is utilized for congregating similar users based on ratings. The authors have integrated the system with a web-speech API. This API will help to access the models which are trained on a vast amount of training data. This API performance is validated on the TSP dataset. The authors have written modules in spaCy and NLTK libraries. The items, names and quantities were extracted using the POS tagging (parts of speech). This has helped to boost accuracy and navigation for the system. The status of the food in preparation is tracked after the items appear on the ordered screen. The authors were able to observe that the whole website worked perfectly. The facial recognition of the customer achieved 89% accuracy.

The authors in [12] have developed an IoT based system in a refrigerator with the help of MQ3 sensor, DHT11 sensor for measuring gas presence, temperature and humidity respectively. The alerts can be sent to the user with the help of SMS and E-mail. Authors in [13] have received a patent on inculcating RFID for monitoring food freshness. I. D. S. Batista et al. [14] have described an intelligent platform with hardware and software that can monitor parameters like temperature, moisture and weight of the food in the buffet. The authors have described initial prototype configuration. The data generated by the sensors was received by the Arduino micro controller and ESP8266. The authors were able to demonstrate the results properly with greater accuracy in monitoring nutritional values of the food. Latencies were observed during the testing of the system as data flowed from the physical layer to view layer.

The allocation of jobs in an intelligent manner is always a concern in the fast food industry as time and efficiency is a substantial factor. K. Aytac et al. [15] have

observed that not much work has been achieved in this particular area. The authors have attempted to solve this by utilizing a genetic algorithm. Different sensors are connected for different tasks to be accomplished with a gateway. This particular gateway acts as a connecting highway between all the devices. Data received from the sensors is preprocessed, redundant data is eliminated and ML algorithms are applied for taking decisions. The order is given to the gateway, and this information is displayed on the order screens. These results are utilized to assign jobs to the restaurant staff after the optimization algorithm has been executed. In this work, IoT architecture has been proposed by the authors to enable fast food restaurants to intelligently allocate different jobs to workers with their own specialities. The unbiased allocation of jobs is done by a genetic algorithm, which uses Bet Prediction Selection. The costs were evaluated for various parameters and it was observed that there was a substantial decrease in the cost when the proposed method was applied as an initialization step. The key benefits of utilizing Internet of Things based smart staff allocator include savings with respect to time, resources and better allocation of tasks to staff members. Additionally, it allows full control and tracking of the task assignment system. The algorithm which is used is a genetic algorithm, which gives better results. The use of bet prediction selection as a selection method does not give substantial results while allocating tasks to staff.

3 Proposed Work

The proposed project is divided into two modules. These modules are menu automation and food quality monitoring. There are different actors who are involved in the functioning of the two modules. These are Customer, Admin, Hall Manager, Head Chef, Chef and Kitchen manager. The menu automation module involves actions from customer, Head chef, Hall Manager and Chef as described in Fig. 1. The Food Quality Monitoring involves the action from kitchen manager and chef. The menu automation module helps the customer to order food based on the available items in the menu with the help of a text based chat bot. The items in the menu are managed by the admin, who is actually the owner of the restaurant. Tasks like table booking and payment are handled by the Hall manager. The head chef and chef are the actors who take care of the order being delivered to the customer. They confirm the order from the customer. The information about the prepared food is communicated to the hall manager. The customer will be served and subsequently payment process is initiated by the customer.

The hardware for the menu automation includes Arduino microcontroller, NodeMCU Wi-Fi module, MFRC522 RFID reader, and buzzer with RFID cards and keychains. The software components used for implementing the menu automation module includes HTML, CSS, JavaScript, and Bootstrap. The databases used are a combination of SQL and NoSQL databases which are PostgreSQL and Firebase. The chat bots are designed with the help of the tool named Dialogflow by Google. The food quality monitoring module is utilized to monitor the quality of the food

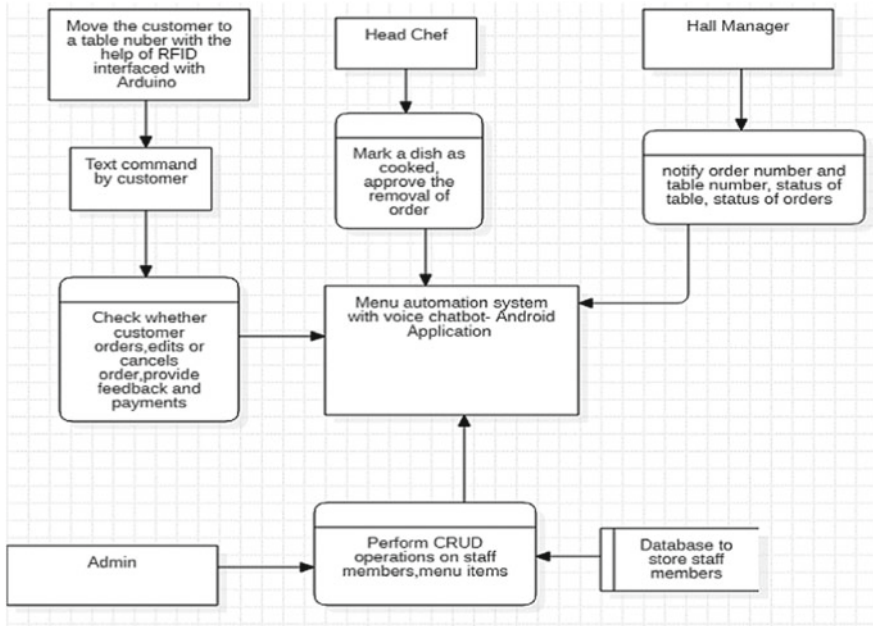


Fig. 1 The system architecture for menu automation module

which is prepared. The kitchen manager is able to check the values of the sensors which are attached to the system that detect the quality of food. The sensors detect the changes in the environment continuously and it can be checked anytime by the Kitchen Manager. The kitchen manager tells the chef to take an appropriate action. The communication between the actors is completed with the help of a text based chat bot which helps the kitchen manager to get the values easily and notify the chef.

The hardware includes sensors to detect the changes in the environment which can affect the quality of food. There are 3 sensors which are used in this work. These are DHT11 temperature and humidity sensor, MQ3 gas sensor and dust particle sensor. These sensors are connected with Arduino microcontroller and NodeMCU Wi-Fi module as shown in the Fig. 2. The software components utilized include HTML, CSS, JavaScript and bootstrap. The database used here is Firebase. The chat bot is designed using Dialogflow by Google.

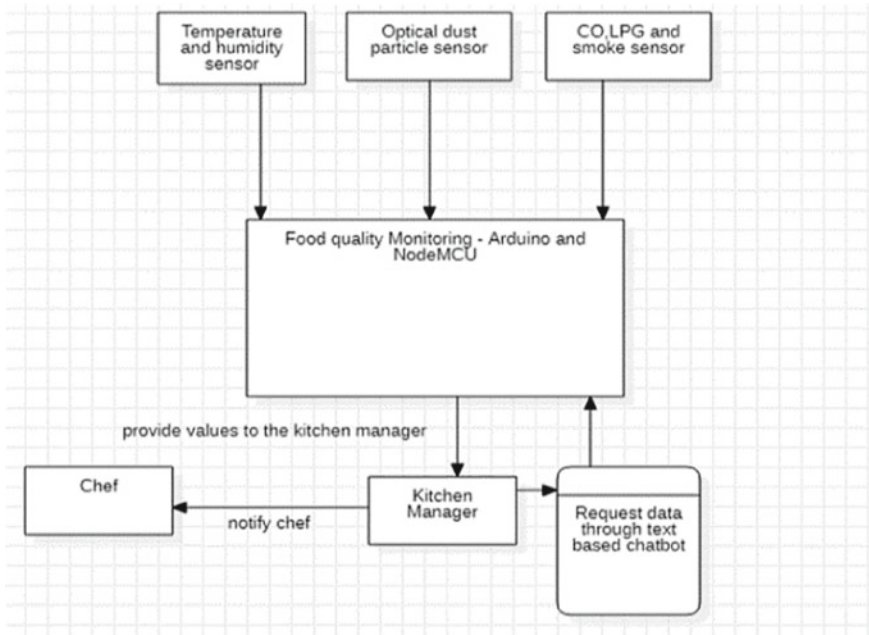


Fig. 2 The system architecture for food quality monitoring

4 Results

The main pipeline where the customer is involved in the system starts from here. The customer books a table in the restaurant so that he can order a dish. A customer is not allowed to order without reserving a table. The customer swipes a card in front of the RFID reader which alerts the hall manager about the occupancy of a particular table. The hall manager accepts the request and the customer is allowed to order.

The chat bot assists the customer to order food items after the reservation has been confirmed as shown in Fig. 3. The chat bot functions in Hindi and English. The chat bot is trained to understand some English words, if they are used in Hindi sentences.

For example, “Chinese” is a cuisine so that will also be used in Hindi. The chat bot is able to detect similar words in both languages. This is shown in Fig. 4.

The customer gives final confirmation and quantity of the items as shown in Fig. 5 and Fig. 6.

The order from the customer now goes to the head chef who confirms the same as shown in the Fig. 7.

The chef can also confirm the same order if he is free. Here it is assumed that head chef or chef can confirm the order depending on their availability. The chef marks the item as “Done”, i.e., the item is cooked, as shown in Fig. 8.

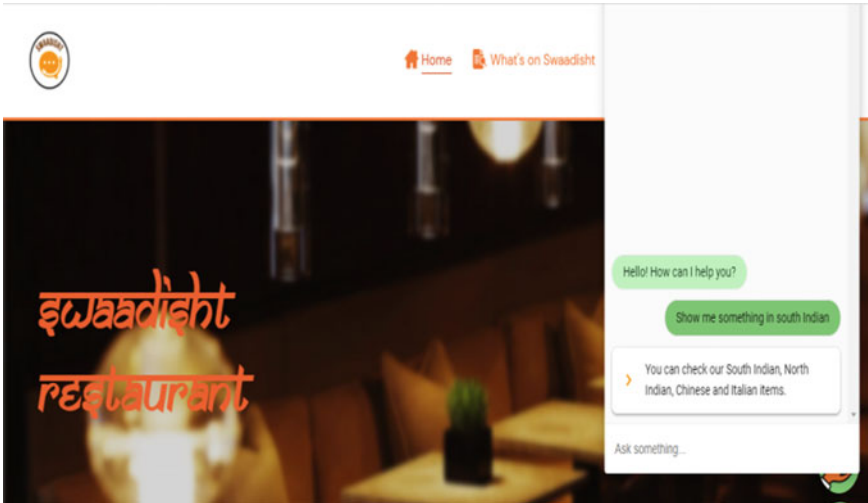


Fig. 3 A customer requesting the chat bot for cuisines

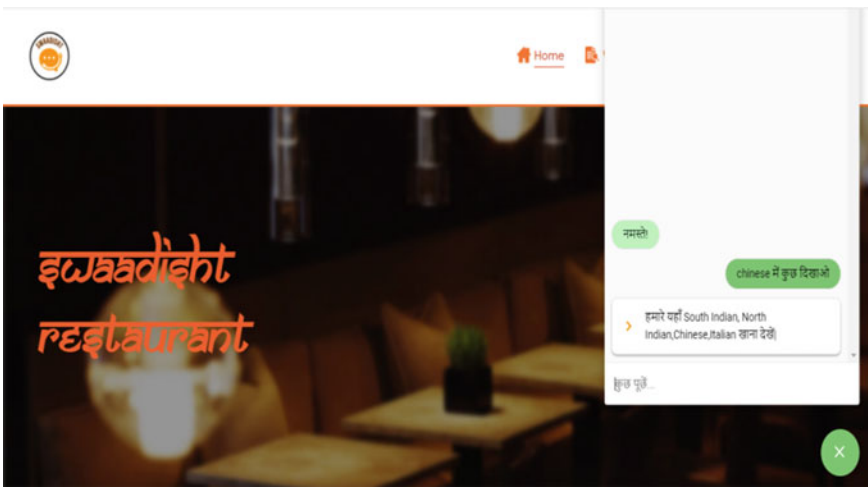


Fig. 4 Customer requesting for the cuisines in Hindi

After the chef marks the item as “Done”, the head chef will confirm the item once more. This is done so that each and every item cooked is approved by the head chef. This is shown in Fig. 9.

After these steps, the hall manager plays the role of billing the customer for the food served to him as shown in Fig. 10.

The customer can pay the bill using several payment options like Cash, card and UPI. The current work described in the paper limits the payment options to Cash.

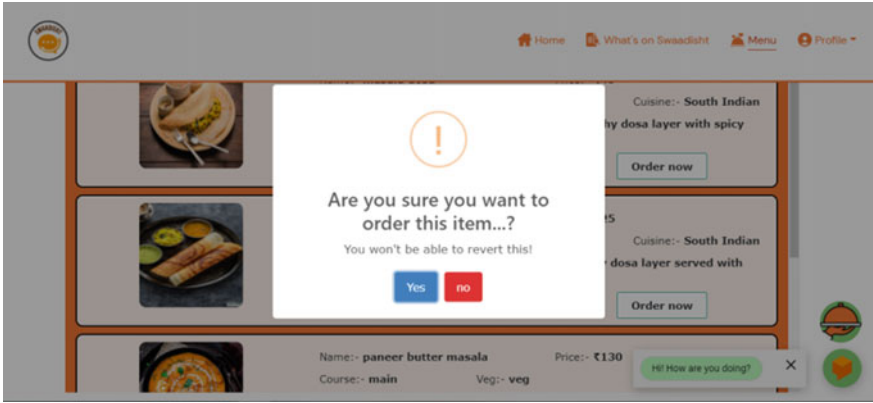


Fig. 5 A customer orders an item after selecting the cuisine

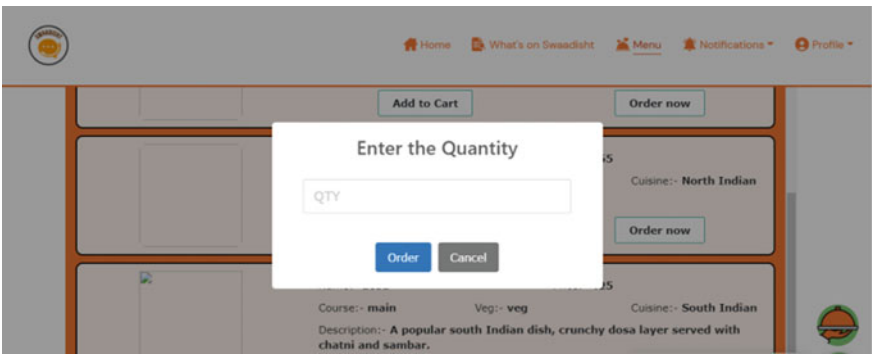


Fig. 6 The customer enters the quantity of the item

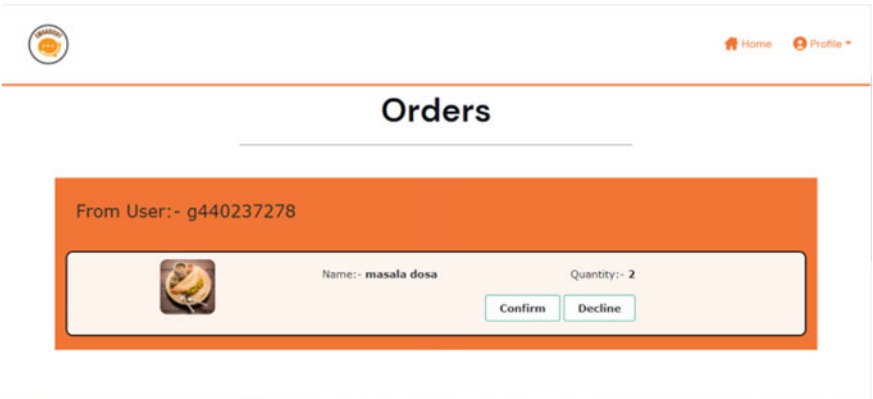


Fig. 7 The head chef checks the order from customer

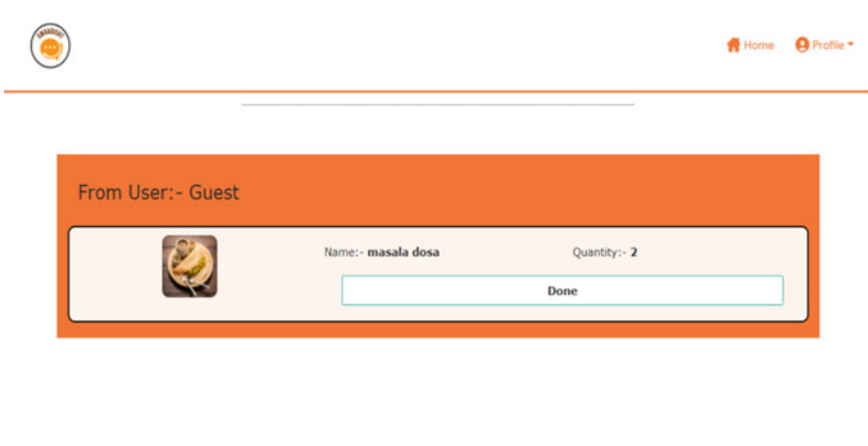


Fig. 8 The chef marks the item as “Done”

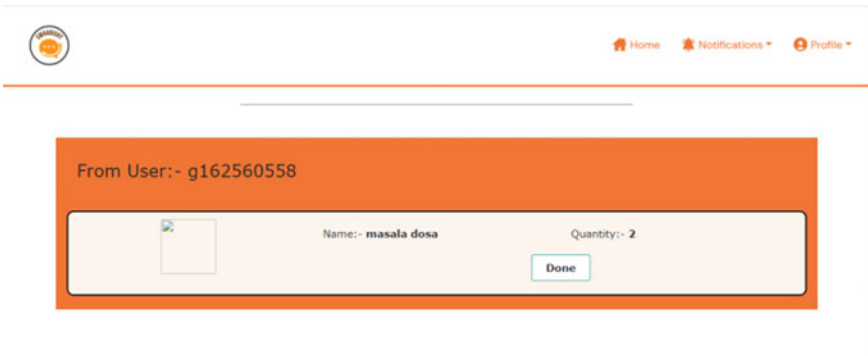


Fig. 9 The head chef marks the dish as done from the chef

Other options like UPI and Card can be included in future versions. The customer will be able to check the payment status as shown in Fig. 11.

The status will change after the hall manager confirms the payment. This is shown in Fig. 12. The pipeline where the customer interacts with the system is completed here.

This system authorizes the admin to manage the menu items and staff as shown in Fig. 13.

The admin is able to edit the items in the menu. The admin has to add the name, image, price, course, Veg or Non-Veg, Cuisine and Description as shown in Fig. 14.

The admin is able to delete the items in the menu as shown in Fig. 15.

The admin can set an item for the menu for today. The menu for today will be visible to the customer when he visits the website. Another feature given to the admin is staff management. The admin is able to add the staff members in the restaurant.



Fig. 10 The hall manager marks the dish as SERVED from here

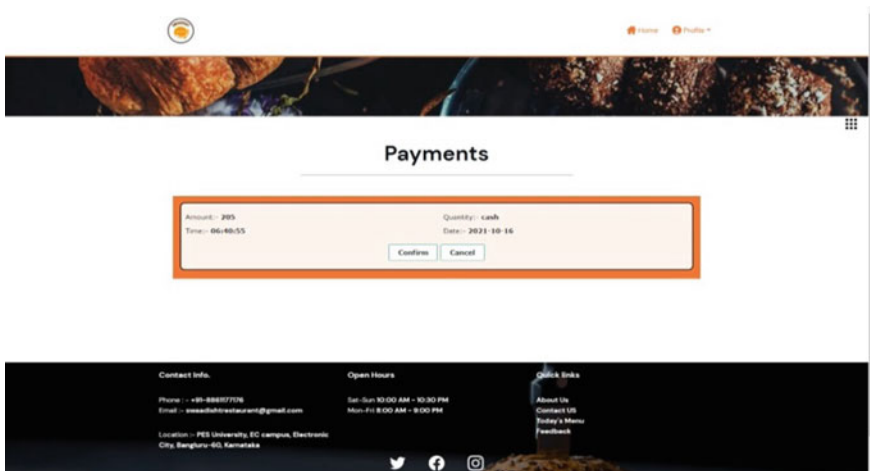


Fig. 11 The hall manager checks the payment by customer

The details of the staff members are added. The admin operations are available for management of staff and menu items. The admin will be able to organize the menu items in the restaurant and provide the customer with the variety of food and choice of cuisine. The admin is the back bone of the menu automation module because the admin is required to provide the menu to the customers in a fast, efficient and different manner.

The admin can delete an employee if required, as shown in Fig. 16.

Another prominent part of the described work is the food quality monitoring module. The kitchen manager has a chat bot which helps him to communicate in

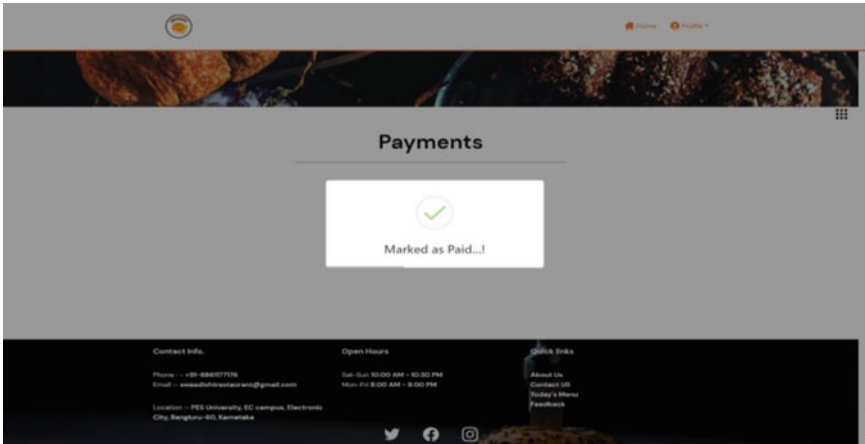


Fig. 12 The hall manager marks the bill as paid

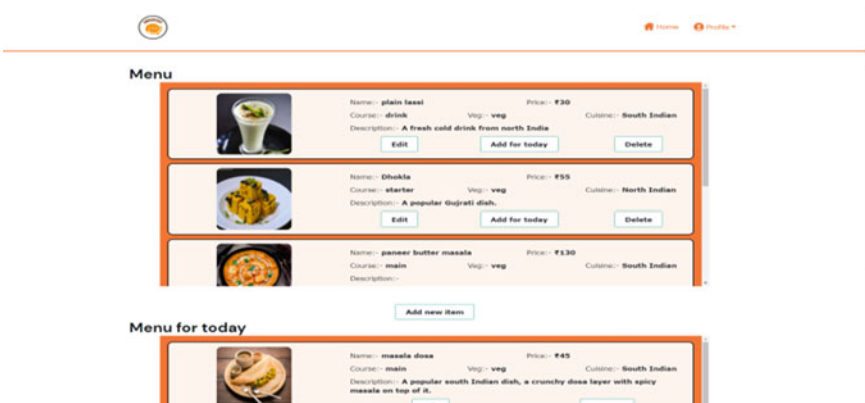


Fig. 13 The admin checks the menu items and menu for today

Hindi and English. This bilingual chat bot also helps the kitchen manager to get the values of sensors and notify the chef accordingly.

Figure 17 depicts the actions performed by the chat bot. The chat bot greets the kitchen manager. The kitchen manager can query the chat bot regarding the situation of the sensors. The chat bot then replies with the relevant data and displays a message that reads “Check the values here”, “Please alert the chef if humidity value is greater than 58, temperature more than 30, and on the presence of any gas or dust concentration”. This information is passed to the kitchen manager by the chat bot.

Figure 18 shows how the kitchen manager alerts the chef on detecting poor quality of the food.

The same procedure is also repeated in Hindi as shown in Fig. 19.

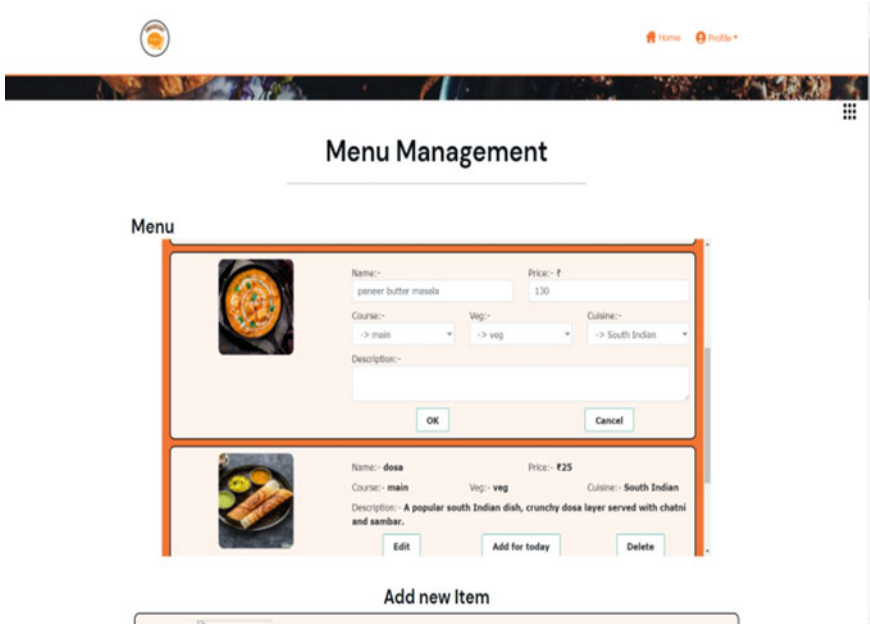


Fig. 14 The admin edits the item in the menu

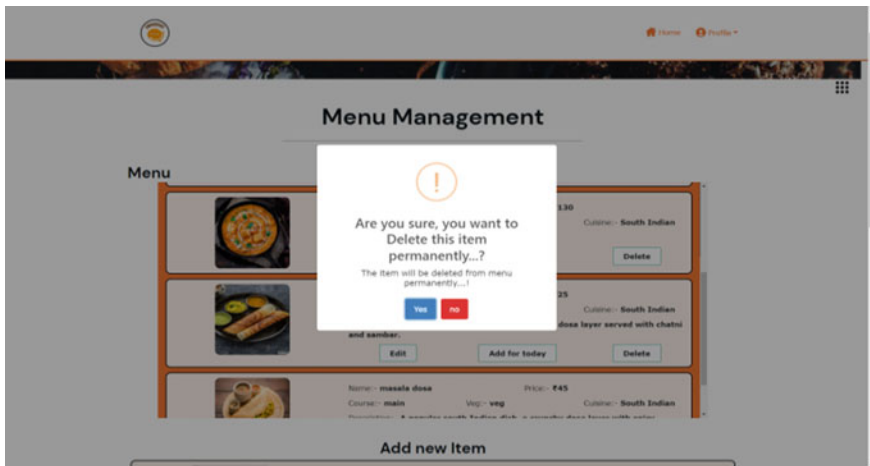


Fig. 15 The admin is able to delete the items

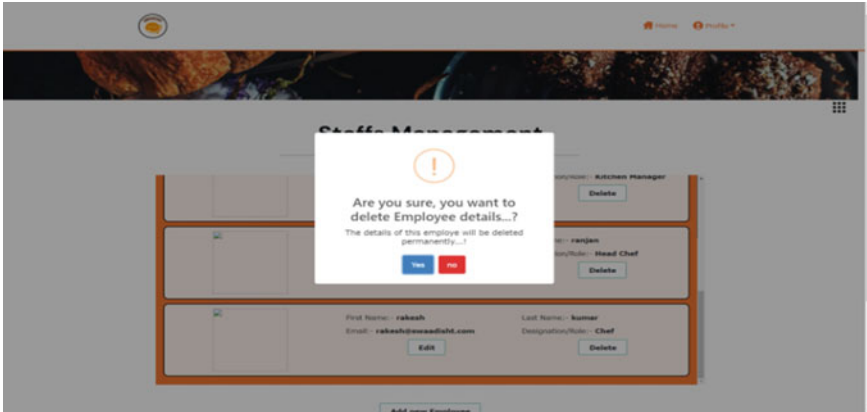


Fig. 16 The admin can remove the employee

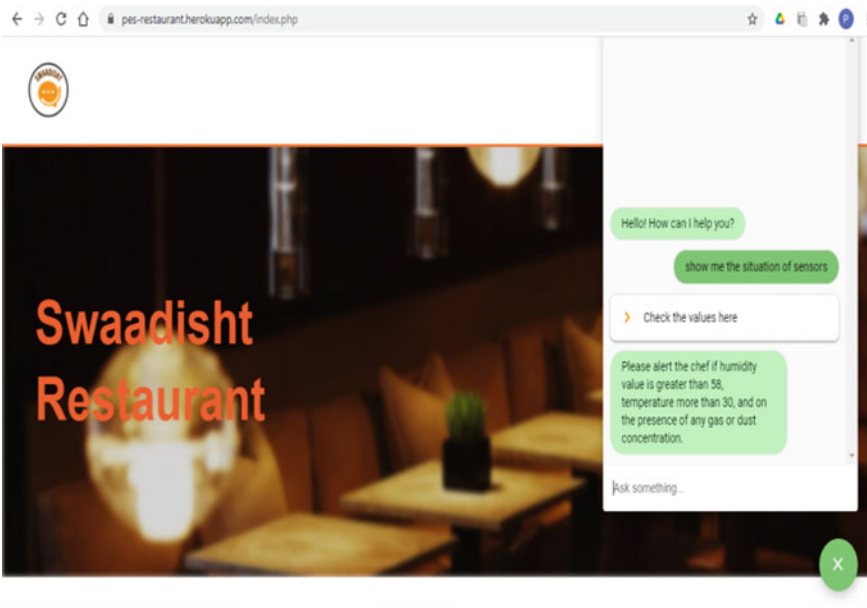


Fig. 17 The chat bot requests the sensor values in English

The chat bot will alert the kitchen manager about the actions to be taken in Hindi. Hindi. The kitchen manager can instruct the chef in Hindi to take action based on the observed sensor values as shown in Fig. 20.

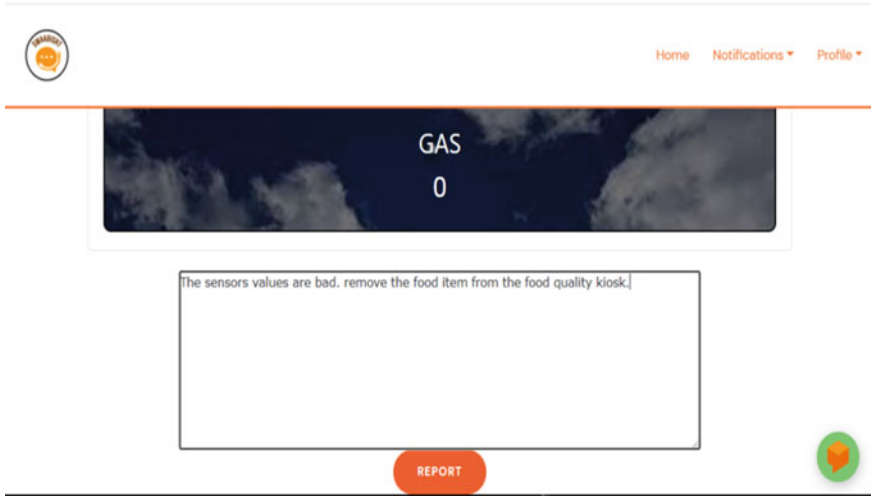


Fig. 18 The kitchen manager alerts the chef in English

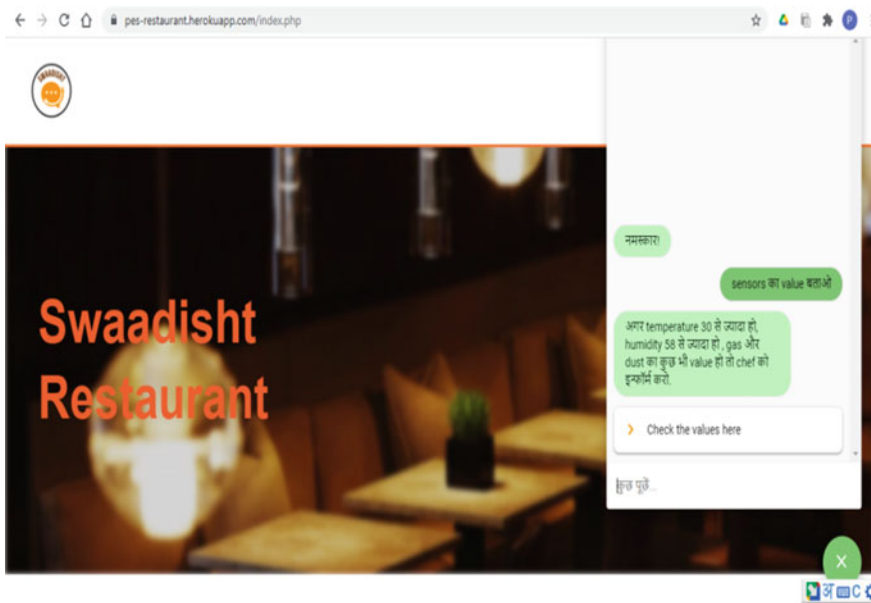


Fig. 19 The chat bot requests the sensor values in Hindi

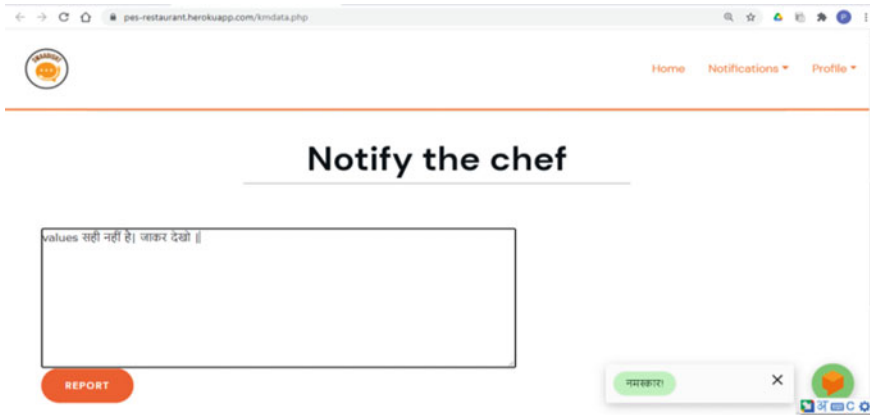


Fig. 20 The kitchen manager alerts the chef in Hindi

5 Conclusion and Future Work

The proposed project implements menu automation and food quality monitoring modules as described above. The project is able to simulate the functions which are performed in a restaurant like booking a table by the customer, ordering by the customer, managing the preparation of food by the staff in the restaurant and payment of bill by the customer. Apart from this, management of staff and menu items by the admin or the owner of the restaurant can also be performed. A very important and unique part of this work is the food quality monitoring system which enables the restaurant staff to monitor the food by checking the values like temperature, humidity, gas and dust concentration. There is also an alert system if there are any issues in the observed values. The proposed project can be improved based on the business needs. Future improvements to this work may include various other functions performed in the restaurant like management of electrical appliances, making a chef to cook a dish according to his/her specialty. A voice recognition chat bot can be included for visually impaired users. There can be a provision for offers, discounts and coupons to regular customers. The monitoring of food quality can be made more precise with the help of additional sensors. The mobile application can also be made compatible for IOS mobile systems.

References

1. Ahmed S, Tan TM, Mondol AM, Alam Z, Nawal N, Uddin J (2019) Automated toll collection system based on RFID sensor. In: 2019 international carnahan conference on security technology (ICCST), Chennai, India, pp 1–3. <https://doi.org/10.1109/CCST.2019.8888429>

2. Kumar NS, Vuayalakshmi B, Prarthana RJ, Shankar A (2016) IOT based smart gar-bage alert system using Arduino UNO. In: 2016 IEEE region 10 conference (TENCON), Singapore, pp 1028–1034. <https://doi.org/10.1109/TENCON.2016.7848162>
3. Harpanahalli J, Bhingradia K, Jain P, Koti J (2020) Smart restaurant system using RFID technology. In: 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), pp 876–880. <https://doi.org/10.1109/ICCMC48092.2020.ICCMC-000162>
4. Amutha KP, Sethukkarasi C, Pitchiah R (2012) Smart kitchen cabinet for aware home. In SMART 2012, The First International Conference on Smart Systems, Devices and Technologies, pp 9–14
5. Salah Uddin M, Khan M, Ali D (2019) Kitchen Grocery Items Monitoring System Based on Internet of Things. *Int J Comput Netw Technol* 7(2)
6. Balaji A, Sathyasri B, Vanaja S, Manasa MN, Malavega M, Maheswari S (2020) Smart kitchen wardrobe system based on IoT. In: 2020 International Conference on Smart Electronics and Communication (ICOSEC), Trichy, India, pp 865–871. <https://doi.org/10.1109/ICOSEC49089.2020.92154591>
7. Barnouti NH (2016) Improve face recognition rate using different image pre-processing techniques. *Am J Eng Res* 5(4):46–53
8. Li X, Areibi S (2004) A hardware/software co-design approach for face recognition. In: The 16th international conference on microelectronics, Tunisia
9. Kaushal A, Raina JPS (2010) Face detection using neural network & gabor wavelet transform. *IJCST* 1(1)
10. Bheleet SG, Mankar VH (2012) A review paper on face recognition techniques. *Int J Adv Res Comput Eng Technol* 1(8):339–346
11. Raju G, Akshay J, Dhavan R, Kumawat A (2020) Speech oriented virtual restaurant clerk using web speech API and natural language processing. *Int J Eng Res* 9. <https://doi.org/10.17577/IJERTV9IS050684>
12. Nasir H, et al (2018) The implementation of IoT based smart refrigerator system. In: 2018 2nd international conference on smart sensors and application (ICSSA). IEEE, pp 48–52
13. Lindsay J (2006) RFID system and method for tracking food freshness. US6982640, Accessed 27 May 2004
14. Batista IDS, Sardina IM, Dantas RR (2019) Monitoring restaurants in real-time. In: 2019 II workshop on metrology for industry 4.0 and IoT (MetroInd4.0&IoT), Naples, Italy, pp 202–206. <https://doi.org/10.1109/METROI4.2019.8792882>
15. Aytac K, Korçak O (2018) IoT based smart staff allocator in quick service restaurants. In: 2018 23rd conference of open innovations association (FRUCT), Bologna, Italy, pp 1–7. <https://doi.org/10.23919/FRUCT.2018.8588019>

Comprehensive Assessment of Big Data in Recommendation Systems



Swati Dongre and Jitendra Agrawal

Abstract Developments in web-based e-commerce platforms cause recommendation systems to gain increasing importance. Recommendation systems are systems developed to provide valuable and personalized recommendations for users. In the age of big data, existing recommendation systems face scalability and efficiency problems in the face of increasing numbers of users and products. Within the scope of this study, a comprehensive and comparative review of big data and recommendation systems has been made. Studies in which big data are used in recommendation systems have been examined in the literature. The necessary pre-processes and methods for applying big data to recommendation system with high performance and success have been discussed in detail.

Keywords Adverse drug reaction · Collaborative filtering · Classification technique · Drug recommendation · Drug repositioning · Healthcare recommendation system · Recommendation system · Social media

1 Introduction

With the developments in hardware and software technologies, data collection processes have become accessible and continuous. Daily transactions performed on the Web using credit card transactions, mobile devices, or personal computers can be stored automatically. Similarly, advances in information technologies enable a large amount of data flowing between IP networks. This large amount of data can be used to extract different patterns from various applications [1].

S. Dongre (✉) · J. Agrawal
Department of Computer Science, School of Information Technology, RGPV, Bhopal, India
e-mail: dswati31@gmail.com

J. Agrawal
e-mail: jitendra@rgtu.net

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
P. P. Joby et al. (eds.), *IoT Based Control Networks and Intelligent Systems*,
Lecture Notes in Networks and Systems 528,
https://doi.org/10.1007/978-981-19-5845-8_11

139

Today, where speed is becoming more and more important in the internet environment, traditional advice systems cannot respond quickly to changes in user preferences. They cannot capture the interests of users in real-time. In the age of big data, advisory systems need to express response times in milliseconds and cope with high computational complexities. For example, “What is the proportion of male users between 20 and 30 years old who clicked on an advertisement from Beijing during the last ten seconds?” Considering the query, it is necessary to calculate the combination of four dimensions: region, age, gender and advertisement. Such questions cause high computational costs in large flowing data [2].

Data are continuously obtained in today’s developing applications, inflowing data rather than data sets stored finely. Such floating data includes stock market data, network traffic measurements, Web server records, and click flows on the Web, measurements obtained from sensor networks and mobile communication records. It can be given as. Operations performed on flowing data differ from traditional data mining approaches as the size of flowing data increases over time, and response times to queries should be short. For this reason, it is not possible to scan all the flowing data to store the musical data entirely and respond to the questions [3].

Several research and computational difficulties arise because the volume of background data is so large. As the volume of data increases, it is impossible to process the same data by passing it many times effectively. The ability to process a data item at most once puts restrictions on the implementation of underlying algorithms. Therefore, algorithms to be used in streaming data mining should be designed to operate with a single pass over data [4]. In most cases, this process has a natural temporal component, as data can evolve. This behavior of flowing data is called temporal locality. Therefore, straightforward adaptation of single-pass streaming data mining algorithms may not be an effective solution for the task. Streaming data mining algorithms need to be carefully designed to focus on the development of the underlying data [4].

Recommendation systems have become increasingly popular with the widespread use of electronic commerce applications in the Internet environment. However, most of the recommendation systems developed are not designed in real-time. Real-time interactions cannot be established in online environments, as existing systems cannot operate effectively in large-scale systems. The main problem in applying big data algorithms to recommendation systems is that the storage space for operations performed on memory is minimal. In addition, big data increases the cognitive load on users, causing scalability and user satisfaction problems for recommendation systems. For this reason, recommendation system should have a scalable structure against the increasing number of users and products in larger data sets and calculation costs without reducing their performance. Within the scope of this study, research has been conducted on the applicability of big data in recommendation systems. With the study in which extensive data are used in recommendation systems. The methods used in big data and extensive data analysis were examined comprehensively by evaluating the recommendation systems from the point of view of big data. Studies in the literature on applying big data technologies such as Hadoop and Spark in large-scale recommendation systems have been evaluated. The methods used in

recommendation systems, big data sources and technologies used in big data analysis have been analyzed comparatively. In the continuation of this chapter, studies in the literature on the application of flowing data algorithms to recommendation systems are reviewed.

In 2016, a new probabilistic neighborhood-based method was developed to provide real-time recommendations. The assumption in referral system implementations using streaming data is that not all evaluations of a particular user or assessment of a specific item can be taken simultaneously. Advisory system implementations are different from multidimensional streaming apps where all dimensions of the same record are constantly retrieved simultaneously. In recommendation system applications, the user can evaluate a particular item at any time. Also, new users or articles can be added to the system at any time.

In general, the number of users and items increases over time, as reviews are never deleted. Therefore, it is assumed that the number of users at time t is $m(t)$, the number of items at time t is $n(t)$, and the size of the evaluation matrix at time t is $m(t) \times n(t)$. The users' evaluations about the items are taken in the form of (User Id, Item Id, Rating Score). User ratings were taken as binary, with $+1$ expressing dislike and 1 expressing dislike. It may be desirable to specify a list of suggestions to be presented to users in online application areas, or a list of users who are interested in a particular item having a single user evaluation update on the flowing data does not produce accurate results in classical neighborhood-based approaches because it requires recalculation of the distance between things. Also, not all evaluation matrix is expected to be kept in memory. The positive ($+1$) evaluations made by the users until the time t are $P(i, t)$ and $P(j, t)$, and the negative evaluations are $N(i, t)$ and $N(j, t)$ for the i and j elements whose neighborhood distances are to be calculated. It is expressed with. The similarity between elements i and j at time t Eq. It is estimated as seen in 1.

$$S^+(i, j, t) = \frac{P(i, t) \cap P(j, t)}{P(i, t) \cup P(j, t)} \quad (1)$$

This similarity, calculated based on positive evaluations, is the Jaccard index. $\alpha = |P(i, t) \cap P(j, t)|$ and $\beta = |N(i, t) \cap N(j, t)|$. By weighing positive and negative evaluations, including and Spouse. 2, is achieved.

$$S(i, j, t) = \frac{\alpha \cdot S^+(i, j, t) + \beta \cdot S^-(i, j, t)}{\alpha + \beta} \quad (2)$$

To predict user evaluations, the similarity of a particular item with all items is first calculated. The weighted average of the ratings of the items most similar to a specific item is envisaged as the user rating for that item. The set of items similar to the item that the user evaluates, and the user represents the evaluation score for item neighbourhood-based prediction score at time t for the user and element i is calculated using 3.

$$P_{u,i}(t) = \frac{\sum_{j \in I_{i(u)}} S(i, j, t) \cdot r_{u,j}}{\sum_{j \in I_{i(u)}} S(i, j, t)} \quad (3)$$

The probabilistic neighbourhood-based methods developed within the scope of the study and the min-hash technique, in which the similarities between the items are calculated probabilistically, are proposed. Similarities are approximated by tracking each user in the min-hash index. The basic idea is to apply sort order to users using a hash function. In this sort of order, the probability that the first user makes a positive evaluation for item i is the first user to make a positive evaluation for item j is similar to the Jaccard index. The $(.)..(.)$ hash functions are applied to users who evaluate j positively and negatively. After the hash functions are implemented, the users who make positive evaluations are kept in the $M +$ data structure. The users who make negative evaluations in the $M -$ data structure. These data structures can be stored in memory because their size is smaller than the user-item evaluation matrix and can be easily updated. The similarity between elements i and j Co. Calculated using 4.

$$S^+(i, j, t) \approx R^+(i, j, t) = \frac{\sum_{s=1}^d \delta(i_s = j_s)}{d} \quad (4)$$

$\delta(.)$ is a function that takes value 0 if items i and j are similar and 1 is not similar [1].

In a study in 2015, limited hardware resources and time constraints were used to provide real-time news advice.

A new method has been proposed, which is optimized in its subjects. The system called PLISTA aims to provide fast and efficient suggestions by responding to suggestion requests within 100 ms. When a user visits a news portal in the developed system, the portal sends a suggestion request to the PLISTA server. The PLISTA server transfers the request to a randomly selected recommendation algorithm. To respond to the request, the chosen recommendation algorithm must submit a suggestion list within 100 ms. With the offline test server called VAGRANT4, the performance of the proposed algorithms is analyzed using interaction data of users who have logged in in the past. The architecture of the developed system is shown in Fig. 1.

One of the limitations of the developed system is identifying unique users since users do not log in to news portals. This makes it difficult for user-based methods to obtain comprehensive data to calculate personalized recommendations.

Considering the limited lifetime of news articles in the developed system, the most reviewed news articles in a predetermined time are determined according to user-item interaction statistics and presented as suggestions. The underlying idea of this approach is that the most popular themes can also be of interest to users who haven't seen them yet. The popularity measure of an article is $r(a)$, where $p(a)$ is the number of users who read the article a and $t(a)$, the time the article a was published. It is calculated as seen in 5.

$$r(a)_{T,P} = \log_{10}(\max(\text{abs}(p(a) - T), 1.0)) + \text{sign}(p(a) - T) \cdot \frac{\text{now} + T(a)}{P}) \quad (5)$$

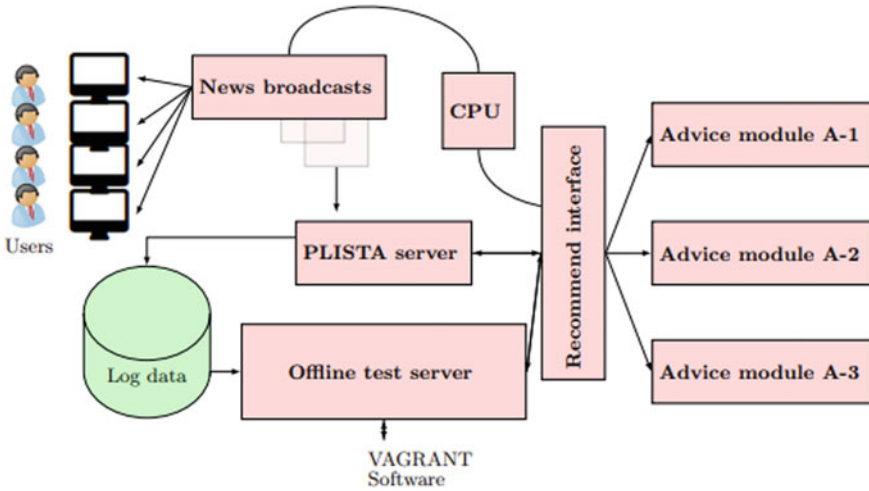


Fig. 1 Architecture of the developed system

Here p determines the effect of article age on the popularity measure. On the other hand, T determines how many users an article should have read before it earned a recommendation value. The abs function returns the absolute value of the expression, and the $sign$ function returns the sign of the given term, positive or negative.

Similar to classical collaborative filtering approaches, relationships between articles are calculated according to standard features. For example, if a user has read two articles, these articles are said to be related. The algorithm provides a set of related articles for each known article. With this approach, when a suggestion is presented to a user, a group of reports obtained by searching for articles related to the paper that the user has read before is used.

The developed system was analyzed in terms of the resource consumption rate of the system when user interactions were most intense, in terms of parallel requests, the maximum number of updates and impressions, response time, and test scenarios for how many portals an algorithm can respond to with the same hardware resources [5, 6].

In the study conducted by D. R. Staçkcli in 2021, a new method for flow-based recommendation systems is proposed by making queries on the flowing data and calculating a set of personalised recommendations. These queries cover relational algebra operations and data mining operations. $U = u_1, u_2, \dots, u_n$.

Set of users, $I = i_1, i_2, \dots, i_m$. To represent the set of items and time T , the result of the interaction of each user u with the item i at time t is obtained directly or indirectly as an evaluation score.

Evaluation score, calculated as $R = [(u, i, r, t)|(u, i, r, t)EUxIxRxT]$.

RECSYS queries are created with the model, which is trained by using the data of the evaluation input flow to prepare the developed model, and suggestion lists are made for each request. The assessment flowing data is divided by the EXTRACT

TEST DATA module into two flowing data, the test data flow and the learning data flow. Learning data is used by the TRAIN RECSYS MODEL module to train a model. With each new learning group, this operator updates the model and makes a copy for subsequent operators. This operator sets the validity intervals of outgoing models according to the following rules:

- At every point in time, there is one valid pattern. This means that when a new model is valid, the previous model should be invalid.
- All learning sequences used to train the model must be within the model's validity range and not a learning group that is not used within the validity range of the model. It also limits the number of evaluations held in memory.

In each proposal request, a series of proposal candidates are determined with the RECOMM CANDIDATES module. These candidates are usually items that the user has not evaluated. The PREDICT RATING (PREDICT RATING) module uses models to estimate the evaluation score of each recommendation candidate. The RECOMMEND module selects the items to be suggested based on the evaluation results. The TEST PREDICTION operator compares estimated and actual evaluation scores by applying an evaluation metric such as RMSE [6, 7].

In the study conducted by S. Zhang in 2021, a new algorithm was developed to analyse user-item interactions in online news portals and provide flowing data-based recommendations to meet quality, robustness, scalability and tight time constraints. Daily news data are collected from news portals and discussion platforms and analyzed. Flowing data from the PLISTA competition (ACM Recsys News Challenge 2013) were analyzed to determine the properties and hidden rules of the flows. PLISTA is a platform where advertisers and publishers come together, allowing researchers to evaluate their algorithms under real-world scenarios. PLISTA's purpose is to suggest interesting articles to users. Every time a user visits a web page on the news portal, article suggestions are created and placed on the suggestion page. In this contest, when a user visits the news site participating in the quiz program, the PLISTA server selects a random advisory team and transmits the request to this team. The algorithm of the chosen team offers six suggestions, depending on the demand. Suggestion requests are answered within 100 ms, including the communication time. Team performance is measured by the number of tips clicked by users. The problems encountered are that it is difficult to identify unique users because they are not logged in. The suggestions submitted by the participating teams affect user behavior. Another problem is the users' feedback time for the proposed suggestions. Some users click the suggestions right away, while others wait for days.

ACM Recsys 2017 data were used as training data within the scope of the study. Training data consists of 84 million user click data and approximately 1 million suggestion click events. The user click-through rate of suggestions and online analysis were used as evaluation criteria. The click rate is generally meagre, as most users are only interested in the articles on the Website and do not pay attention to the recommendations offered. Due to the placement of suggestions on the web page, the habits

of the user groups, and the influence of time, the click-through rate is mainly dependent on the portal. Online evaluations determine to what extent the recommendation algorithm accurately predicts user clicks and calculates the precision value.

With the community-based approach implemented, the most appropriate algorithm is selected for suggestion requests. It is decided with which algorithm the incoming requests will be answered based on context with a decision tree. A trend-based approach was used to determine whether news articles were suitable as suggestions. In this approach, based on the assumption that the algorithm's success shortly will be booming in the future, the data of the last 60 min are used. With the filling approach used in creating the suggestion list, each element of the suggestion list consisting of 6 elements is forwarded to several different suggestion algorithms. The highest-ranked recommendation for each recommendation agent is used for a request in the result set. The underlying idea of community strategy is that collecting recommendations from different recommendation algorithms will lead to diversity, as each recommendation algorithm uses its specific criteria when calculating.

Analysis results showed that in offline assessments (without tight time constraints), the community strategy achieved approximately 5% better recommendation precision. The advantage of the community strategy is that the system constantly adapts to changes in the user's behaviour. The disadvantage is that the final result can only be completed after all suggestion algorithms have completed their calculations [7].

P. Liu, In 2019, a time-based advice system called *TeRec* was proposed by expanding online scoring approaches. In *TeRec*, users can receive hashtag recommendations based on their real-time interests while tweeting and generating quick feedback for requests. *TeRec* provides a browser-based client interface that allows users to access real-time topic suggestions and processes and stores real-time streaming data on the server-side. Working in a streaming data environment (WSDE), *TeRec* provides real-time advice to its users based on their preferences at any time. *TeRec* models users and items using matrix factorisation to get more accurate results. The basic idea of *TeRec* is to use hashtags as proxies for exciting topics. When a user is about to post a tweet, the system predicts the user's current interests and suggests various topics (hashtags) they might want to use in this tweet.

TeRec offers a browser-based service that helps users choose the appropriate hashtags when they tweet. The server side of the system models and stores user preferences, calculates evaluation score estimates between users and hashtags and provides a hashtag suggestion list when users tweet. The developed system consists of three layers, as seen in Fig. 2. The first layer is the user interface, where interactions between users and data, such as displaying suggestion results and receiving user feedback, are carried out.

A matrix of user preferences and item properties is kept in the storage layer. The third layer is the layer where suggestions between the storage layer and the user interface are created, and the model is updated. The working process of *TeRec* works as follows: In 1 step, the user interface takes the requests from the users and asks for recommendations. The recommendation model calculates the evaluation scores and returns the suggestion list to the user interface in the second step. In step 3, the

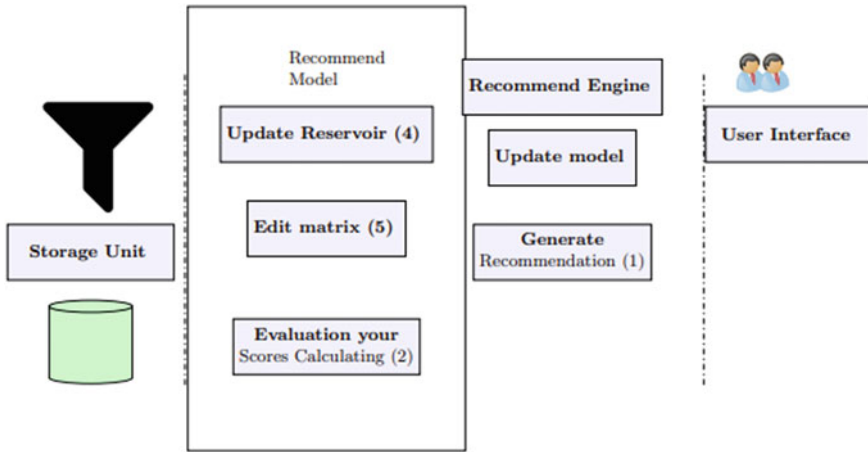


Fig. 2 TeRec architecture

user interface updates the suggestion model by receiving user feedback. In step 4, the recommendation model updates the data store (reservoir) of past entries. In step 5, the recommendation model updates the matrix using the updated pool.

Given that the data can be extensive and useless, a subset of informative inputs (reservoir) needs to be sampled to speed up model updates. For this purpose, a reservoir sampling technique has been used. When the size of data entries reaches c , in Vitter’s algorithm t , data c/t will likely be preserved. In the reservoir sampling mechanism presented within the scope of the study, t . data will be kept with a probability of $1 - c/t$. The decision to put new data in the reservoir given, the location of the data sample in the reservoir is probability of augmentation $1 - P(s_i \in R_{t-1})$. It is calculated using 6.

$$P(s_i \in R) \propto \exp \frac{1}{t - i} \tag{6}$$

$1 - P(s_i \in R_{t-1})$ The exponential decay function commonly used in time series analysis, $t-i$, denotes the difference between the current time order t and the time order i of the data sample received by the system.

In the developed system, the reservoir, which stores the samples of the last preferences of the users, is dynamically updated with each new data and always produces its suggestions according to the latest updated model. Analysis results showed that the TeRec system is more successful in offering real-time hashtag suggestions for tweet streams. The reservoir, which stores the samples of the users’ last preferences in the developed system, is dynamically updated with each new data and always produces its suggestions according to the latest updated model. Analysis results showed that the TeRec system is more successful in offering real-time hashtag suggestions for tweet streams [8].

2 Big Data Based Recommendation Systems

Recommender systems use users' past purchases and evaluation data to provide personalised recommendations. The volume of data currently available requires a reassessment of the methods used to generate recommendations in recommendation systems [9]. Parallel and distributed data processing, which is the basis of big data, should form the basis of algorithm design. Traditional parallel computing environments such as OpenMP and MPI and distributed computing platforms such as MapReduce and Spark are used for this purpose [10].

The two main problems addressed by recommendation systems are assessment score estimation and creating a best-N recommendation list. The purpose of the review score estimation is to calculate the rating a user will give for an item. The purpose of making the Best-N suggestion list is to present a suggestion list consisting of N items that the users will be interested in and probably will like [11]. The two most commonly used approaches to calculating suggestions are neighborhood-based and hidden factor model-based. Neighborhood-based methods are used to calculate recommendations based on similarities between items or users. In hidden factor model-based approaches, users and items are matched within the same hidden area, and the items closest to a user in this area are presented as suggestions. Remote factor model-based approaches are widely used to obtain evaluation estimates, while neighborhood-based methods are widely used to get the best-N suggestion list [12].

The increasing amount of data causes significant data analysis problems in recommendation systems. Recommendation systems often experience scalability and inefficiency issues when processing or analysing large-scale data [13]. In this section, studies in which big data analysis are performed in the recommendation systems in the literature are examined.

Meng et al. In 2014, a Keyword-Aware Service Recommendation Method called KARS was developed to provide users with personalised and practical suggestions. Keywords were used to indicate users' preferences, and a user-based collaborative filtering method was used to generate appropriate recommendations. In the developed system, keywords extracted from close users' reviews were used to determine user preferences. To increase scalability and efficiency in the big data environment, the KARS system has been developed on the Hadoop platform using the MapReduce parallel processing paradigm. In the developed method, two data structures, namely the keyword candidate list and the specialised domain name dictionary, were used to obtain the users' preferences. The keyword candidate list expressed as $K = k_1, k_2, \dots, k_n$, where n is the number of keywords in the keyword candidate list, is a set of keywords related to users' preferences and multiple criteria of candidate services, and a can be words. Preferences of active users and close users are matched to keyword clusters. Calculation of users close to the active user Jaccard similarity is used for [14].

S. Balakrishna et al. In 2020, it was aimed to generate useful information from these data by managing unstructured extensive healthcare data in a secure environment and turning the data into a useful, practical model. Within the scope of the study,

it was aimed to create an application system for the early diagnosis of diseases. The application system was created by using the Naive Bayes (NB) classification algorithm working on Apache Mahout to make recommendations about the users' health conditions, provide treatment optimisation, and prevent adverse events [15].

B. HOSSEINI et al. In 2019, a distributed matrix computing library based on Spark, was proposed in the study conducted by 2015. Matrix computing, which is the basis of data analytics applications such as social network mining, recommendation systems, and natural language processing, is inadequate in terms of such data sizes and calculations, as matrix scales grow in the big data age. The developed Marlin library provides high-level matrix calculation with the distributed matrix processing algorithms it contains. Experimental results have shown that Marlin is faster than distributed matrix processing algorithms based on R and MapReduce [16].

G. S. BHATHAL et al. In 2109, a study conducted by Hadoop in 2015, a recommendation system provides many data such as reviews, previews, opinions, complaints, statements, feedback and comments on any item such as products, events, individuals and services on the Web was developed. In the developed system, a hybrid filtering technique was used to filter different data types such as examination, opinion, explanation, comment and complaint. Recommendations offered to users are based on user reviews, content, the behaviour of the reviewer and the timing of studies generated by different users. The developed system was tested with additional sized files of the MovieLens dataset on the Hadoop platform. The resulting graph shows that parallel to the increase in file size, the data size in the form of evaluation, analysis and feedback also increased. Still, the data processing time did not grow at the same rate [17].

W. Yin et al. In 2021, a recommendation system was developed to provide users with personalized route suggestions using extensive route data. Updates using the developed system models and driving preferences of different drivers. To create a personalized route proposal, it is aimed to develop the most effective subset routes for the route to be determined according to the source and destination specified by the driver and the departure time. Fifty-two thousand two hundred eleven taxi drivers analyzed the system designed in Beijing. Test results have shown that the system developed provides efficient and effective advice [18].

S. Zhang et al. In 2021, a system called TencentRec was developed using the Storm platform to provide real-time and accurate recommendations on big data. Streaming data was analysed using Storm with a designed data access component and a data storage component. Item-based collaborative filtering, content-based filtering, and filtering methods based on demographic features have been applied for different applications. Suggestion changes are presented in real-time using the real-time data collection and processing process. The developed system consists of pre-processing layer, algorithm layer and storage layer components. The preprocessing layer parses the received data, filters the unqualified data, and sends it to the algorithm layer. The algorithm layer is responsible for the main algorithm calculations. Collaborative filtering, content-based filtering, and demographic-based filtering methods are applied at the algorithm layer. The storage layer filters the results produced by

the algorithm layer according to the rules of different applications and updates the calculation results [7].

K. Rattanaopas et al. In 2017, a new recommendation system was developed with collaborative filtering methods on Apache Hadoop using the MapReduce paradigm due to the scalability and productivity problems traditional recommendation systems face in the analysis of increasing user and product data. The developed system consists of Hadoop nodes, distributed recommendation engine and HBase storage. Amazon product dataset is used in the developed system. Similarities were calculated using the Pearsons Correlation Coefficient method in the collaborative filtering method. User recommendations generated are stored in the HBase distributed database. HBase uses Bloom Filter, which reduces over search on disk [19].

D. Dessa et al. In 2019, A micro-video recommendation system were developed in the study conducted by 2016. The developed system provides the producer of the video with information such as how many users liked the video. It also includes video suggestions to users by analysing users' favorite videos and viewing history. Data were collected from video sites and forums using Web crawler software in the developed system. Using the data obtained, a micro-video model and user model were created. The data obtained with the web crawler software was stored on the Hadoop platform, and Mahout was used to processing the data. A slope algorithm has been used, enabling collaborative filtering methods to be performed on Mahout [20].

M. Jakomin et al. In 2018, the sRec system, which can update in real-time for infinite and variable size flowing data, was developed. In the developed system, the input flow is modelled as user feedback activities, new users and new items. sRec provides real-time recommendations by determining the changing number of users and products and content shifts in user preferences. sRec consists of an online recommendation module and an offline parameter learning module. The system constantly updates its model to capture user dynamics and provide real-time recommendations. In the sRec system, user-item evaluations are modelled with a time-based function. Offline parameter updates are performed only when events occur and based on previous occasions. On the other hand, the online suggestion module offers suggestions according to the last probability distribution of the preferences made by the users using the user-item-time models created [21].

In a study conducted by Prando et al. In 2017, a new recommendation system based on users' interactions on social networks was developed to determine users' preferences on e-commerce platforms. The developed recommendation system analyses the social network data of new users with content-based filtering techniques and assigns users to specific categories. User preferences have been obtained using the posts directly made by the users, the shares they like and the pages they want. The developed system has been tested on an e-commerce platform to solve the cold start problem. The results of the analysis showed that the developed system yielded successful results for new users accessing the e-commerce platform for the first time [22].

R. A. Hamid et al. In 2021, a new method was developed to determine the relationships between users and interests, using a vector called the user position vector. The developed system consists of a user information collection module, user clustering module, location information collection module, user-location vector calculation module and location profile module. User profile information and information such as age and gender are collected from Facebook using the user information collection module. The user information obtained is clustered to determine close users by using the k-means algorithm. Location information is collected from trip advisors and travel blogs. User location vectors are calculated according to location information and user profile information. On the other hand, the location profile module matches the interests of the users extracted from the user profiles with the locations visited by the close users according to their interests [23].

E. Inan et al. In 2018, a movie recommendation system was developed using the item-based collaborative filtering method and Hadoop programming model. The developed system aimed to store the increasing data volume by using the distributed file system DFS and MapReduce and increase the algorithm's performance and the response speed of the system by processing the data in parallel. The developed system has been tested on the Movielense database. Experimental results show that the system provides high efficiency and reliability in large data sets compared to classical methods [24].

S. Mudda et al. In 2019 A new method for calculating the affinity between users in social networks is presented in a study conducted by 2017. By analyzing the big social data on Twitter, a personalized recommendation system based on friendship has been developed that suggests topics or interests to users. In the developed system, comparative experiments were performed according to precision, recall, f-criterion and average absolute error metrics using one-month Twitter data. Experimental results revealed that the developed method is more successful in determining the degree of affinity between users and calculating personalised suggestions [25].

M. Almaghrabi et al. In 2020, a new deep learning-based model was developed as a solution to the cold start problems experienced by collaborative filtering approaches. In the developed model, the problems encountered in cases where user evaluations about the products are missing or not at all have been tried to be overcome. A neural network-based deep learning architecture was used to extract the content properties of the items. The collaborative filtering model, which models the temporal dynamics of user preferences and item properties, has been modified to predict the content properties of items with cold start problems using Singular Value Decomposition (SVD). The developed system has been tested on a large data set containing Netflix evaluations. Test results showed that the developed model performs better than classical models in estimating the evaluation scores of the items with cold start problem [26].

2.1 *Data Collection Used in Recommendation Systems Approaches*

During the estimation phase of the items to be offered to the users as suggestions, information such as the attributes of the users, their behaviors or the contents of the resources accessed by the users are used to create user profiles and models. For example, in e-learning platforms, cognitive skills, mental abilities, learning styles, interests, preferences and interactions with the system are used to create user profiles [6].

The quality and success of the recommendations presented in the recommendation systems are directly related to the user profiles to be created. It is necessary to have as much information about the users as possible to provide suggestions that can meet the users' expectations. It is based on the fact that indirect feedback obtained through observing users' behavior will reflect user preferences more accurately [27].

Indirect Feedback. Indirect feedback is obtained by observing user actions such as users' purchasing history, browsing history, time spent on web pages, links clicked by the user, and clicks on the user interface. Indirect feedback reduces user burden by removing user preferences from users' interactions with the system. Indirect feedback stands out as a more objective approach, as it does not require any user effort and is obtained through direct analysis of user behaviour [28].

Direct Feedback. In systems where direct feedback is used, users are expected to evaluate the items in the system via a user interface. The accuracy of the recommendations made in these systems depends on the number of reviews made by the user. The shortcoming of this method is that it requires user effort, and the users are not always ready to provide enough information. Even though direct feedback requires more user effort, inferences from user behavior are unnecessary, so the data obtained is seen as more reliable and transparency is ensured in the recommendation process [29].

Hybrid Feedback. These methods can be used together to minimise the weaknesses of direct and indirect feedback and to achieve the best performance. Hybrid feedback can be performed by indirect feedback added as a control mechanism over the immediate feedback or by taking the users' opinions on the items while receiving indirect feedback [30].

2.2 *Methods Used in Recommendation Systems*

It is vital to use effective and correct suggestion techniques to provide valuable and quality advice to users. The features and potentials of different proposal submission methods should be determined correctly and evaluated according to the system to be used.

Content-Based Filtering. Content-based filtering methods are the domain-dependent algorithms used, and the analysis of item properties is of great importance for generating predictions. Content-based filtering techniques show higher success in application areas such as web pages, broadcasts and news suggestions. In content-based filtering methods, user profiles are created according to the properties of the items that users have evaluated in the past. Mostly, items with characteristics close to items that users have positively assessed are presented as suggestions. Content-based filtering methods use different models to calculate similarities between items to produce meaningful recommendations. He can use probabilistic models such as the Vector Space Model such as TF/IDF or the Naive Bayes Classifier, Decision Trees and Artificial Neural Networks to model the relationships between different items. These techniques make recommendations through statistical analysis or by learning the model based on machine learning methods. Content-based filtering methods can organize suggestions to be presented quickly against possible changes in user profiles. The major disadvantage of this method is that it requires in-depth knowledge and explanation of the properties of the elements [31].

Collaborative Filtering. Collaborative filtering methods are estimation method used in application areas with content that cannot be quickly and sufficiently defined, such as movies and music. Collaborative filtering works by creating a database (user-item matrix) in which users have their evaluations of items. It matches the relevant and relevant preferences of the users by calculating the similarities between the user profiles at the stage of submitting suggestions. The active user has not seen before but is positively evaluated by the users in his neighbourhood are presented as suggestions. Guesswork and recommendations can be generated by collaborative filtering. The estimate is expressed with $R_{i,j}$, which indicates the evaluation that user i will make for product j . The recommendation is an item list with N elements that users will like most, as seen in Fig. 3 [32].

User – Based Methods

In user-based methods, users like the target user are designated as the user's neighbours. Similarities between users are usually calculated using the cosine similarity between the vectors representing the evaluation scores of the two users or the Pearson correlation coefficient. To denote K neighbouring users, the equation seen in Eq. 7 is used to estimate the evaluation score of the user u for item i .

$$\hat{r}_{ui} = \mu_u + \frac{\sum_{v \in N_u^k(i)} \text{sim}(u, v)(r_{vi} - \mu_v)}{\sum_{v \in N_u^k(i)} \text{sim}(u, v)} \quad (7)$$

$\text{Sim}(u, v)$ refers to the similarity between u and v users, and μ_u and μ_v indicate the mean of the evaluation scores of the u and v users. On the other hand, k refers to the set of close users who evaluate item i [33].

Item – Based Methods

Item-based methods calculate the evaluation score of the u user on the target item by using the similarities of the items scored by a particular user to the target item.

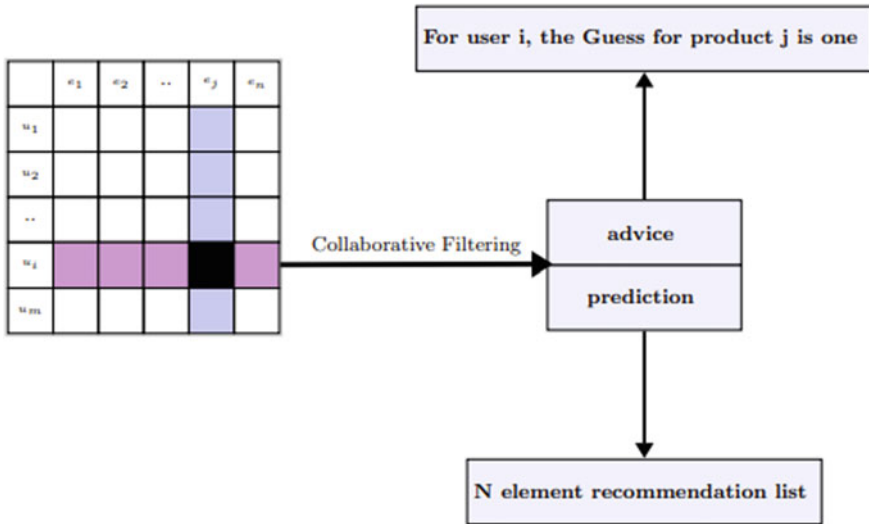


Fig. 3 User-item evaluation matrix

Similar to user-based methods, cosine similarity or Pearson correlation coefficient can be used to calculate the similarity between items. The evaluation score of user \$u\$ about item \$i\$ can be calculated using the equation seen in Eq. 8.

$$\hat{r}_{ui} = \mu_i + \frac{\sum_{j \in N_i^k(u)} sim(i, j)(r_{uj} - \mu_j)}{\sum_{j \in N_i^k(u)} sim(i, j)} \tag{8}$$

\$Sim(i, j)\$ expresses the similarity between the \$i\$ and \$j\$ items, and \$\mu_i\$ and \$\mu_j\$ represent the average of the evaluation scores for the \$i\$ and \$j\$ items. On the other hand, it refers to \$k\$ close items similar to \$i\$, which user \$u\$ evaluates.

Collaborative Filtering Based on Hidden Factor Model

Hidden factor-based methods are based on finding properties that define the properties of objects. Item features and user preferences are expressed with numerical factor values, as seen in Fig. 4. Estimates of user evaluations are derived from the models obtained by combining a smaller number of parameters.

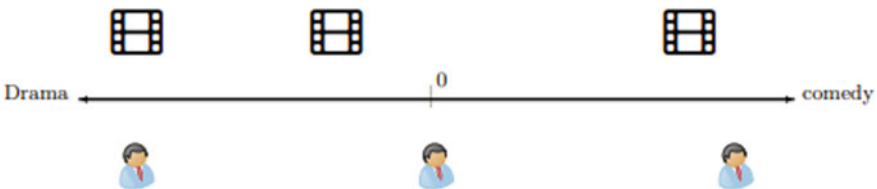


Fig. 4 Hidden factor models

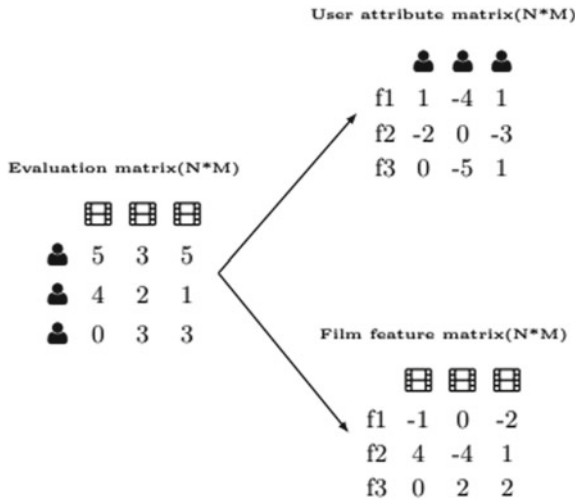


Fig. 5 Item and user attribute matrices

Matrix factorisation methods are used in hidden factor model-based approaches. Matrix factorisation methods are used for evaluation score estimation. It is taken as the basis that the user and element matrix P and Q can be calculated using R, which denotes the user evaluation matrix. The evaluation score of the user u on the i item is determined by associating the items and users with the factor vector as shown in Fig. 5. It is calculated using 9.

$$\hat{r}_{ui} = P_u^T q_i \tag{9}$$

Spouse, The matrix filled using 9 is used to determine the items that the users did not evaluate [34].

Hybrid Methods. Hybrid methods are created by combining different recommendation techniques to eliminate some of the limitations and problems of recommendation systems and achieve better system optimisation. The idea behind hybrid methods is that another will eliminate the disadvantages of one algorithm, and combinations of multiple algorithms can provide more accurate and compelling recommendations than a single algorithm. Hybrid systems can be created by applying algorithms separately and combining the result, using content-based filtering methods in collaborative filtering methods, or creating a unified recommendation system that combines both approaches [35].

2.3 Contextual Suggestions

Apart from the users' ratings, there is a great deal of contextual data that can improve the quality of the recommendation. Contextual data can be time, location, or additional information associated with users, items, or ratings. Contextual suggestions allow the modelling of relationships by considering each context as a different dimension. Thus, instead of a two-dimensional evaluation matrix consisting of users and items, it creates a multi-dimensional relationship model (user, item, time, location) [36].

2.4 Neighbourhood-Based Approaches for Big Data Scaling

Filtering and near-nearest neighbour-based approaches use data structures to reduce the number of computed similarities. Other methods select a user or subset of items to identify neighbours with no guarantee of high-quality results. The performance increase is achieved through multithread and distributed systems. For the determination of neighborhoods, k-nearest neighbor and e-closest neighbor approaches are used. The K-nearest neighbor (kNN) course aims to find k objects in the closest community to a query object. The e-nearest neighbor (eNN) or similarity search approach seeks to find all things with at least e similarity to the query.

Nowadays, closest neighbouring methods have been proposed for sparse vectors in which non-neighbouring object pairs are ignored or filtered with an increasing amount of data. The vector representing a user or item evaluation is sparse as users generally do not evaluate most of the items. Search methods use inverse index structure to prevent comparison of objects with no properties. The index structure creates an array, one for each property among all entities this way j . For the feature, a non-zero r_i , i elements with a value of j and a list of j (i, r_i, j) is obtained [37].

The methods used to determine the neighbourhoods work memory-based. A memory shared parallel data processing methods aim to minimise the working time of threads and resource competition. The distributed solutions available to identify the closest neighbours often use the MapReduce framework. With the MapReduce framework, objects can be divided into smaller subgroups, and the nearest neighbour search methods can be applied among block pairs. Some block comparisons can be eliminated based on block-level filtering techniques.

2.5 Evaluation Score Estimation

Recommendation systems mainly aim to estimate the user evaluation scores for the rest of the items using their ratings on certain items. A system consisting of n users and m items is expressed by an $N \times M$ size R matrix containing the r, u, i evaluations

made by u users for i items. The R matrix is sparse due to items that users have not yet evaluated.

Matrix decomposition techniques are used to divide the user-item evaluation matrix into low-level matrices. By this means, it is aimed to complete the missing values in user-item evaluation matrices. To use matrix factorisation methods in recommendation systems, least squares and gradient descent algorithms are used. Least squares is one of the matrix completion algorithms applied to large scale data. The factor vectors can represent elements and users to represent the number of m users and n the number of items. The purpose of the Least squares method is to find an estimate of parameters that fit a function. On the other hand, the Gradient descent method is an optimisation algorithm that is widely used in the field of machine learning and makes many iterations for low computational complexity. It performs a simple and iterative optimisation process that assumes a cost function and initial values for the optimization variables. The gradient descent algorithm aims to find the minimum point of a function [38].

2.6 Evaluation of Recommendation Systems Metrics Used

The metric to measure the quality of recommendation algorithms may vary according to the algorithm. Metrics that measure the accuracy of recommendation systems are divided into two as statistical measures and decision support steps. The suitability of each metric depends on the characteristics of the data set and the types of tasks the system will perform.

Statistical accuracy measures evaluate the accuracy of the recommendation algorithm used by directly comparing the estimated evaluation scores with the evaluation scores of real users. Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and correlation are generally used as statistical accuracy metrics. The mean absolute error refers to the measure of the deviation of the recommendation from the user's specific value and Equivalent. It is calculated as seen in 10.

$$MAE = \frac{1}{N} \sum_{u,i} |p_{u,i} - r_{u,i}| \quad (10)$$

The $p_{u,i}$ prediction score for user u and item i , the evaluation score $r_{u,i}$ that user u gave for item i , and the value N represents the total number of evaluation points in the dataset. The lower the MAE value, the more accurately the recommendation system predicts user reviews. The square root of the mean square error Equivalent. As seen in 11, it is similar to the mean absolute error metric but gives higher weight to more significant deviations.

$$RMSE = \sqrt{\frac{1}{N} \sum_{u,i} (p_{u,i} - r_{u,i})^2} \quad (11)$$

Decision support measurements are sensitivity (Precision), sensitivity (Recall) and F-criterion (F - Measure). These criteria help users select high-quality items from the items available in the system: tenderness, Co. As seen in 12, it determines whether the items presented as suggestions are related to the user.

$$Precision = \frac{\text{no. of items selected from the suggestion list}}{\text{size of the suggestion list}} \quad (12)$$

Sensitivity, Co. As seen in 13th, it determines how many of the items chosen by the users were offered to them as suggestions.

$$Recall = \frac{\text{no. of items selected from the suggestion list}}{\text{total number of items selected}} \quad (13)$$

F-criterion is Spouse. As seen in the 14th, it enables the precision and sensitivity metrics to be calculated within a single metric [39].

$$F - \text{criterion} = \frac{2 * Precision * Recall}{Precision + Recall} \quad (14)$$

3 Conclusion

Recommendation systems have been used for a long time in most platforms serving on the Web, especially e-commerce. Recommendation systems aim to alleviate information overload on users by offering personalised content that may be of interest to users. In the age of big data, it is impossible to process the same data repeatedly due to the increasing user clicks in the recommendation systems and the increase in the amount of product and the volume of data in the background. A data item can be processed at most once causes restrictions on the classical recommendation system algorithms in the experience. For this reason, traditional data mining approaches used in recommendation systems are ineffective.

Within the scope of this study, studies using big data in recommendation systems were analysed, and the methods used in big data and big data analysis were examined comprehensively by evaluating the recommendation systems from the perspective of big data. Neighbourhood-based approaches and hidden factor model-based approaches used in recommendation systems are discussed in traditional recommendation systems and big data recommendation systems. As a result of the literature researches, it was seen that the studies were developed using MapReduce and Spark, which are parallel processing platforms. It has been observed that the matrix-based methods used in data analytics applications such as classical social network mining, recommendation systems and natural language processing are insufficient due to

the distributed structures of big data, so parallel and distributed data processing environments have come to the fore in the age of big data.

Parallel processing techniques for static large data sets are recommended for researchers who will work on big data and recommendation systems. More practical, valuable and fast recommendations can be obtained using all existing user and product reviews. It is recommended to use a decision model that can be updated according to the incoming data samples. It will not be possible to process all of the data repeatedly in the recommendation systems to be carried out with flowing data.

References

1. Al-Rifai SS, Shaban AM, Shihab MSM, Mustafa AS, Al-Halboosi HA, Shantaf AM (2020) Paper review on data mining, components, and big data. In: 2020 international congress on human-computer interaction, optimization and robotic applications (HORA), pp 1–4
2. Muangprathub J, Boonjing V, Chamnongthai K (2020) Learning recommendation with formal concept analysis for intelligent tutoring system. *Heliyon* 6(10):e05227
3. Alam I, Khusro S, Khan M (2021) Personalized content recommendations on smart tv: challenges, opportunities, and future research directions. *Entertain Comput* 38:100418
4. Churyumov G, Tokarev V, Tkachov V, Partyka S (2018) Scenario of interaction of the mobile technical objects in the process of transmission of data streams in conditions of impacting the powerful electromagnetic field. In: 2018 IEEE second international conference on data stream mining processing (DSMP), pp 183–186
5. José EF, Enembreck F, Barddal JP (2020) Adadrift: an adaptive learning technique for long-history stream-based recommender systems. In: 2020 IEEE international conference on systems, man, and cybernetics (SMC), pp 2593–2600
6. Staqckli DR, Khobzi H (2021) Recommendation systems and convergence of online reviews: the type of product network matters! *Decis Support Syst* 142:113475
7. Zhang S, Liu H, He J, Han S, Du X (2021) A deep bi-directional prediction model for live streaming recommendation. *Inf Process Manag* 58(2):102453
8. Liu P, Zhang L, Gulla JA (2019) Real-time social recommendation based on graph embedding and temporal context. *Int J Human-Comput Stud* 121:58–72
9. Zhang Y (2016) Grorec: a group-centric intelligent recommender system integrating social, mobile and big data technologies. *IEEE Trans Serv Comput* 9(5):786–795
10. Zhao X (2019) A study on e-commerce recommender system based on big data. In: 2019 IEEE 4th international conference on cloud computing and big data analysis (ICCCBDA), pp 222–226
11. Mohamed MH, Khafagy MH, Ibrahim MH (2019) Recommender systems challenges and solutions survey. In: 2019 international conference on innovative trends in computer engineering (ITCE), pp 149–155
12. Ahuja R, Solanki A, Nayyar A (2019) Movie recommender system using k-means clustering and k-nearest neighbor. In: 2019 9th international conference on cloud computing, data science engineering (Confluence), pp 263–268
13. Arote SS, Paikrao RL (2018) A modified approach towards personalized travel recommendation system using sentiment analysis. In: 2018 international conference on advances in communication and computing technology (ICACCT), pp 203–207
14. Meng S, Dou W, Zhang X, Chen J (2014) Kasr: a keyword-aware service recommendation method on mapreduce for big data applications. *IEEE Trans Parallel Distrib Syst* 25(12):3221–3231

15. Balakrishna S, Thirumaran M (2020) Chapter 7 - semantic interoperability in IoT and big data for health care: a collaborative approach. In: Balas VE, Solanki VK, Kumar RK, Khari M (eds) Handbook of data science approaches for biomedical engineering. Academic Press, pp 185–220
16. Hosseini B, Kiani K (2019) A big data driven distributed density based hesitant fuzzy clustering using apache spark with application to gene expression microarray. *Eng Appl Artif Intell* 79:100–113
17. Bhathal GS, Singh A (2019) Big data: hadoop framework vulnerabilities, security issues and attacks. *Array* 1–2:100002
18. Yin W, Sun Y, Zhao J (2021) Personalized tourism route recommendation system based on dynamic clustering of user groups. In: 2021 IEEE Asia-Pacific conference on image processing, electronics and computers (IPEC), pp 1148–1151
19. Rattanaopas K (2017) A performance comparison of apache tez and mapreduce with data compression on hadoop cluster. In: 2017 14th international joint conference on computer science and software engineering (JCSE), pp 1–5
20. Dessa D, Fenu G, Marras M, Recupero DR (2019) Bridging learning analytics and cognitive computing for big data classification in micro-learning video collections. *Comput Human Behav* 92:468–477
21. Jakomin M, Curk T, Bosnić Z (2018) Generating inter-dependent data streams for recommender systems. *Simul Model Pract Theory* 88:1–16
22. Choi S-M, Jang K, Lee T-D, Khreishah A, Noh W (2020) Alleviating item-side cold-start problems in recommender systems using weak supervision. *IEEE Access* 8:167747–167756
23. Hamid RA, Albahri A, Alwan JK, Al-qaysi Z, Albahri O, Zaidan A, Al-noor A, Alamoodi A, Zaidan B (2021) How smart is e-tourism? a systematic review of smart tourism recommendation system applying data management. *Comput Sci Rev* 39:100337
24. Inan E, Tekbacak F, Ozturk C (2018) Moreopt: a goal programming based movie recommender system. *J Comput Sci* 28:43–50
25. Mudda S, Zignani M, Gaito S, Giordano S, Rossi GP (2019) Timely and personalized services using mobile cellular data. *Online Soc Netw Media* 13:100048
26. Almaghrabi M, Chetty G (2020) Multilingual sentiment recommendation system based on multilayer convolutional neural networks (MCNN) and collaborative filtering based multistage deep neural network models (CFMDNN). In: 2020 IEEE/ACS 17th international conference on computer systems and applications (AICCSA), pp 1–6
27. Yan C, Xian J, Wan Y, Wang P (2021) Modeling implicit feedback based on bandit learning for recommendation. *Neurocomputing* 447:244–256
28. Hu Y, Xiong F, Lu D, Wang X, Xiong X, Chen H (2020) Movie collaborative filtering with multiplex implicit feedbacks. *Neurocomputing* 398:485–494
29. Tiwari S, Saini A, Paliwal V, Singh A, Gupta R, Mattoo R (2020) Implicit preferences discovery for biography recommender system using twitter. *Procedia Comput Sci* 167:1411–1420
30. Sun X, Meng L, Liang J, Li S (2019) Hybrid excitation synchronous motor feedback linearization decoupling sliding mode control. In: 2019 22nd international conference on electrical machines and systems (ICEMS), pp 1–5
31. Palomares I, Porcel C, Pizzato L, Guy I, Herrera-Viedma E (2021) Reciprocal recommender systems: analysis of state-of-art literature, challenges and opportunities towards social recommendation. *Inf Fusion* 69:103–127
32. Ignatov DI, Nikolenko SI, Abaev T, Poelmans J (2016) Online recommender system for radio station hosting based on information fusion and adaptive tagaware profiling. *Expert Syst Appl* 55:546–558
33. Alhijawi B, Kilani Y (2020) A collaborative filtering recommender system using genetic algorithm. *Inf Process Manag* 57(6):102310
34. Tewari AS (2020) Generating items recommendations by fusing content and user-item based collaborative filtering. *Procedia Comput Sci* 167:1934–1940
35. Rasheed F, Wahid A (2021) Learning style detection in e-learning systems using machine learning techniques. *Expert Syst Appl* 174:114774

36. Chen R, Hua Q, Chang Y-S, Wang B, Zhang L, Kong X (2018) A survey of collaborative filtering-based recommender systems: From traditional methods to hybrid methods based on social networks. *IEEE Access* 6:64301–64320
37. Aljunid MF, Dh M (2020) An efficient deep learning approach for collaborative filtering recommender system. *Procedia Comput Sci* 171:829–836
38. Chen J, Fang J, Liu W, Tang T, Yang C (2020) CLMF: a fine-grained and portable alternating least squares algorithm for parallel matrix factorization. *Future Gener Comput Syst* 108:1192–1205
39. Banihashemi S, Li J, Abhari A (2019) Scalable machine learning algorithms for a twitter followee recommender system. In: 2019 spring simulation conference (SpringSim), pp 1–8

Review on Application of Wireless Technology Using IoT



Deekshitha S. Nayak, N. Akshaya Krishna, Sahana Shetty, Sukanya D. Naik, V. Sambhram, and Krishang Shetty

Abstract Wireless technology is one of the means of communication that sends data that has been multiplexed across a large frequency range. Wireless networking is more preferable and also convenient than wired networking because it saves time and removes a variety of cable-related barriers. Internet of Things (IoT) system consists of a network of smart devices linked to a cloud platform. The Internet of Things (IoT) is a promising platform for taking existing technology to the next level. The different IoT applications clearly demonstrate how important technology is in everyday life. This paper reviews about the application of wireless technology used in the Internet of Things.

Keywords Internet of Things · Structural damage detection · Structural health monitoring · Wireless sensor networks · Detection of bridge damage

1 Introduction

Internet of Things (IoT) has infiltrated numerous factors of human life in recent years, including cities, residences, universities, industrial facilities, organisations, agricultural settings, hospitals, and medical centres. Several characteristics, such as data generation/consumption and online services, improve everyday life and activities in IoT environments around the world. In the IoT ecosystem, a range of applications run to provide facilities and smart services. Innovative apps for monitoring, managing, and automating human activities are being developed as user needs grow [1]. Cloud computing is also used in IoT applications to generate appropriate mix of services for service-based applications in IoT environments by combining existing

D. S. Nayak (✉) · S. Shetty · S. D. Naik

Department of Electronics and Communication Engineering, Mangalore Institute of Technology & Engineering, Moodabidri, Mangaluru, Karnataka, India
e-mail: deekshitha.nayak06@gmail.com

N. Akshaya Krishna · V. Sambhram · K. Shetty

Department of Civil Engineering, Mangalore Institute of Technology & Engineering, Moodabidri, Mangaluru, Karnataka, India

atomic services. IoT scenarios are used in smart device applications, and consumers apply them to their daily activities in various locations [2]. Smart cities include automated transportation, water distribution, energy management, urban security, pollution monitoring, and other great features in the real time IoT. Smart City is one of the powerful IoT applications that addresses many of the city's primary issues. IoT applications also offer the advantage of allowing users to choose the best option for them, regardless of whether they decide, manage, or monitor cloud resources for the environment. Despite the fact that application domains have diverse objectives, they all have the same aims. It's all about providing smart services to improve the quality of people's lives. Meeting quality of service (QoS) criteria is the key focus of IoT applications. The Internet of Things (IoT) refers to physical objects (or groups of such objects) as shown in Fig. 1 that are equipped with sensors, processing power, software, and other technologies, and that connect to and exchange data with other devices and systems over the Internet or other communication networks. This paper mainly presents the application of internet of things in construction industry where maintenance plays a major role in monitoring the entire architecture of the structure in order to avoid future destruction. IoT application smart services must fulfil user requirements in terms of security, cost, service time, energy usage, dependability, and availability [3]. Wireless technology is one of the means of communication that sends data that has been multiplexed across a large frequency range. Wired networking is inconvenient compared to wireless networking because wireless networking saves time and removes a number of obstacles that are cable-related [4]. IoT system is made up of sensors and devices that communicate with the cloud via some form of connectivity. Once the data reaches the cloud, software analyses it and may decide to take action, such as issuing an alert or automatically altering the sensors/devices without the user's involvement.

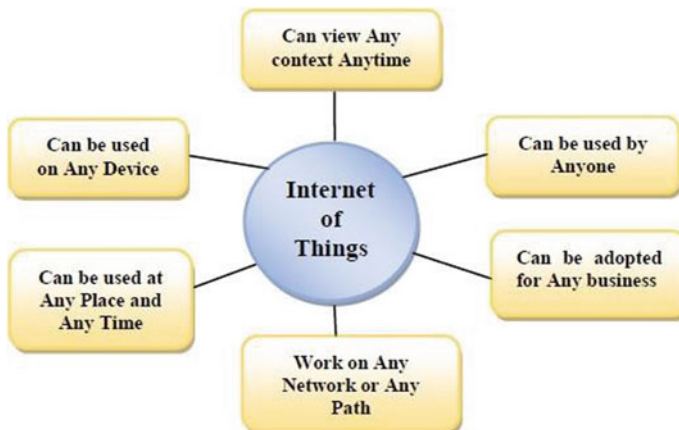


Fig. 1 Consents of IoT

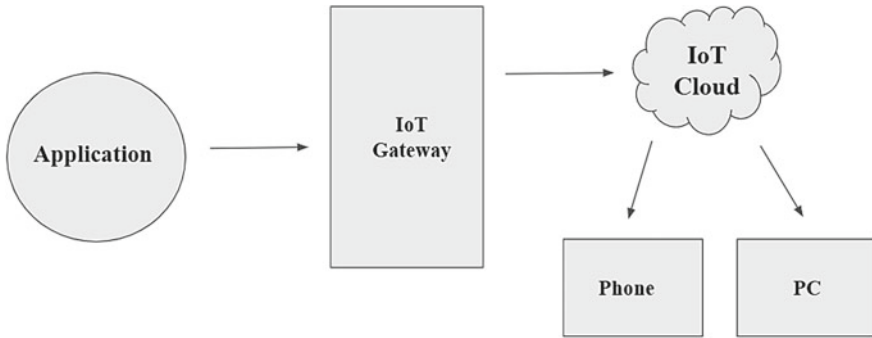


Fig. 2 Architecture of Internet of Things

IoT is currently popular when more businesses invest in the development of various IoT solutions, the technology's potential to improve corporate operations and save costs grows. Many smart devices communicate with one another through the Internet, requiring little human intervention. In practise, the IoT system is made up of a number of smart devices as in Fig. 2 (sensors, controllers, and other devices) as well as the cloud platform to which they are all linked. The mechanism works on a number of levels:

- Physical devices are referred to as hardware.
- The cloud infrastructure is in charge of storing and processing the data that is received.
- A user-facing management software for controlling IoT devices (for a smartphone, tablet, or PC).

The cost of each infrastructure component is determined by the sort of solution are developing and the degree of its complexity [5, 9]. Cloud computing is a software architecture based on applications that stores data on remote servers accessible via the internet. To access data stored in the cloud, a user can utilise an internet browser or cloud computing software. On the other hand, because it is responsible for securely storing data and information, is the most critical aspect of cloud computing. It includes servers, computers, databases, and central servers. One of the main goals of IoT device management is to monitor the IoT devices remotely. It mainly involves viewing the status of your devices and other network activities, gain important insights into data points that the user has defined. Measurements such as the intensity of the vibration can be included in these data points and lastly create user-defined event notifications to aid important choices and preventative maintenance.

2 Application of Wireless Technology

Currently because of rapid technological advancements, the modern world is becoming shorter and faster. Due to developments in wireless technologies, the globe has experienced amazing changes in wireless communication during the last few decades. Wireless technologies have been widely adopted and allow people to communicate without the use of cables [6]. Mobile and wireless technology has advanced significantly, resulting in wireless gadgets that are both convenient and economical. Wireless networking is especially useful in situations when physical media installation is not possible and on-the-spot access to information is required [7]. Access to both speech and data is feasible because to wireless networking. Advanced wireless communication technologies are employed in a variety of personal and business voice and data service applications [8]. Based on the intensity of the vibration which are detected from the sensors damage can be categorised into three types minor damage, moderate damage and sever damage. Based on these three classifications the sort of damage can be identified and can be rectified accordingly.

2.1 Structural Health Monitoring

The method of analysing the health of the structure in service using an automated monitoring system is known as structural health monitoring (SHM), and it's an important part of low cost maintenance based on conditions strategies [9]. SHM system must include a sensing network, data processing and analysis, damage assessment, and decision-making as shown in the Fig. 3. The potential for SHM system to provide major economic and personal safety considerations gains is enormous. SHM technology is being applied to real-world civil engineering structures, on the other hand, is still in its early phases, and because of its multidisciplinary nature, it demands breakthroughs in a range of areas [10]. To make certain that infrastructure management profit from this developing technology, extensive additional work is required.

Structural health monitoring is mainly a method of utilising a damage assessment and evaluation of health tool for engineering structures. SHM makes use of sensing devices and related hardware and software for structural performance monitoring and working conditions of engineered structures [11]. SHM entails monitoring a structure that evolves over a period of time on a regular basis responses from the structure and operating measurements of the environment from the sensor network, followed by an assessment of the structure's current status and future performance [12]. The output of this procedure for long-term SHM is frequently updated information about the structure's ability to fulfil the purpose it was created for, taking into account the unavoidable degradation and aging that occurs in working environments. Furthermore, SHM is used for quick condition evaluation in order to offer timely and accurate information about the structural integrity following catastrophic occurrences such as

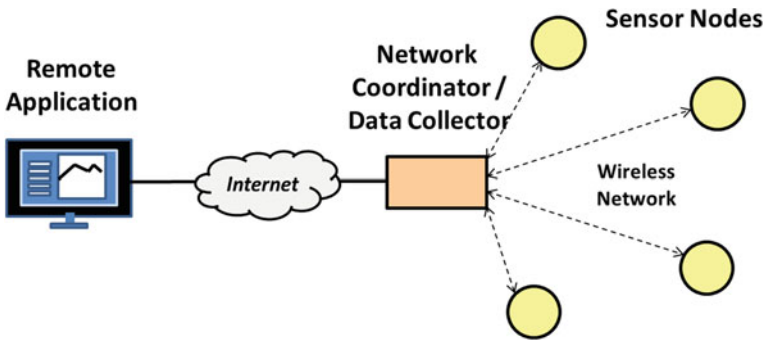


Fig. 3 Structural health monitoring

blast or an earthquake [13]. SHM intends to use monitored data to discover structural problems and assess the structure's health. In the Structural health monitoring system (SHM), other indices or measures based on SHM data include frequency change, flexibility, curvature, and displacement-based indices are all examples of displacement-based indices. Various tests and experimental measurements can be used to keep track of one's health on a constant basis. To talk about performance, you need to first identify objective and quantitative criteria for a limit-event. During the lifecycle of a structure, the desired performance is obtained by avoiding exceeding the limit states. The most typical limit-states that can be evaluated are a structure's or a structural component's safety, serviceability, and durability.

2.2 Structural Damage Detection

Maintenance should be prioritized in the construction industry because these structures are exposed to a variety of environmental variables, making maintenance and damage detection essential. The detection, location, and analysis of a structure structural deterioration is known as structural damage detection [14]. Each structure is distinct in terms of material, shape, and behaviour, which might change through time as a result of age, use, or environmental conditions. Bridges, dams, nuclear power plants, and energy utilities, for example need to have constant monitoring and timely maintenance [15]. Structural health monitoring is a new approach that is being used around the world, particularly in buildings that are in environments that are hostile [16]. The data from the structure is collected using sensors so that the deterioration can be seen and propose a solution. As the complexity of the structures grows, an approach that is both efficient and cost-effective is necessary [17]. In the Structural Damage Detection (SDD) as shown in Fig. 4, there is an approach for structural damage detection approach based on transmissibility in the time domain. The method updates finite element models based on the difference in transmissibility of structure response in the time domain before and after damage. By minimising

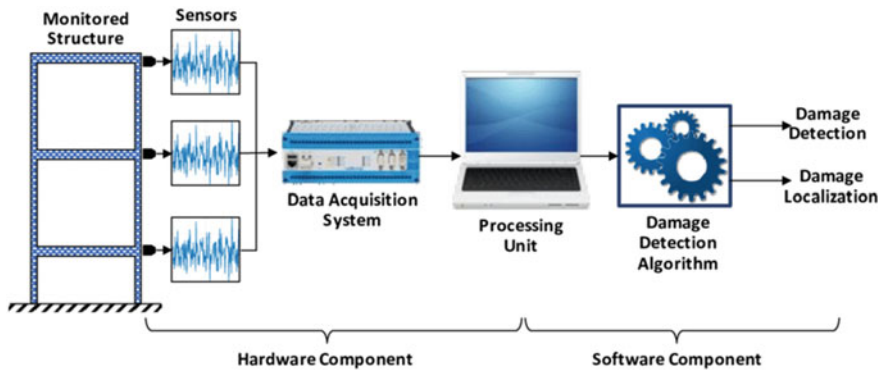


Fig. 4 Structural damage detection

the discrepancy between the measurements at gauge locations and the reconstruction response extrapolated by the finite element model, the damage is detected and quantified through iteration.

The following five levels describe the goals of a structural damage detection technique.

- Damage detection, which gives a qualitative indicator that the structure may be damaged.
- Damage localization, which tells you where the damage is most likely to be.
- Damage categorization gives information on the sort of damage.
- Damage appraisal provides a rough estimate of the damage.
- Damage prognostication provides the data on the structure's safety, such as estimated remaining usable life [18].

Computers are used sparingly in the building industry. Finding a technique to assess the durability and safety of concrete constructions is tough [19]. In order to examine the real-time behaviour of physical structures, IoT becomes important to display residents' consents and possessions that could be linked at any time, at any place, with any person, employing some network plus some internet utility.

2.3 Detection of Bridge Damage

Bridges are critical components of any surface transportation system. Damage to a major bridge may result in long-term economic loss in addition to the expense of repair or replacement due to partial or entire closure of the route. The aftermath of a major earthquake, the survival of bridges is also critical in order to assist rescue operations [20]. As a result, it has become standard practise to conduct a severe examination of bridge integrity and safety at regular period of time and promptly following catastrophic events like earthquakes.

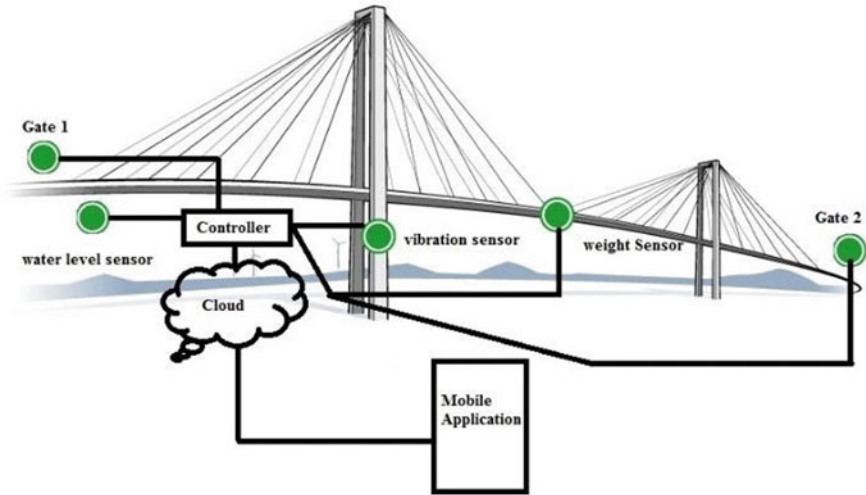


Fig. 5 Detection of bridge damage

Detection of structural degradation in bridges is a study issue that has gotten a lot of attention in recent years as in Fig. 5. An outdated infrastructure on the rail and road that is susceptible to traffic loads that far exceed the original design criteria, is the fundamental cause for its rise in popularity. This enormous rise in weight and loading causes structural deterioration, which shortens the working life of the structure [21]. Furthermore, as bridge infrastructure ages and deteriorates, the frequency of inspections must rise to compensate for the decrease in structural safety.

Bridge condition deteriorates over time because of variety of the degradation mechanisms, including corrosion, creep, and also cyclic loads. Vibration-based bridge damage detection systems have traditionally centred on keeping track of differences in modal parameters [22]. Because of their changes in sensitivity to different operating and environmental circumstances, these techniques are frequently mistaken for structural damage [23]. More improved computational tools have emerged in recent study, allowing not just for the evaluation of data that is noisier and more difficult, but also for research to move away from simply monitoring changes in modal parameters [24]. Sensors mounted on various areas of the bridge monitor vibration, traffic, vehicle weight, water level monitoring etc. If any of these metrics When their threshold value is exceeded, the communication system sends a message to the management centre, allowing preventative steps to be taken [22]. All of the bridge's parameters are collected by the processor and transfers them to another module that is only a short distance away. The receiver module receives the parameters from the transmitter and sends a message to a database centre with all of the parameters. In the existing work the damage at different locations of the bridge was detected however the severities of the damage could not be determined. The sensors play a vital role in the damage detection. The accuracy is lacking in the old analogue sensors. From

the paper damage detection of bridge using wireless sensors, classification of the damage is mandatory when it comes to service life of a structure[25, 26]. Damage at different locations of the bridge was detected however the severities of the damage could not be determined in this system. From the paper structural health monitoring: An IoT Sensor System for Structural Damage Indicator Evaluation, the technique was used to determine the extent of damage in an aluminium bar held in a bench vice. The system would not be cable enough to be installed in the case of real structures, such as buildings. From the paper development in vibration based structural damage detection technique, sensors play a vital role in the damage detection. With the use of digital accelerometers, the drawbacks of old analogue sensors utilised in ordinary monitoring systems can be eliminated. Thus the system could have been made more cost efficient by making use of such accelerometers.

3 Conclusion

Wireless technologies enable low-cost access to digital information and connect a variety of computing and telecommunications equipment without the need to buy, transport, or connect wires. In the future, it will be the primary technology for exchanging digital information. IoT is an abbreviation for Internet of Things; it is a network of interconnected devices that can be accessed via the Internet. Wireless sensor networks (WSN) are regarded as the IoT's eyes and ears. WSN serves as a bridge that connects the real and digital worlds. IoT is regarded as the primary evolution of the Internet; use IoT in a variety of applications that have an impact on daily lives.

The Internet of Things (IoT) is a promising platform for advancing existing technology. The various IoT applications are reviewed with the importance of technology in daily lives. Because of the integration of intelligence into the electrical things that surround us, this evolving networking prototype will have an impact on every aspect of the lives, from automated homes to smart health and environmental monitoring. Field structural behaviour is better understood in Structural Health Monitoring. Damages should be detected as soon as an issue arises. By Structural Health Monitoring inspection and repair times are cut in half. It also assists in the development of sound management and maintenance practises. Structural Damage Detection on the other hand mainly detects the damage in the first place and allows the users to take effective measures before it's too late. By making use IoT this can be done efficiently without any complications. It is feasible to discover potential failure mechanisms by continuously monitoring bridge motions, vibrations, and structural changes.

Conflicts of Interest The authors declare no conflict of interest.

References

1. Scuro C, Sciammarella PF, Lamonaca F, Olivito RS, Carni DL (2018) IoT for structural health monitoring. *IEEE Instrum Meas Maga* 21(6):4–14
2. Gubbi J, Buyya R, Marusic S, Palaniswami M (2013) Internet of things (IoT): a vision, architectural elements, and future directions. *Future Gener Comput Syst* 29(7):1645–1660
3. Serov A (2017) Cognitive sensor technology for structural health monitoring. *Procedia Struct Integrity* 5:1160–1167
4. Avci O, Abdeljaber O, Kiranyaz S, Hussein M, Inman DJ (2018) Wireless and real-time structural damage detection: a novel decentralized method for wireless sensor networks. *J Sound Vibr* 424:158–172
5. Arcadius T, Gao B (2017) Structural health monitoring framework based on IoT. *IEEE Internet Things J* 4(3):619–635
6. Chanv B, Mehta V (2017) Structural health monitoring using IoT and wireless technologies. In: *IEEE international conference on intelligent communication and computational techniques (ICCT)*, Manipal University Jaipur, 22–23 December 2017, pp 151–157
7. An H, Youn BD, Kim HS (2022) A methodology for sensor number and placement optimization for vibration-based damage detection of composite structures under model uncertainty. *IEEE Compos Struct* 279:12–21
8. Lamonaca F, Scuro C (2018) Synchronization of IoT layers for structural health monitoring. *IEEE*, pp 89–84
9. Abdelgawad A, Yelamarthi K (2018) Internet of Things platform for structural health monitoring. School of Engineering & Technology, Central Michigan University, ET100, Mount Pleasant, MI 48859, USA. *IEEE*
10. Myers A, Mahumud MA (2016) Toward integrating structural health monitoring with internet of things. School of Engineering & Technology Central Michigan University, ET100 Mount Pleasant, MI 48859, USA. *IEEE*, pp 438–441
11. Pallares FJ, Betti M, Bartoli G, Pallares L (2021) Structural health monitoring (SHM) and Nondestructive testing (NDT) of slender masonry structures. *Constr Build Mater* 297:123–138
12. Barontini A, Masciotta MG, Ramos LF, Amado-Mendes P, Lourenço PB (2017) An overview on nature-inspired optimization algorithms for structural health monitoring of historical buildings. *Procedia Eng* 199:3320–3325
13. Worden K, Cross EJ (2018) On switching response surface models, with applications to the structural health monitoring of bridges. *Mech Syst. Signal Processing* 98:139–156
14. Spencer BF, Bridge J (2014) Structural damage detection using smart sensors. In: *Proceedings of SPIE, sensors and smart structures technologies for civil, mechanical, and aerospace systems*, vol 69. *IEEE*, pp 2–18
15. Yana YJ, Chengb L, Wua ZY, Yam LH (2007) Development in vibration-based structural damage detection technique. *Mech Syst Signal Process* 21:2198–2211
16. Doebling SW, Farrar CR, Prime MB, Shevitz DW (2019) Damaged identification and health monitoring of the structural and mechanical systems from changes in their vibration characteristics, vol 27. *IEEE*
17. Kiranyaz S, Hussein M, Gabbouj M, Inman DJ (2004) A review of vibration-based damage detection in civil structures: from traditional methods to machine learning and deep learning applications. *Mech Syst Signal Process*, 12–28
18. Day-Lewis FD, Slater LD, Robinson J, Johnson CD, Terry N, Werkema D (2017) An overview of geophysical technologies appropriate for characterization and monitoring at fractured-rock sites. *J. Environ Manag.* 204:709–720
19. Carni DL, Scuro C, Lamonaca F, Olivito RS, Grimaldi D (2017) Damage analysis of concrete structures by means of b-Value Technique. *Int J Comput* 16(2):82–88
20. Roy K, Ogai H, Bhattacharya B, Ray-Chaudhuri S, Qin J (2016) Damage detection of bridge using wireless sensors. *Int Feder*, 23–30
21. Dhivya A, Hemalatha M (2013) Structural health monitoring system-an embedded system approach. *Int J Eng Technol (IJET)* 5:273–281

22. Gao ZF, Du YL, Su MB, Chen B (2006) Network sensor and its application in structure health monitoring system. In: Proceedings of the 1st international conference on innovative computing, information and control, pp 68–71
23. Mutillo M, Stornelli V, Alaggio R, Paolucci R, Di Battista L, de Rubeis T, Ferri G (2007) Structural health monitoring: an IoT sensor system for structural damage indicator evaluation. Department of Industrial and Information Engineering and Economics (DIIIE), pp 1–15
24. Alavi AH, Hasni H, Lajnef N, Chatti K (2016) An intelligent structural damage detection approach based on self-powered wireless sensor data. *IJSR* 4(12). ISSN:2319–7094
25. Karuppusamy P (2018) A sensor based IoT monitoring system for electrical devices using blynk framework. *J Electron Inf* 2(3):182–187
26. Chen JIZ, Lai KL (2020) Internet of Things (IoT) authentication and access control by hybrid deep learning method-a study. *J Soft Comput Paradigm (JSCP)* 2(04):236–245

Comparative Study of Conditional Generative Models for ISL Generation



Marrivada Gopala Krishna Sai Charan, S. S. Poorna, K. Anuraj,
Choragudi Sai Praneeth, P. G. Sai Sumanth,
Chekka Venkata Sai Phaneendra Gupta, and Kota Srikar

Abstract Deep Generative Models are widely used to generate data that look similar to training data. Synthesis and sampling of data using Deep Generative Models can be useful for certain situations where generating data by hand would be expensive or time consuming. Data-sets for Indian Sign Language often are of small size, which hinders training of Deep Learning models to a good accuracy. In this project we attempt to compare various state-of-the-art Conditional Generative Models for Indian Sign Language Recognition task and evaluate them using various performance metrics.

Keywords VAE · CVAE · Indian Sign Language Generation · CGAN

1 Introduction

Convolutional Neural Networks(CNN) are widely used for recognition tasks. One major drawback of training CNNs is that they require $\sim 10^4$ – 10^6 images for training, which, in most cases is an expensive proposition. This situation further exacerbates for Indian Sign Language(ISL) datasets, for which datasets are of very small size. One major problem with small datasets is overfitting. This is a serious problem since, as the network memorizes the training data and it will perform poorly on unseen examples. Dataset Augmentation is one of the standard solutions against overfitting. Training classifier using noise, rotation etc. will make it immune to distortions which preserve semantics. But creating data by hand is often a cumbersome task. Several Conditional Generative models can be used to do image generation. Datasets for ISL Recognition

M. G. K. S. Charan · S. S. Poorna (✉) · K. Anuraj · C. S. Praneeth · P. G. S. Sumanth ·
C. V. S. P. Gupta · K. Srikar
Department of Electronics and Communication Engineering, Amrita Vishwa Vidyapeetham,
Amritapuri, India
e-mail: poornass@am.amrita.edu

are often of small size, which makes training Deep learning models difficult. In this work, we attempt to compare various Conditional Generative Adversarial Networks (GANs) for ISL Generation task.

2 Literature Review

In [1], the authors attempted to use a mixture of GANs rather than increasing depth or width of the model. They obtained probability density function of the composite Generator as weighted sum of probability density functions of different generators. They also made Discriminator output as a weighted sum of outputs of various discriminators. The weights w_i of the Generator and v_i of the Discriminator are the learnable parameters. They tested their models on Cifar-10 dataset which contained images of size 32×32 . Authors used Tensor Processing Units (TPUs) for training their models. They obtained a Fréchet Inception Distance (FID) score of 3.60 and Inception score of 10.21 for Cifar-10 dataset. Training a mixture of GANs on multi-modal data is computationally expensive and requires state of the art hardware. In order to fit multi-modal data, mixture containing more GANs is required and will be equal to the number of modes.

The authors in [2] proposed solutions to Auxiliary Classifier GAN (ACGAN) which suffers from exploding gradients. They used feature normalization with a layer before classification layer for feature extraction. Authors introduced Data-to-Data Cross Entropy loss function for training GANs. Various types of invertible and differentiable data augmentations were used to improve the performance of GANs. StyleGAN2 architecture was adopted along with Differentiable Augmentations for GAN model. The datasets: CUB200, Tiny Imagenet, Imagenet, Cifar-10 and AFHQ datasets were used for analysis and reported 13.9% improvement in accuracy compared to StyleGAN2-ADA. ReACGAN requires additional hyperparameter tuning for parameters of D2DCE(Data-to-Data Cross Entropy) loss function. Sometimes ACGAN generates well classifiable images without considering the fidelity and diversity of images and hence the images for which classifier gives highest classification accuracy will be poor.

Improvements to StyleGAN2 is proposed in [3]. An Adaptive Discriminator, which automatically augments data of the existing dataset was introduced in this paper. This prevents GAN from memorizing or overfitting the data. Discriminator never sees real data and is only evaluated on augmented images, the authors proposed certain conditions under which this approach works. They performed ROTATE2D, SCALE2D and TRANSLATE2D operations on Images. Instead of using deconvolutions, up-sampling with bilinear interpolation was used in generator. This architecture works for images of size $2^n \times 2^n$. The architecture is similar to StyleGAN2 except that here discriminator is trained on augmented images. FID score of 2.42 on Cifar-10 dataset was obtained. The downside is when data is multi-modal results may not be good.

In [4], the authors attempted to improve diversity of samples generated by GAN using Variational Auto Encoder (VAE). The authors added a balanced pre-training stage to GAN components. They used Conditional Variational Auto-encoder in pre-training stage for GAN initialization. They proposed a new objective function to capture the true distribution and integrated this new objective function along with pre-training stage to enhance training stability. Sum of KL divergence, MSE and Cross Entropy loss was used as the new objective function. The model was trained on MNIST, Fashion-MNIST and Cifar-10. The authors were able to get good diversity of samples even with data sets containing an imbalance of 50–100. However, FID score of generated samples was high and the inception score (IS) was less.

Conditional Variational Autoencoder for reconstruction of images was proposed in [4]. The learning algorithm learns a conditional sampler similar to a Random Number Generator (RNG). The latent variable used was similar to a random seed in a RNG and is generated using recurrent process modelled by a network. Two probabilistic encoders for encoding input output and forward map \mathcal{A} were used in this architecture. At each step recurrent unit reuses the observation data y and forward map for refinement. This overall process differs from a deterministic mapping. Deterministic mapping takes back-projected data and outputs a refinement, where as probabilistic mapping employs current sample and quality measures and then decides the refinement strategy. The authors were able to get reconstructed images similar to that of ground truth. However, compared to ground truth generated images are blurrier/softer. In [5], authors compared traditional machine learning methods with Deep Convolutional neural networks for melanoma detection. In [6], the authors compared various state-of-the-art Convolutional Neural Network architectures like ResNet-152, ResNet101, VGG16, VGG19, AlexNet for image classification problem. In [7], the discussed about various aspects of Convolutional Neural Networks like convolutional layer, various types of pooling, various types of activations, various popular frame works available for efficient implementation of Neural Networks. In [8], the authors used Microsoft kinect for Three-dimensional(3D) geometric processing for Indian Sign Language Recognition (Fig. 1).

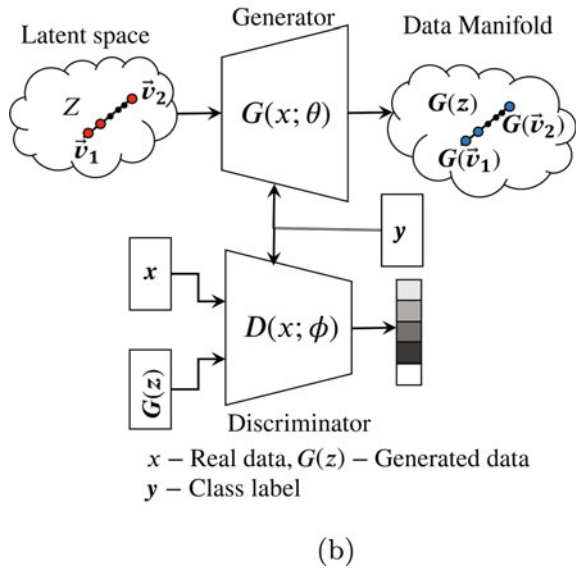
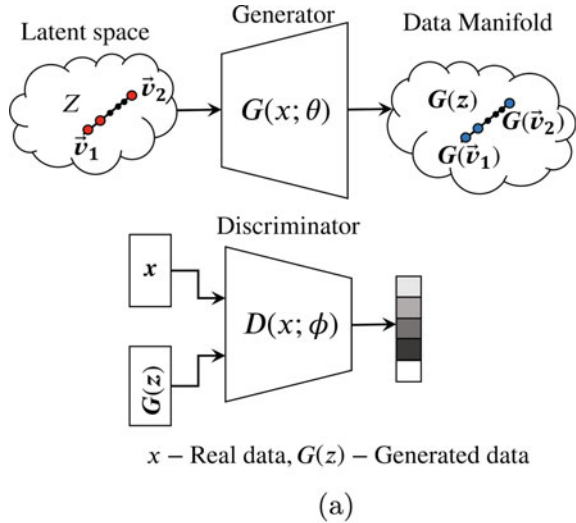
3 CGAN Preliminaries and ISL Dataset

3.1 Conditional GAN

The idea of GANs is to sample from a distribution of data. Real word data distribution \mathcal{W} is a very high dimensional manifold, hence sampling is an extremely complex task, instead we can sample from a simpler data distribution \mathcal{Z} for example $z \in \mathcal{N}(0, 1)$ and learn a transformation $G : z \rightarrow w, w \in \mathcal{W}$. This transformation is learned by the *Generator* net. It does so with the help of *Discriminator* net which receives both real and fake data and classifies them. If class label information is not provided for Generator and Discriminator, then such a type of GAN is called *Uncon-*

ditional GAN, if class information is provided then such a type of GAN is called *Conditional GAN*. Types of GANs is shown in Fig. 1 a and b. Conditional GANs use one of the 2 approaches for injecting of class label information into Generator and Discriminator. *Projection Based* in which class label information is embedded using an inner product, and *Auxiliary Classifier* based in which Discriminator is made to predict class rather than real/fake. We can train the discriminator more number of times for giving better feedback to the generator,

Fig. 1 a Unconditional GAN, b Conditional GAN



3.2 ISL Dataset

ISL(Indian Sign Language) dataset is used in this work¹ for the link to the dataset). This dataset contains $\approx 42,000$ images of resolution 128×128 . It contains Indian Sign Language signs of Alphabets(26),and Numerals(1–9).

4 Training Configuration of Models

In this work, various Conditional GAN models which are shown in Table 2 are trained and compared using ISL dataset. For training of models ADAM optimizer is used. Learning rate of Generator g_l and Discriminator d_l are set to be $5e-5$ and $2e-4$ respectively. The parameters:First order moment β_1 is 0 and the second order moment β_2 is 0.999. All the models are trained for 50k epochs. Parameters which are used for ADAM optimizer is summarized in the Table 1.

For better convergence, exponential moving average is used in all generators, except ACGAN since it suffers from exploding gradients problem. If w_t are the weights of the generator at t^{th} iteration and w_{t-1} are the weights of the generator at $t - 1^{th}$ iteration, Exponential Moving Average (EMA) update is applied as in Eq.(1). Here β is the smoothing factor.

$$w_t = \beta(w_t - w_{t-1}) + w_{t-1} \quad (1)$$

For performing frequency analysis average of Fast Fourier Transform (FFT) of real and fake images are calculated and the resultant FFT is shifted to origin. For training the models, mixed precision data types are used. For comparison of CGANs, FID and IS are used as metrics. The following approaches are adopted for injecting label information in Generator and Discriminator

- (i) Conditional Batch Normalization (cBN) is used in all the Generators except StyleGAN2-ADA, StyleGAN2-DiffAug, StyleGAN3.
- (ii) Conditional Adaptive Instance Normalization (cAdaIN) is used in Generators of StyleGAN2-ADA, StyleGAN2-DiffAug, StyleGAN3.
- (iii) StyleGAN Projection Discriminator (SPD) is used in Discriminators of StyleGAN2-ADA, StyleGAN2-DiffAug, StyleGAN3.
- (iv) Projection Discriminator (PD) is used for Discriminator of BigGAN

Table 1 Training parameters of ADAM optimizer

Batch size	g_l	d_l	β_1	β_2	n_{iter}
64	$5e-5$	$2e-4$	0	0.999	50k

¹ The dataset which used in this work can be found [here](#)

Table 2 Model parameters of various GANs

Model	GI ^a	DI ^b	Loss	EMA ^c
BigGAN	cBN	PD	Hinge	✓
ACGAN	cBN	AC	Hinge	✗
StyleGAN2- <i>DiffAug</i>	cAdaIN	SPD	Logistic	✓
StyleGAN2- <i>ADA</i>	cAdaIN	SPD	Logistic	✓
ContraGAN	cBN	2C	Hinge	✓
ReACGAN	cBN	D2DCE	Hinge	✓
StyleGAN3	cAdaIN	SPD	Logistic	✓
MHGAN	cBN	MH ^d	MH	✓

^{a b} GI/DI represent how class label information is injected in Generator and Discriminator

^c EMA-Exponential Moving Average(for Generator only)

^d MH- Multi-Hinge loss

- (v) Auxiliary Classifier (AC) is used for Discriminator of ACGAN(Auxiliary Classifier GAN)
- (vi) Conditional Contrastive (2C) loss is used for Discriminator of ContraGAN
- (vii) Data-to-Data Cross Entropy (D2DCE) loss is used for Discriminator of ReACGAN(Rebooting Auxiliary Classifier GAN)
- (viii) Multi-Hinge (MH) loss is used for Discriminator of MHGAN(Multi-Hinge GAN).

BigGAN architecture is used as background architecture for all the models except StyleGANs(StyleGAN2-*ADA*, StyleGAN2-*DiffAug*, StyleGAN3). Both StyleGAN2-*ADA* and StyleGAN2-*DiffAug* use same backbone architecture. StyleGAN2-*ADA* is trained on augmented images which means the network never sees real data, where as, StyleGAN2-*DiffAug* is trained on images+ transformed images. Augmentations applied to StyleGAN2-*DiffAug* are differentiable for back-propagating gradients. Linear interpolation of vectors is carried out for all GANs, Linear interpolation of 2 latent vectors \vec{v}_1, \vec{v}_2 is given by Eq.(2).

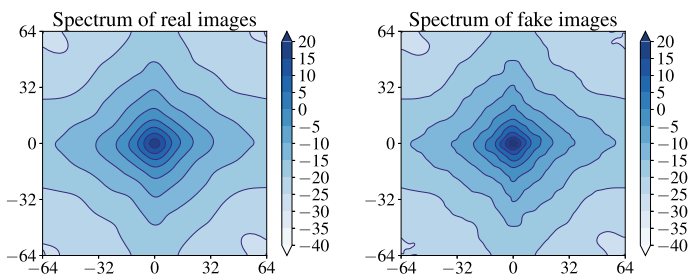
$$\vec{v} = \alpha \vec{v}_1 + (1 - \alpha) \vec{v}_2 \quad (2)$$

where, \vec{v}_1 and \vec{v}_2 are two latent vectors corresponding to different signs Parameters of Conditional GANs used in this work is shown in Table 2. All the GANs are trained on ISL (See footnote 1) dataset

In Table 3, ResBlock up block is residual block with Upsampling, ResBlock down is residual block with downsampling, ResBlock(without up or down) is residual block with identity connections without Up/Down sampling. ch is channel width multiplier.

Table 3 (left) Generator architecture, (right) Discriminator architecture of BigGAN for 128×128 images

$z \in \mathbb{R}^{120} \sim \mathcal{N}(0, 1)$	RGB image $x \in \mathbb{R}^{128 \times 128 \times 3}$
Embed(y) $\in \mathbb{R}^{128}$	ResBlock down $ch \rightarrow 2ch$
Linear($20+128$) $\rightarrow 4 \times 4 \times 16ch$	Non-Local Block (64×64)
ResBlock up $16ch \rightarrow 16ch$	ResBlock down $2ch \rightarrow 4ch$
ResBlock up $16ch \rightarrow 8ch$	ResBlock down $4ch \rightarrow 8ch$
ResBlock up $8ch \rightarrow 4ch$	ResBlock down $8ch \rightarrow 16ch$
ResBlock up $4ch \rightarrow 2ch$	ResBlock down $16ch \rightarrow 16ch$
Non-Local Block (64×64)	ResBlock $16ch \rightarrow 16ch$
ResBlock up $2ch \rightarrow 2ch$	ReLU, Global sum pooling
BN, ReLU, 3×3 Conv $ch \rightarrow 3$	Embed(y) $\cdot h + (\text{linear} \rightarrow 1)$
Tanh	

**Fig. 2** FFT Spectrum of real and fake images for ContraGAN

5 Results of Conditional GANs on ISL Dataset

5.1 Results of ContraGAN on ISL Dataset

The same architecture in [9] for IMAGENET dataset is used in this work. ContraGAN uses BigGAN as backbone architecture. After finishing training of our model, we performed Spectral analysis on real and fake images. Results of Frequency analysis for ContraGAN is shown in Fig. 2.

. The results of latent interpolation with z -axis fixed for ContraGAN is shown in Fig. 3. 8-Nearest Neighbour (8NN) Analysis is performed with the generated images. The results of 8NN analysis for ContraGAN is shown in Fig. 4. The images shown with a dashed blue line are generated images which are nearest to the images in the dataset (Nearest Neighbours).

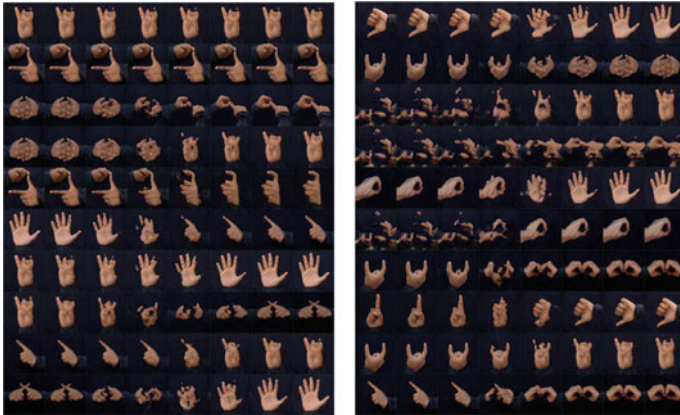


Fig. 3 Images obtained by linear interpolation of latent vectors with z -dim fixed for ContraGAN

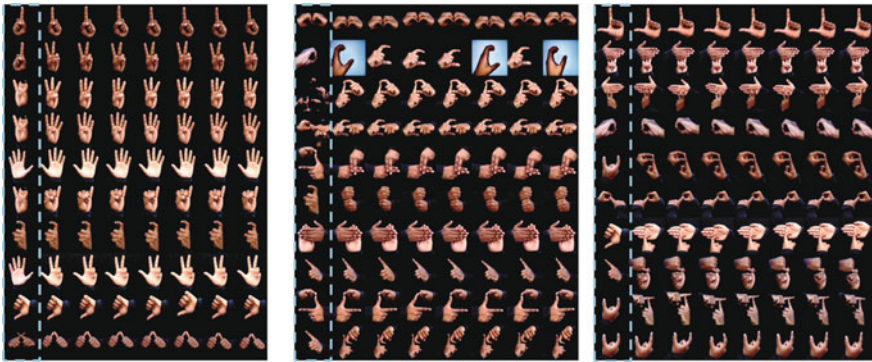


Fig. 4 (left) 8-NN Analysis on 10 Classes, (middle) 8-NN Analysis on 20 classes (right) 8-NN Analysis on 30 Classes for ContraGAN

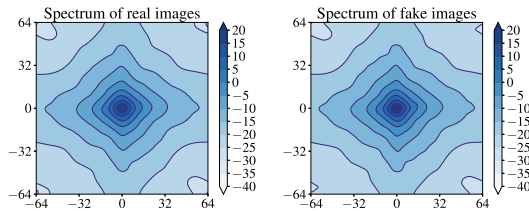


Fig. 5 FFT Spectrum of real and fake images for ACGAN

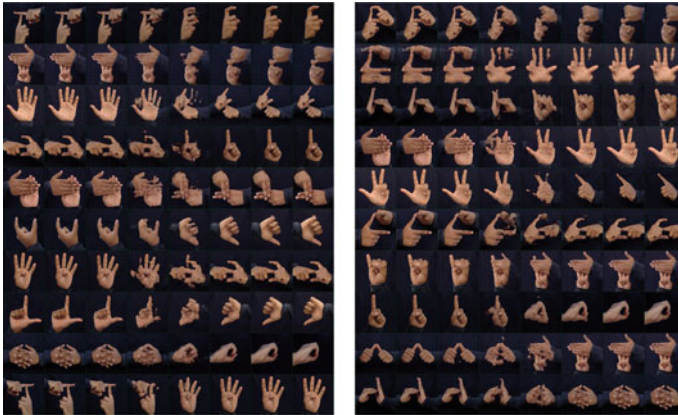


Fig. 6 Images obtained by linear interpolation of latent vectors with z -dim fixed for ACGAN

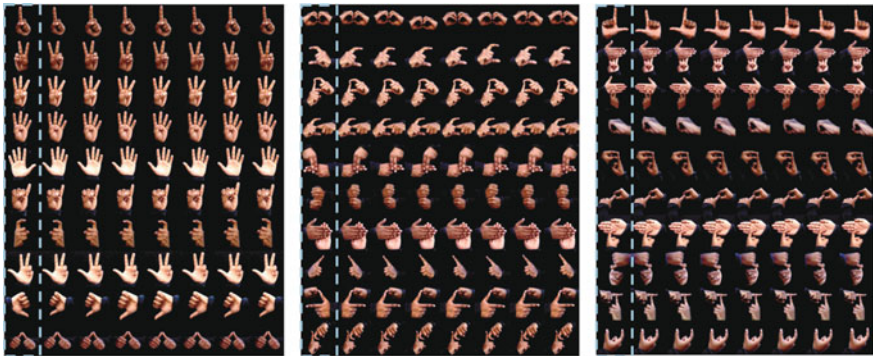


Fig. 7 (left) 8-NN Analysis on 10 Classes, (middle) 8-NN Analysis on 20 classes (right) 8-NN Analysis on 30 Classes for ACGAN

5.2 Results of ACGAN on ISL Dataset

Similar to ContraGAN discussed earlier, the same analysis is followed for other Conditional GANs as well. Results of Frequency Analysis for ACGAN is shown in Fig. 5. Images obtained by linear interpolation of latent vectors for ACGAN is shown in Fig. 6. Results of 8-Nearest Neighbour Analysis for ACGANs is shown in Fig. 7. For implementation of this ACGAN, BigGAN architecture [10] is used.

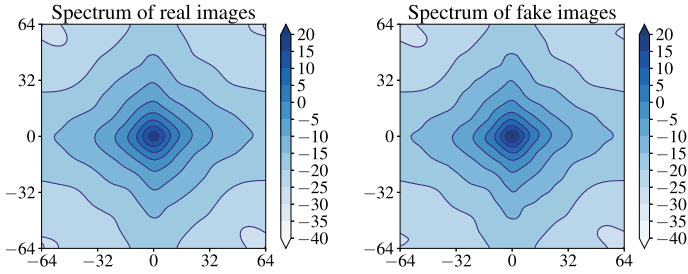


Fig. 8 FFT Spectrum of real and fake images for ReACGAN

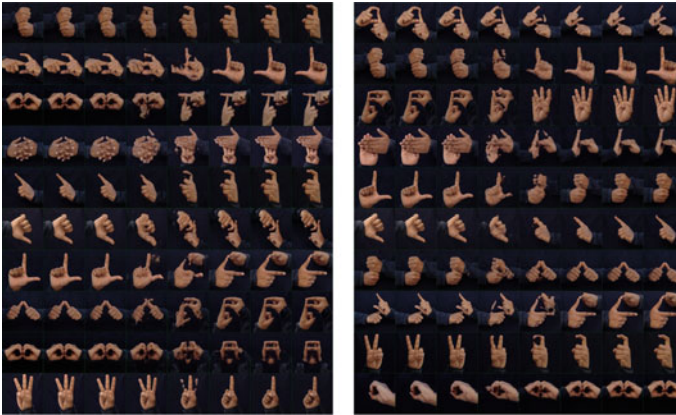


Fig. 9 Images obtained by linear interpolation of latent vectors with z -dim fixed for ReACGAN

5.3 Results of ReACGAN on ISL Dataset

Same architecture which is used for IMAGENET(128^2) dataset in [2] is used in this work. Results of Frequency analysis for ReACGAN is shown in Fig. 8.

Images obtained by Linear interpolation of latent vectors is shown in Fig. 9. Results of 8NN(Nearest Neighbour) analysis for ReACGAN is shown in Fig. 10.

5.4 Results of MHGAN on ISL Dataset

For implementation of this GAN, BigGAN architecture is used as Backbone architecture and Multi-Hinge loss for injecting class label information [11]. Results of Frequency analysis for real and fake images of MHGAN(Multi-Hinge GAN) is shown in Fig. 11.

Results of linear interpolation of latent vector for MHGAN is shown in Fig. 12.

Results of 8-NN(Nearest Neighbour) Analysis for MHGAN is shown in Fig. 13.

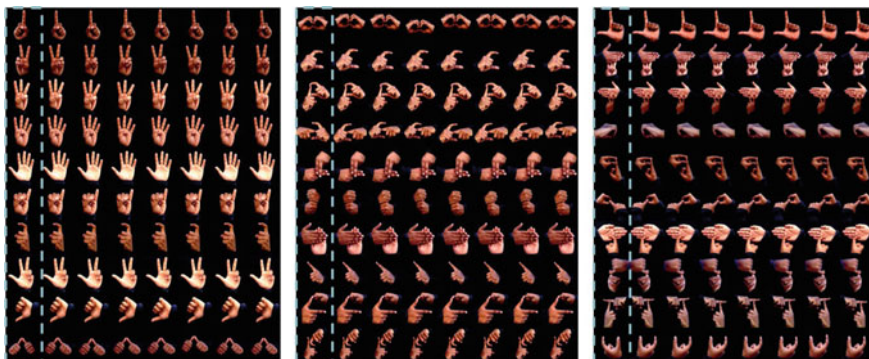


Fig. 10 (left) 8-NN Analysis on 10 Classes, (middle) 8-NN Analysis on 20 classes (right) 8-NN Analysis on 30 Classes for ReACGAN

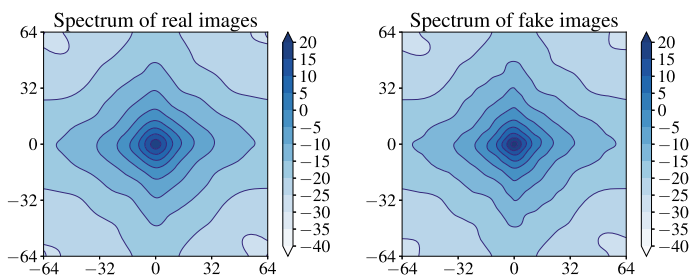


Fig. 11 FFT Spectrum of real and fake images for MHGAN

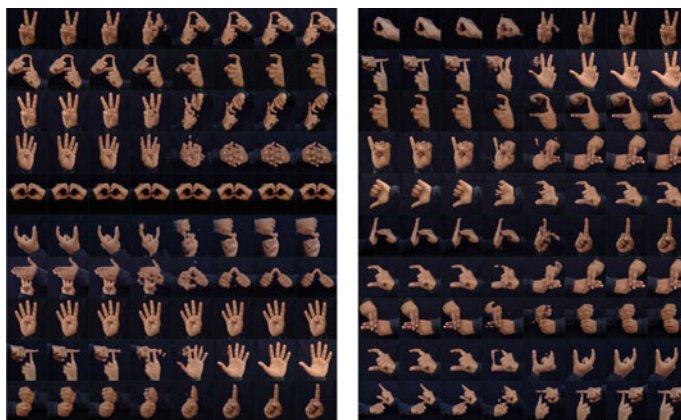


Fig. 12 Images obtained by linear interpolation of latent vectors with z -dim fixed for MHGAN

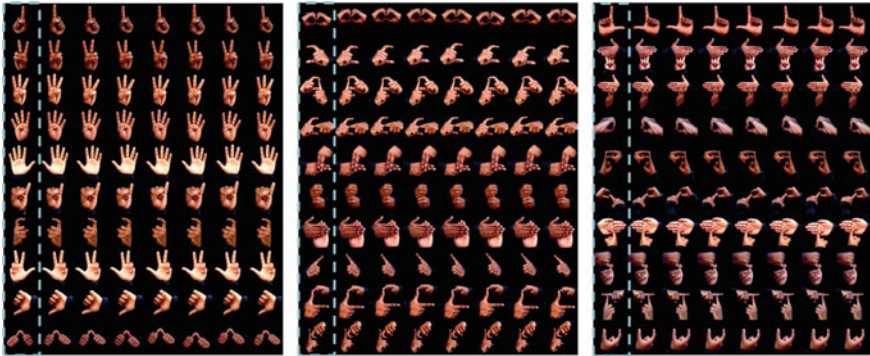


Fig. 13 (left) 8-NN Analysis on 10 Classes, (middle) 8-NN Analysis on 20 classes (right) 8-NN Analysis on 30 Classes for MHGAN

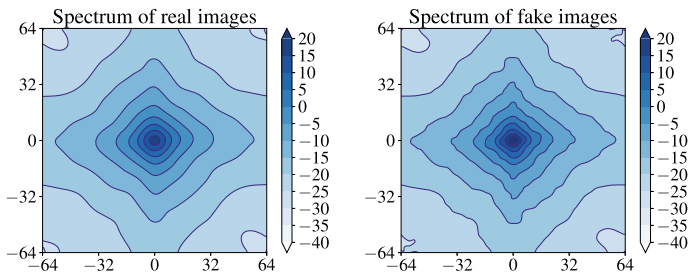


Fig. 14 FFT Spectrum of real and fake images for BigGAN

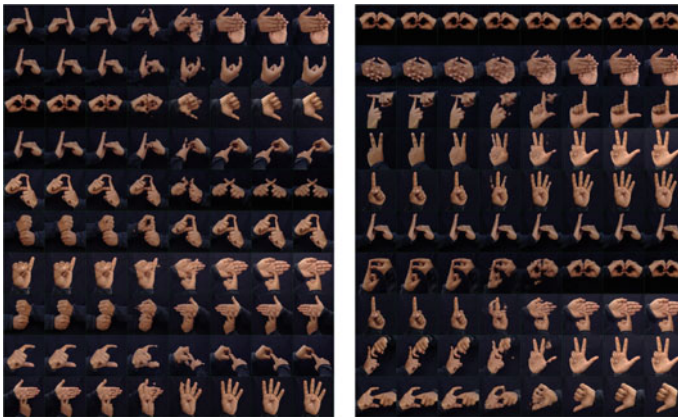


Fig. 15 Images obtained by linear interpolation of latent vectors with z -dim fixed for BigGAN

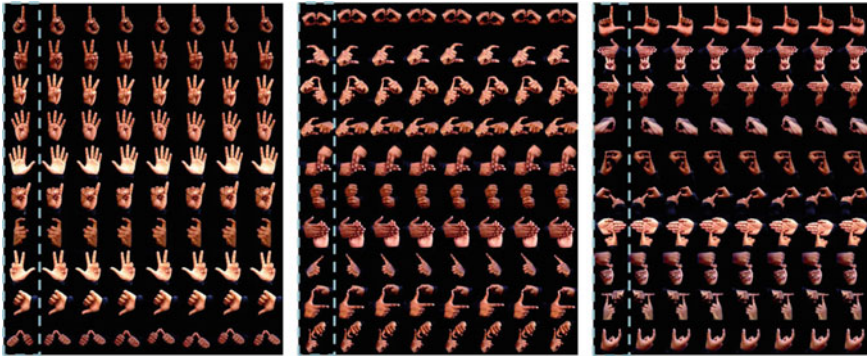
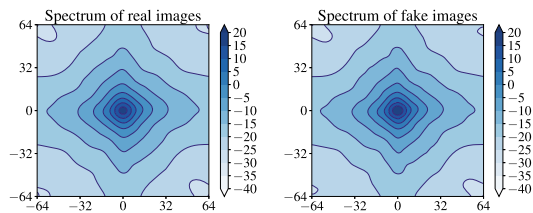


Fig. 16 (left) 8-NN Analysis on 10 Classes, (middle) 8-NN Analysis on 20 classes (right) 8-NN Analysis on 30 Classes for BigGAN

Fig. 17 FFT Spectrum of real and fake images for StyleGAN2-ADA



5.5 Results of BigGAN on ISL Dataset

GAN architecture which is used for IMAGENET(128²) in [10] is used in this work. BigGAN architecture used in [10] forms as backbone architecture for all the CGANs in this work except StyleGANs. Results of frequency analysis for BigGAN is shown in Fig. 14.

Results of linear interpolation of latent vector for BigGAN is shown in Fig. 15.
 Results of 8-NN(Nearest Neighbour) Analysis for BigGAN is shown in Fig. 16.

5.6 Results of StyleGAN2-ADA on ISL Dataset

In StyleGAN2-ADA [3], ROTATE2D,SCALE2D,TRANSLATE2D transformations are applied to the dataset. All these transformations are linear in nature and original image can be recovered from the transformation by applying an inverse transformation to the transformed image. StyleGAN2-ADA is trained only on these transformed images. Results of Frequency Analysis is shown in Fig. 17.

Results of linear interpolation of latent vector for StyleGAN2-ADA is shown in Fig. 18.
 Results of 8-NN analysis for StyleGAN2-ADA is shown in Fig. 19.

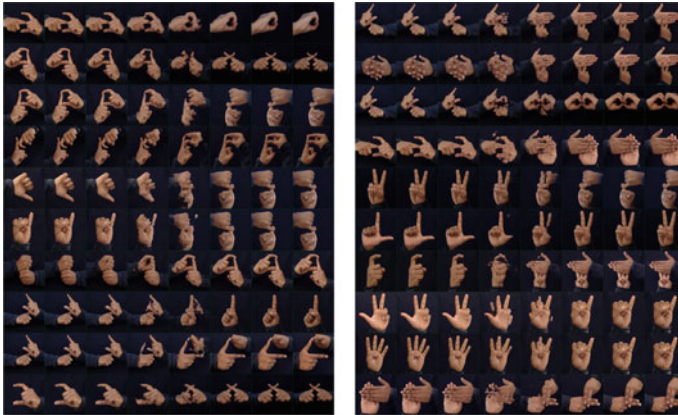


Fig. 18 Images obtained by linear interpolation of latent vectors with z -dim fixed for StyleGAN2-ADA

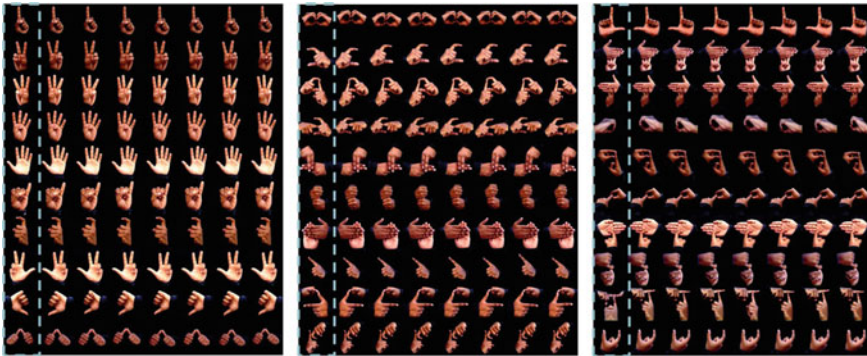


Fig. 19 (left) 8-NN Analysis on 10 Classes, (middle) 8-NN Analysis on 20 classes (right) 8-NN Analysis on 30 Classes for StyleGAN2-ADA

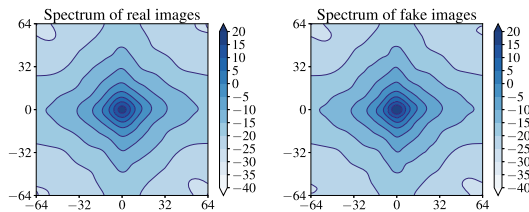


Fig. 20 FFT Spectrum of real and fake images for StyleGAN2-DiffAug

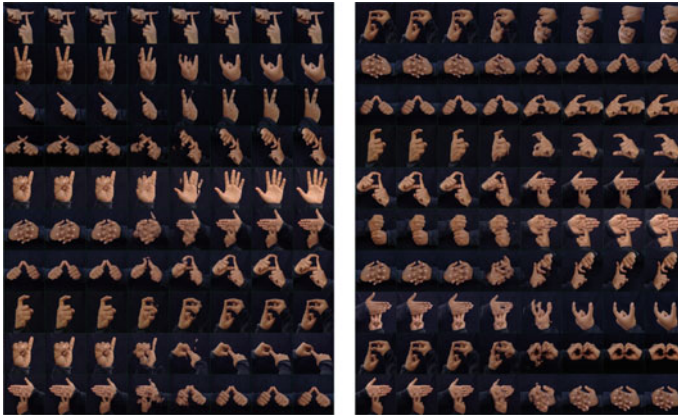


Fig. 21 Images obtained by linear interpolation of latent vectors with z -dim fixed for StyleGAN2-DiffAug

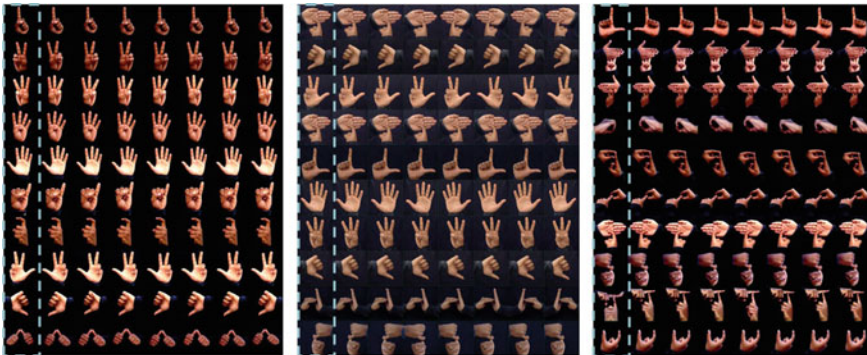


Fig. 22 (left) 8-NN Analysis on 10 Classes, (middle) 8-NN Analysis on 20 classes (right) 8-NN Analysis on 30 Classes for StyleGAN2-DiffAug

5.7 Results of StyleGAN2-DiffAug on ISL Dataset

COLOR,TRANSLATION,CUTOUT [12] policy is applied to images in the ISL dataset. It is important that the transformations applied to the images to be differentiable for back-propagating the gradients to the input. For implementation of this CGAN, StyleGAN2 architecture is used along with transformations described earlier. Results of Frequency analysis for StyleGAN2-DiffAug is shown in Fig. 20. Results of linear interpolation of latent vector is shown in Fig. 21. Results of 8-NN analysis for StyleGAN2-DiffAug is shown in Fig. 22.

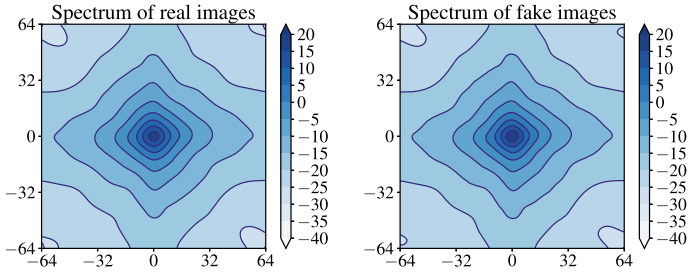


Fig. 23 FFT Spectrum of real and fake images for StyleGAN3

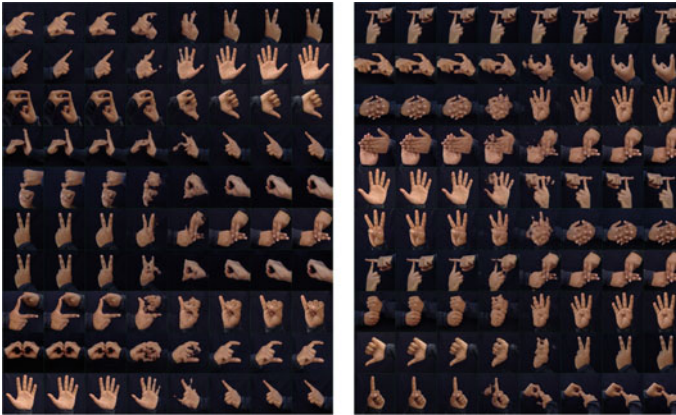


Fig. 24 Images obtained by linear interpolation of latent vectors with z -dim fixed for StyleGAN3

5.8 Results of StyleGAN3 on ISL Dataset

StyleGAN3-R architecture which is proposed in [13] is adopted in this work. It aims to solve aliasing artifacts in the images which stem from feature maps that are discretely sampled and Non-linearities which are point wise. All the network layers in StyleGAN3 are entirely redesigned to move feature maps from discrete domain to Continuous domain. This include Convolutions, Upsampling/Downsampling and Non-linearities. Results of Frequency analysis for StyleGAN3 is shown in Fig. 23. Results of linear Interpolation of latent vector for StyleGAN3 is shown in Fig. 24. Results of 8-NN Analysis for StyleGAN3 is shown in Fig. 25.

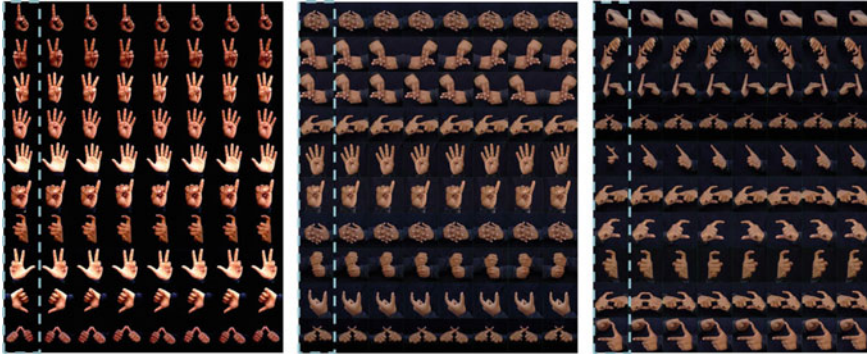
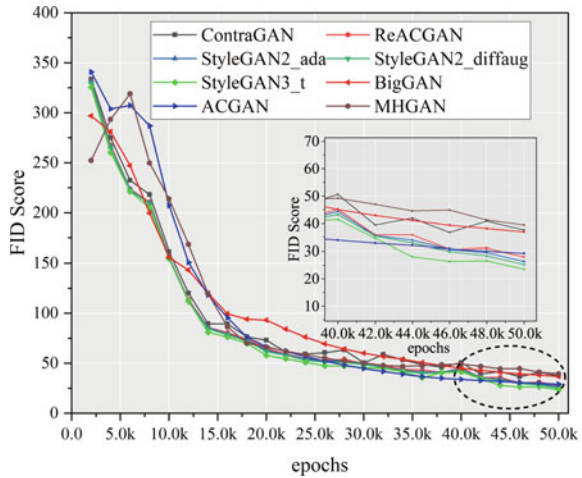


Fig. 25 (left) 8-NN Analysis on 10 Classes, (middle) 8-NN Analysis on 20 classes (right) 8-NN Analysis on 30 Classes for StyleGAN3

Fig. 26 FID score vs epochs for all the CGANs shown in Table 2



5.9 FID Score and IS of CGANs on ISL Dataset

FID score vs epochs for all the CGANs is shown in Fig. 26. Inception Score of all the models trained is shown in Fig. 27. Best Inception Score is given by StyleGAN3(15.65).

Best FID scores and Inception scores at the end of 50k epochs of all the CGAN models trained is shown in Table 4. Best FID and Inception score are marked with bold face.

Fig. 27 Inception score vs epochs for all the CGANs shown in Table 2

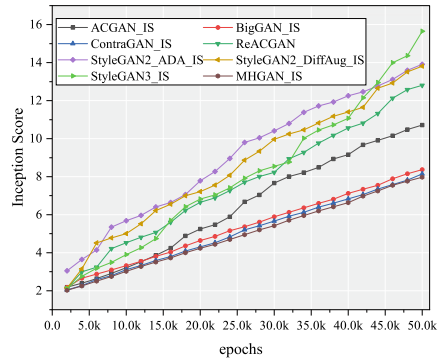


Table 4 FID Score and Inception Score of all the models trained

Model	ISL dataset	
	FID↓	IS↑
BigGAN	36.88	8.37
ACGAN	29.28	10.72
StyleGAN2-DiffAug	25.30	13.91
StyleGAN2-ADA	26.38	13.81
ContraGAN	37.55	8.14
ReACGAN	28.01	10.56
StyleGAN3	23.55	15.65
MHGAN	39.51	7.97

6 Conclusion

One of the bottleneck issues in the research on Indian Sign Language recognition employing Deep Learning architectures is the scarcity of data. Generative Adversarial Models helps to resolve this problem to an extent. This paper proposes a comparative analysis of synthetic data generation using GAN architectures viz. BigGAN, ACGAN, StyleGAN2-DiffAug, StyleGAN2-ADA, ContraGAN, ReACGAN, StyleGAN3 and MHGAN. The analysis shows that StyleGAN3 provides synthetic data with lowest FID and highest IS, for 50000 epochs. An augmented dataset will increase the recognition accuracies of DL models and could be extended to real-time assistive applications with sign language.

References

1. Tang S (2020) Lessons learned from the training of gans on artificial datasets. IEEE Access 8:165044–165055

2. Kang Mi, Shim W, Cho M, Park J (2021) Rebooting ACGAN: auxiliary classifier GANs with stable training. In: *Advances in neural information processing systems*, vol 34 (2021)
3. Karras T, Aittala M, Hellsten J, Laine S, Lehtinen J, Aila T (2020) Training generative adversarial networks with limited data. *Adv Neural Inf Process Syst* 33:12104–12114
4. Yao Y, Wangr X, Ma Y, Fang H, Wei J, Chen L, Anaissi A, Braytee A (2022) Conditional Variational Autoencoder with Balanced Pre-training for Generative Adversarial Networks. arXiv preprint [arXiv:2201.04809](https://arxiv.org/abs/2201.04809)
5. Poorna SS, Reddy MRK, Akhil N, Kamath S, Mohan L, Anuraj K, Pradeep HS (2020) Computer vision aided study for melanoma detection: a deep learning versus conventional supervised learning approach. In: *Advanced computing and intelligent engineering*. Springer, Singapore, pp 75–83
6. Chandra BVB, Naveen C, Kumar MMS, Bhargav MSS, Poorna S, Anuraj K (2021) A comparative study of drowsiness detection from Eeg signals using pretrained CNN models. In: *2021 12th international conference on computing communication and networking technologies (ICCCNT)*, pp 1–3. <https://doi.org/10.1109/ICCCNT51525.2021.9579555>.
7. Aloysius N, Geetha M (2017) A review on deep convolutional neural networks. In: *2017 international conference on communication and signal processing (ICCSP)*, pp 0588–0592. <https://doi.org/10.1109/ICCSP.2017.8286426>.
8. Geetha M, Manjusha C, Unnikrishnan P, Harikrishnan R (2013) A vision based dynamic gesture recognition of Indian Sign Language on Kinect based depth images. In: *2013 international conference on emerging trends in communication, control, signal processing and computing applications (C2SPCA)*, pp 1–7. <https://doi.org/10.1109/C2SPCA.2013.6749448>
9. Kang M, Park J (2020) Contragan: contrastive learning for conditional image generation. *Adv Neural Inf Process Syst* 33:21357–21369
10. Brock A, Donahue J, Simonyan K (2018) Large scale GAN training for high fidelity natural image synthesis. arXiv preprint [arXiv:1809.11096](https://arxiv.org/abs/1809.11096)
11. Kavalerov I, Czaja W, Chellappa R (2021) A multi-class hinge loss for conditional GANs. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp 1290–1299
12. Zhao S, Liu Z, Lin J, Zhu J-Y, Han S (2020) Differentiable augmentation for data-efficient GAN training. *Adv Neural Inf Process Syst* 33:7559–7570
13. Karras T, Aittala M, Laine S, Härkönen E, Hellsten J, Lehtinen J, Aila T (2021) Alias-free generative adversarial networks. *Advances in Neural Information Processing Systems* 34:1–12

A Novel Approach for Segmenting Coronary Artery from Angiogram Videos



K. Kavipriya and Manjunatha Hiremath

Abstract This paper addresses the research focuses on coronary artery disease; it is one of the major heart diseases affecting the people all around the world in the recent era. This heart disease is primarily diagnosed using a medical test called angiogram test. During the angiogram procedure the cardiologist often physically selects the frame from the angiogram video to diagnose the coronary artery disease. Due to the waning and waxing changeover in the angiogram video, it's hard for the cardiologist to identify the artery structure from the frame. So, finding the keyframe which has a complete artery structure is difficult for the cardiologist. To help the cardiologist a method is proposed, to detect the keyframe which has segmented artery from the angiogram video.

Keywords Video · Coronary angiogram · Segmentation · Keyframe · Coronary artery

1 Introduction

Extracting the coronary artery region from the angiogram video is a challenging one. Till studies are going on this area. In this research work author focus on coronary artery disease. This disease is caused because of the blockage in the lumen of the artery, which lessens the blood flow to the heart [1], if the blood flow is not proper to the heart muscles it will lead to stroke or heart attack [2]. In the heart, three main arteries are track over the heart's exterior portion and supply the blood. Among them one right side artery and two left side arteries. Left side artery is the vital because it supplies more blood to the heart and it located in front portion of the heart. If the stenosis blocked the left anterior descending (LAD), it causes a critical stage called Widow maker. To avoid such situation finding the blockage as fast is necessary to say the life of the people. According to the identified Blockage percentage in the artery the doctor starts the treatment. If the blockage is under 50% its mild stage, then the

K. Kavipriya (✉) · M. Hiremath
CHRIST (Deemed to Be University), Bangalore, India
e-mail: kavipriya.k@res.christuniversity.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
P. P. Joby et al. (eds.), *IoT Based Control Networks and Intelligent Systems*,
Lecture Notes in Networks and Systems 528,
https://doi.org/10.1007/978-981-19-5845-8_14

191

physician will suggest the medication. If the blockage is above 75% [3], it's a critical stage then physician strongly considered for the surgery mostly. For diagnosing this coronary artery disease doctors mainly use the angiogram test. During an angiogram test, a catheter is inserted into the wrist, when it reaches the spot, a dye is inoculated through it. During this test, the x-ray machine will start capturing the image at a different angle. While capturing the picture in a different position the artery shape will alter over time due to the heartbeat. It's difficult for the doctor to identify the particular frame where the artery is visible completely. To solve this problem a method is proposed to detect the keyframe which extract the blood vessel region from the coronary Angiogram video.

The keyframe is the frame where the blood vessel is fully visible, over this the doctor can verdict the degree of the stenosis. Keyframe extraction is the major part of the detection of stenosis in the angiogram video. Once the keyframe is detected properly it helps the researcher to go for the further steps like segmentation, feature extraction, and detection of stenosis.

In this paper Sect. 1 presented the introduction about the research area and Sect. 2 presented the existing method and the research gap. In Sect. 3 the Dataset is described, while in Sect. 4 Proposed research method is presented and in Sect. 5 states the experimental result and analysis. Finally, Sect. 6 shown the conclusion and future direction.

2 Literature Review

In [4], author presented a method to select the frame automatically using two different methods: Frangi and parallel curves. The author analysed the artery visibility in each frame of an angiogram video sequence to find the best frame in it. However, in this paper the Keyframe extracted are not optimal. In [5], author proposed a method to extract the high-contrast coronary angiogram image automatically using a Hessian method and adaptive feature transformation method. In this research work, 20 angiogram frame sequences were used to experiment. Through this segmentation method, the author was able to distinguish the artery and the background in the X-ray images. However, in this research work the normalization of artery region in the image is missing. In [6], author presented a segmentation method to find the arteries in the coronary angiogram images. The author used a single-scale Gabor filter using the Boltzmann univariate marginal distribution algorithm in angiogram images to segment the artery. For the experiment author used 40 angiogram images as a training set and got an accuracy of 95.02%. This research work's segmentation accuracy is less than the other existing method. In [7], author presented a method to segment and track the artery using graph-based Formulation besides the temporal priors technique. In this method, superpixel groups are also used to extract the coronary artery. 12 sequence of angiogram images was used for the experiment to show the robustness and efficiency of the method. However, the artery tracking leads to average performance and the result is not optimal to the research work. In [8], author

proposed a method to segment the vascular structure from angiogram frames. In the first phase, a morphological operator is applied to the frames to remove the noise from the image. In the second phase by using robust principal component analysis quasi-static structures from the frames were removed. In the third phase, the frames were smoothed and then, by using the multi-feature combination vascular structure is segmented. In this research 22 X-ray coronary angiogram image sequence was used to demonstrate performance of the segmentation method. In [9], the researcher created a framework to extract the artery from the sequence of images. In this, the author considered the artery area as a moving part and background as a static part to form a matrix decomposition model. Once the background was removed and the artery region was enhanced by the Hessian matrix, spatiotemporal energy function was used to fine-tune the artery structure. The researcher used 40 angiogram videos, captured using X-ray machine. This paper concentrates only on the segmentation part, here keyframe selection is not automatic. In the Existing methods, the researcher tried different techniques to identify the keyframe from the angiogram video, but the results of these papers are more than 15 keyframes. But in our proposed method, minimum keyframes (less than 10) are identified compared to other methods and the resulted keyframe has complete and clear artery structure. By using these results, researchers can find the stenosis area easily. Our proposed method will overcome all the limitations in other existing methods.

3 Dataset

In this research, the data used are collected real-time from the Rahavendar hospital with proper permission. 50 patients datas are collected in the hospital from 2019 to 2020. Among the 50 patients 44 are men and 6 are women. The data is in video format. The medical expert rotates the C-Arm device in different angles to capture the Angiogram images as a frame.

- Dataset Type: Angiogram Videos
- Sample size: 50
- Type of machine: Philips Allura Xper FD20 X-ray C-arm system.
- Video rate: 10–15 fps.
- Angles Considered: Caudal, Cranial
- Projection: Left Anterior Oblique (LAO), Right Anterior Oblique (RAO), Postero-Anterior (PA or AP).

4 Proposed Research Method

In the proposed method, the initial stage frames are extracted from the angiogram video, then the black area surrounding the key part of the frames is removed based

on the threshold value. Once the black area in the frames is removed, the required region is extracted using Frangi method. Finally, the keyframe is identified.

4.1 *Extracting the Frames from the Angiogram Video*

The coronary angiogram videos are captured using the X-ray machine at different angles. These videos are then stored to form the dataset. Each video VI_i is read from the dataset as shown in Eq. (1):

$$VI_i = \{VI_1, VI_2, VI_3, \dots, VI_n\} \tag{1}$$

where $1 \geq VI \leq n$.

The frames from the angiogram video are extracted and stored in a directory for further analysis. It is shown in Eq. (2):

$$FI_i = \{FI_1, FI_2, FI_3, \dots, FI_N\} \tag{2}$$

where $1 \geq FI \leq N$.

N denotes the number of Frames present in the Angiogram video. Read the video through the video reader object in MATLAB [10] and then the number of frames is calculated using the object's property. For each frame (FI) in the video is incremented 1 till it reaches the N . Then one by one the frame is read and stored in the secondary memory for further process.

From the angiogram video, 45 frames are extracted as shown in Fig. 1.

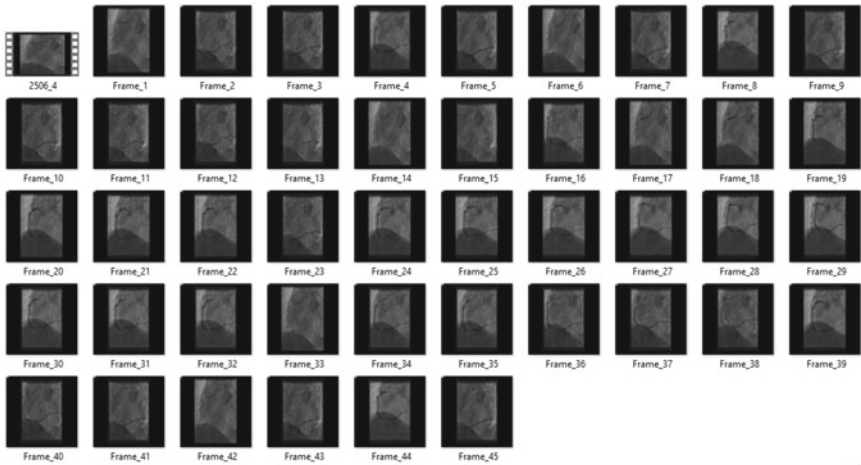


Fig. 1 Sample of the frames extracted from the video



Fig. 2 Sample of the preprocessed frames

4.2 Frame Pre-processing

Once the frames are extracted from the video then all the frames are converted into the grayscale format. Converting the frames into grayscale will better the diagnostically substantial information in the image. This is analogous to deploying the [11] sensitometry curve of X-ray images to enhance radiographic information in the image. The extracted frames consisted of an irrelevant black area, which needed to be removed. This helped in enhancing the frame quality. Generally, by using the thresholding method light or dark regions in the image can be separated. Hence, the same thresholding method [12] is used for the removal. By keeping the threshold value as 60 the black region in the frames is removed using the following Eq. (3). The output of the same is depicted in Fig. 2.

$$G(x, y) = \begin{cases} 1 & \text{if } FI(x, y) \geq T \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where $G(x,y)$ is an enhanced Frame of $FI(x,y)$ frame at threshold T .

4.3 Segmenting the Region of Interest

Image segmentation is possible by considering the similarity and discontinuity of the object. Segmenting the foreground from the background is an essential part of image processing. Here, angiogram image detecting the curvilinear structure [13] is

an important step for segmenting the artery. The direction and strength of the artery can be represented by the eigenvectors and eigenvalues of the Hessian matrix [14]. The Frangi method was used to extract tube-like structures [15, 16], hence it is used in this proposed method to segment the artery from the image. This method helps to get the direction of the artery structure and to premeditate the probability of the local artery. This method takes full contemplation of all the eigenvalues and the artery to interpret the spontaneous geometry features. By using these features artery can be detected.

$$H = \begin{bmatrix} G_{xx} & G_{yy} \\ G_{yx} & G_{yy} \end{bmatrix} \quad (4)$$

$$\lambda_{1=K-\sqrt{K^2-Q^2}}, \lambda_{2=K+\sqrt{K^2-Q^2}} \quad (5)$$

where $G(x,y)$ enhanced the image,

$$K = (G_{xx} + G_{yy})/2 \quad (6)$$

$$Q = \sqrt{G_{xx}G_{yy} - G_{xy}G_{yx}} \quad (7)$$

The values obtained from the λ_1 and λ_2 helps to identify the coronary artery structure.

From the following function, the structure of the vessel can be computed.

$$VI(\sigma) = \begin{cases} 0 & \text{if } \lambda_1 > 0 \\ \exp\left(-\frac{R_D^2}{2\beta^2}\right) \left(1 - \exp\left(\frac{-S1^2}{2c^2}\right)\right) & \text{otherwise} \end{cases} \quad (8)$$

$$R_D = \frac{\lambda_2}{\lambda_1}, \quad 1 = H_F = \sqrt{\lambda_1^2 + \lambda_2^2} \quad (9)$$

where β and c are the parameters that control the Frangi method sensitivity to R_D and $S1$ respectively [17]. Based on the scale-space theory [18], the segmented output is the highest width of artery matches to an appropriate scale factor σ . VI is calculated in various scale factor σ at every pixel, then take the highest one as the concluding output of this filter.

$$Z(x, y) = \max[VI(x, y; \sigma)] \quad (10)$$

From the segmented output to reduce the artifact [19] Gaussian filter and [20] morphological operation is applied.

4.4 Detecting the Keyframe

To identify the frame which has the full structure of the artery, the connected component properties are used. The connected components in a segmented image are a set of pixels that form a related group [21]. For all the frames in the dataset, the connected component properties and their relation is calculated.

Since the artery has an elongated structure, it can be easily identified in each frame using the highest connected components. The entire steps are shown in Fig. 3.

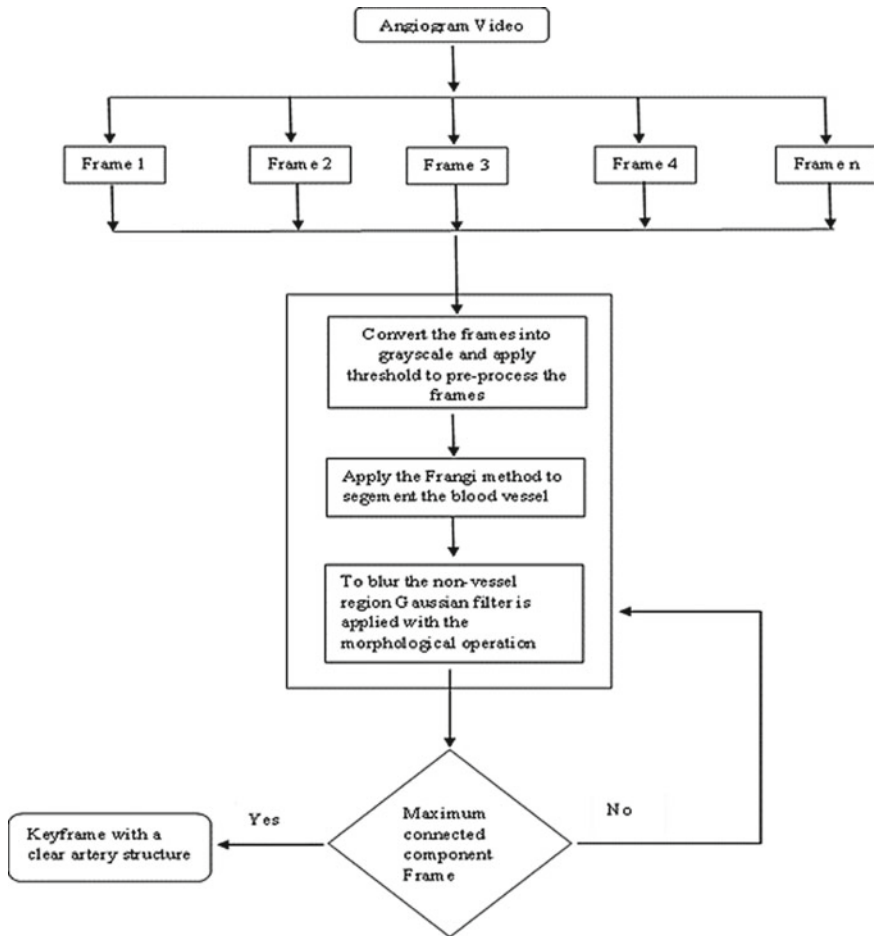


Fig. 3 Flow diagram of the proposed method

5 Experimental Results and Analysis

This work is an experiment on coronary angiogram dataset. In the first module, the Angiogram video is separated into multiple frames and stored as a folder. Then, in the second module black area in the frames is removed by using the threshold value. In the third module the Frangi filter is used on the frames to segment the foreground (artery). Further frames are sharpened and blurred by a Gaussian filter. Then, morphological operator is used to enhance the frames. In the fourth module number of connected component in each frame are calculated. Among them, the frame containing the highest connected component is considered as a keyframe and is stored in the directory for further processes like stenosis detection.

Table 1 represent the result evaluation of the proposed method. In this table first column represents the video ID and second column for video size, the third column for the number of frames extracted from the video using this proposed keyframe extraction method also shown in Fig. 4, and the last column state the segmentation accuracy of the keyframe.

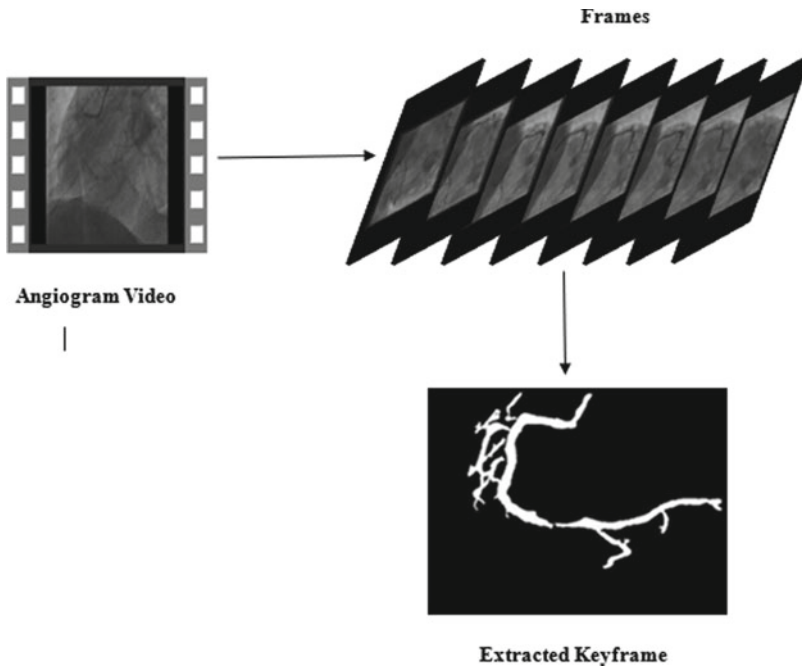


Fig. 4 The proposed method with the result

Table 1 Evaluation of the proposed method

Video	Video size (MB)	Frames	Key frames	Segmentation accuracy
1	28.3	63	4	0.94
2	20.6	46	3	0.95
3	14.8	33	1	0.96
4	15.8	36	1	0.96
5	18.0	40	2	0.95

6 Conclusion

Coronary artery disease is the major heart disease in recent days, so finding the disease in its early stage is important to save human life. Cardiologists are facing difficulties in diagnosing the stenosis from the angiogram video, because of the artifacts present in it. Also, identifying the frame which has a complete structure of the artery is a challenging task. To solve this problem, an automatic method to detect the keyframe and the structure of the artery is proposed. This work can be further extended to identify the exact location of the stenosis.

References

1. Zifan A, Liatsis P ((2016)) Patient-specific computational models of coronary arteries using monoplane X-ray angiograms. *Comput Math Methods Med*
2. Badila E, Calmac L, Zamfir D, Penes D, Weiss E, Bataila V (2017) The cardiovascular system and the coronary circulation. In: Itu L, Sharma P, Suciuc C (eds) *Patient- specific Hemodynamic Computations: Application to Personalized Diagnosis of Cardiovascular Pathologies 2017*. Springer, Cham, pp 13–59
3. Jiangping S, Zhe Z, Wei W, Yunhu S, Jie H, Hongyue W, Hong Z, Shengshou H (2013) Assessment of coronary artery stenosis by coronary angiography: a head-to-head comparison with pathological coronary artery anatomy. *Circ Cardiovasc Intervent* 6:262–268
4. Syeda-Mahmood T, Beymer D, Wang F, Mahmood A, Lundstrom RJ, Shafee N, Holve T (2010) Automatic selection of keyframes from angiogram videos. In: 2010 20th international conference on pattern recognition. IEEE, pp 4008–4011
5. Tsai YC, Lee HJ, Chen MYC (2015) Automatic segmentation of vessels from angiogram sequences using adaptive feature transformation. *Comput Biol Med* 62:239–253
6. Cervantes-Sanchez F, Cruz-Aceves I, Hernandez-Aguire A, Avila-Cervantes JG, Solorio-Meza S, Ornelas-Rodriguez M, Torres-Cisneros M (2016) Segmentation of coronary angiograms using gabor filters and boltzmann univariate marginal distribution algorithm. *Comput Intell Neurosci* 2016
7. M'hiri F, Duong L, Desrosiers C, Leye M, Miró J, Cheriet M (2016) A graph-based approach for spatio-temporal segmentation of coronary arteries in X-ray angiographic sequences. *Comput Biol Med* 70:45–58
8. Song S, Du C, Chen Y, Ai D, Song H, Huang Y, Yang J (2019) Inter/intra-frame constrained vascular segmentation in X-ray angiographic image sequence. *BMC Med Inf Decis Mak* 19:1–11
9. Xia S, Zhu H, Liu X, Gong M, Huang X, Xu L, Guo J (2019) Vessel segmentation of X-ray coronary angiographic image sequence. *IEEE Trans Biomed Eng* 67:1338–1348

10. Nachamai M, Paulose J, Marandi S (2018) A Comparative analysis of the efficiency of video reader object for frame extraction in MATLAB. *J Multimedia Process Technol* 9:16–21
11. Barnes GT, Lauro K (1989) Image processing in digital radiography: basic concepts and applications. *J Digit Imaging* 2:132–146
12. Norouzi A, Rahim MSM, Altameem A, Saba T, Rad AE, Rehman A, Uddin M (2014) Medical image segmentation methods, algorithms, and applications. *IETE Tech Rev* 31:199–213
13. Cui H, Xia Y, Zhang Y (2019) 2D and 3D vascular structures enhancement via improved vesselness filter and vessel enhancing diffusion. *IEEE Access* 7:123969–123980
14. Canero C, Radeva P (2003) Vesselness enhancement diffusion. *Pattern Recogn Lett* 24:3141–3151
15. Frangi AF, Niessen WJ, Vincken KL, Viergever MA (1998) Multiscale vessel enhancement filtering. In: *International conference on medical image computing and computer-assisted intervention*. Springer, Heidelberg, pp 130–137
16. Jerman T, Pernuš F, Likar B, Špiclin Ž (2015) Beyond frangi: an improved multiscale vesselness filter. In: *Medical Imaging 2015: Image Processing*. vol 9413. SPIE, pp 623–633
17. Kerkeni A, Benabdallah A, Manzanera A, Bedoui MH (2016) A coronary artery segmentation method based on multiscale analysis and region growing. *Comput Med Imaging Graph* 48:49–61
18. Zhang F, Zhang X, Liu X, Cao K, Du H, Cui Y (2014) Blood vessel enhancement for DSA images based on adaptive multi-scale filtering. *Optik* 125:2383–2388
19. Li Z, Zhang Y, Liu G, Shao H, Li W, Tang X (2015) A robust coronary artery identification and centerline extraction method in angiographies. *Biomed Signal Process Control* 16:1–8
20. Bai X, Zhou F (2013) A unified form of multi-scale top-hat transform based algorithms for image processing. *Optik* 124:1614–1619
21. Singh RP, Agarwal P (2013) Extraction of region of interest through e-learning videos with matlab. *Int J Comput Appl* 68:38–40

Analysis of Different Cryptographic Algorithms in Cloud-Based Multi-robot Systems



Saurabh Jain, Shireen Rafat Alam, and Rajesh Doriya

Abstract Nowadays, securing data over the network become an important concern. When the data is travelled between Robots and the Cloud then there is high chance of the eavesdropping of the data by the attacker. To protect the data secrecy and confidentiality that arises because of the threat, cryptography is used. At first, data is converted into unreadable text (Cipher Text) known as encryption and after receiving the data, receiver performs reverse encryption (decryption). There are many algorithms for performing the encryption and decryption process. That's why it is necessary to find out which algorithm performs better with resource-constrained and limited computing ability devices. In this paper, various encryption techniques like DES, 3DES, AES, Blowfish and ECIES are analysed based on encryption and decryption time. This encryption mechanism consumes significant number of resources like memory utilisation, CPU time, computation time, etc. In this paper, we compared different cryptographic algorithms in terms of key size, block size, Power consumption and speed. The simulation results will be calculated in terms of encryption and decryption time of the algorithms.

Keywords Robots · Symmetric Key Cryptography · Asymmetric Key Cryptography · Cloud robotics · ECIES

1 Introduction

In today's world, a digital communication system has become one of the most important requirements. Data and information may now be transmitted electronically, thanks to the rapid advancement of digital communication. System security is

S. Jain (✉) · S. R. Alam · R. Doriya
National Institute of Technology Raipur, Raipur, India
e-mail: sjain.phd2017.it@nitrr.ac.in

R. Doriya
e-mail: rajeshdoriya.it@nitrr.ac.in

an important aspect that ensures the passage of information and data in a communication system. Today, there are numerous approaches for ensuring the security of information [1]. The goal of those solutions is to protect the data from being tapped or stolen while it is being exchanged. For example, Effective security procedures must be established because the data represented by medical photographs is a vital source of information concerning the privacy of patients. Healthcare businesses should pay more attention to installing secure cryptosystems in the Internet of Healthcare Things (IoHT) era to avoid security breaches [2]. The patient's data must be securely stored and sent. Security of data is not an afterthought in data transmission; it is a basic procedure of total protection against all forms of threat. The implementation of a cryptographic algorithm is one of many ways to ensure system security. When a data transmission system communicates through a medium that cannot be trusted, a cryptographic method is required. The use of an encryption algorithms ensures data and information's confidentiality, authentication, non-repudiation and integrity.

Information security in a system is essential for protecting data against 3rd parties who lack the capability to make choices regarding its content [3]. If high-value data was stored on the system and afterwards retrieved by a third party, it would be exceedingly dangerous. One aspect of information protection is password security. We might now use software invoices as passwords to open information that has already been widely available. Another way to encode data is to use cryptography to encrypt it so that it could not be read, damaged, or altered by others [4]. In general, Cryptanalyst considers a cryptographic method to be a purely mathematical thing [3]. The following factors support the assumptions:

- The intruder will choose the ciphertext and plaintext pairs at random.
- The method structure is publicly available.
- The attacker does not know the key.

When a cryptographic mechanism is implemented on hardware, cryptanalysts and IC designers are frequently unaware of these flaws. As a result, the fundamental premise from traditional cryptanalysis cannot be applied in this circumstance. Cryptography is a technology for ensuring the confidentiality of messages. The word "composing mystery" has a unique significance in Greek [4]. Plaintext as cypher text is the working cycle that changes throughout the actual communication. Cryptography nowadays ensures that data transferred is protected, with the goal of assuring that the receiver can access this information through authorized sources; cryptography can be considered an old technology which has been used until now [5]. Cryptography is used by billions of people all around the world to encrypt data and information. The evolution of encryption typically leads to an endless array of conceivable outcomes. Even though it is difficult to prevent hacking, but we can assure the security of our sensitive data although it is hacked by implementing encryption solutions [6].

1.1 *Cryptography Security Services*

The primary focus of this study is on evaluating the performance of current encryption algorithms such as AES, DES, and blowfish. After receiving the measurements and findings, the encryption methods will be evaluated for their ability to withstand various attacks. The encryption procedure and data decryption using various algorithms will be explained in this research. This study is expected to be beneficial in terms of protecting data and maintaining data confidentiality. The following parts such as privacy, verification, integrity of data, methodology that guarantees the messages are delivered between parties, data accessibility rights, availability of data should be evaluated in terms of data security [7].

Security of information is needed while sending the data over the network, as it provides confidentiality, integrity and availability of the data [8]. Due to large evolving data, it can be stored on cloud. This data can be very sensitive and is not available to public. Robot generates huge amount of data which can be offloaded on cloud for processing. Robots can be used for various tasks such as in military services for navigation and map building, diffusion of hazardous bombs where there is risk of human life. Robots can also be used in hospitals, industries, home for helping in daily chores. Applications of robots are Simultaneous Localization and Mapping (SLAM), path planning, grasping, object detection and many more. This application can generate huge data and because of limited memory and processing power of robots, it is offloaded on the cloud to store the data and also to increase the processing speed. Robots used in military services, hospitals and industries will generate sensitive data which will have to be secured while sending over the network [9]. This can be achieved by various encryption methods. Encryption is a method in which plaintext is converted to cipher text using a key. There are two types of encryption algorithms Symmetric and Asymmetric key algorithms.

Our research is structured as follows: The literature review is discussed in Sect. 2. Section 3 discusses symmetric and asymmetric algorithms, followed by Sect. 4, which concentrates on ECIES techniques utilised in previous studies. Section 5 focuses on the findings and debates, while Sect. 6 wraps up the study and discusses future work.

2 Literature Review

Sonal Sharma et al. [11] proposed a limited domain natural number technique that is similar to but not identical to the RSA technique. They increased the cryptosystem's security by using two natural numbers for every pair of keys (public, private). The programmes are written using Java library functions. SRNN with a modulus value of 1024 bits, as per the findings, delivers greater speed and security. Rajni Meelu and Punita Meelu [10] created RSA on JAVA Eclipse Platform utilising multiple text sizes. RSA encryption and decryption results are provided in seconds. They arrive

at the conclusion that encryption takes longer than decryption for the identical text size and key.

For performance evaluation, DES and AES [11] have been built, with memory requirements, simulation time, and the avalanche effect taken into account. The avalanche impact is less for AES but greater for DES, as per research. There is no focus on the number of bytes encrypted at any particular time. [14] shows the energy usage of various typical symmetric key encryptions on mobile devices. Only after 600 Triple-DES encryptions of a 5 MB file, the remaining battery power is revealed to be 45 percent, and future encryptions are not possible because the battery expires soon.

In [12] a comparison of six standard encryption algorithms on three different platforms was done only on the basis of execution time. The trials were carried out on both text and image data. All algorithms were found to run faster on the Windows XP platform. There are no performance metrics that can be used to compare its performance. They concentrated solely on platform performance.

This paper [13] presents tri-brid encryption algorithm, here the first AES key is encrypted using RSA algorithm, and the original file is converted into encrypted file using AES to protect a file stored in the cloud. The important point of this work is the key and the file both are encrypted thus provides better security. Patnaik et al. [14] have used symmetric key cryptography algorithm as well as steganography. Data security is ensured by AES, blowfish, RC6, and RSA algorithms and the performance is improved by introduction of the LSB steganography method. This method provides security while the data is stored in cloud but does not provide security while transmitting the data to the cloud. The author of [18] presents an effective Hybrids and Adapting Cryptographic (HAC)-based safe authentication framework for IoT authentication. To achieve authentication in IoT devices, the suggested approach employs cryptographic operations such as hybrid cryptography, a hashing function, and the exclusive-or operation.

The study and implementations of certain symmetric cryptographic methods are presented in this paper [19]. It selects the most effective symmetric method and integrates it with an asymmetric strategy to demonstrate the value of a hybrid encryption architecture for cloud security. This paper explains how to achieve cloud privacy, secrecy, and integrity. It also defends against bot attacks and minimises the server's load. According to the author of [20], using a single method to ensure high-level data security in cloud computing is ineffective. To address this, the study proposes a new security mechanism based on a built-in secured cloud infrastructure, which makes the system more resistant to security flaws. Anas et al. [21] proposed a method for cloud security utilising ECC, as well as reviewing recent experiments and discussing several phases of ECC, which will be useful for future research.

Gurav et al. [22] developed a solution based on decentralised Information Flow Control (DIFC) to address cloud security concerns. The newly developed Two-Fold Improved Poor Rich Optimization Algorithm was used to determine the best key for the hybrid AES-ECC encryption architecture (TF-IPRO). The proposed model was compared to different mechanisms and current models such as DIFC with Hybrid AES-ECC + SSA. Based on the results, the suggested approach has achieved the

shortest encryption time for the chess dataset. Chinnasamy et al. [23] gave a thorough explanation of hybrid encryption access control for reliable HER retrieval in the healthcare cloud. The research proposes a new hybrid cryptography technique to provide safe cloud storage, in which the Improved Key Generation System of RSA (IKGSR) algorithm is used to encrypt health data and the Blowfish algorithm is utilised for key encryption. They also used steganography-based access control with substring indexing and keyword search mechanisms to efficiently extract the encrypted material for key sharing.

3 Background

Encryption is a vital tool for safeguarding confidential data. The objective of encryption in communications is to protect privacy (by prohibiting disclosure or confidentiality). It is the process of encrypting data, a method of conversing with someone while others are listening. However, some are unable to comprehend what you are saying [15]. Data security is mostly dependent on encryption methods, protection against harmful attacks. Security is extremely important with mobile devices. To prevent this, various types of algorithms are utilised. Malicious attack on data transmission. Algorithm for encryption can be divided into two types: symmetric key (private) and asymmetric key (public) [16].

3.1 *Symmetric Key Cryptography*

Symmetric-key algorithms use the same key for encrypting the plain-text and also for decrypting the ciphertext as shown in Fig. 1. It is considered as more secure than asymmetric-key algorithms. It is very fast therefore it is generally used to encrypt large amount of data [17]. Some of the common encryption algorithms are DES, 3DES, AES, IDEA, Blowfish, two fish and RC5.

3.1.1 **Data Encryption Standard (DES)**

Data Encryption Standard is an encryption algorithm which encrypts data in 64-bit block. This algorithm converts 64-bit plain text into 64-bit cipher text. DES is symmetric algorithm where same key is used for both encryption and decryption process [18]. The length of the key used is 56-bit. This algorithm is combination of basic confusion and diffusion techniques of encryption. DES describes the number of combinations of these techniques known as rounds. There is total 16 rounds in DES algorithm.

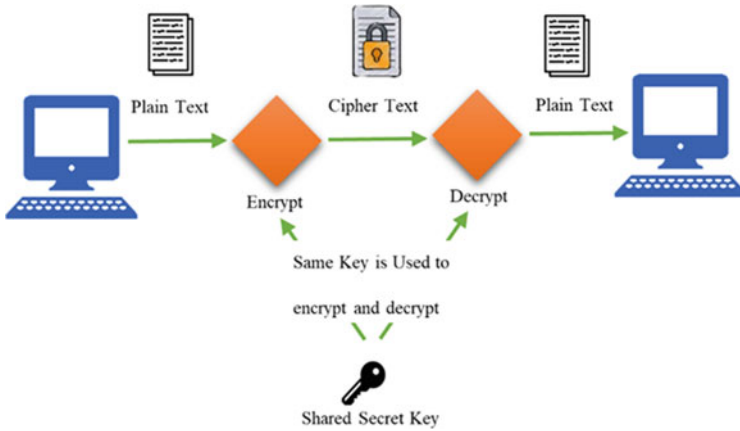


Fig. 1 Symmetric-key algorithm

3.1.2 Triple-DES

In Triple DES, three times block-cipher mechanism is applied to each data blocks. To increase additional security while encryption, key size is increased in Triple DES refers to a packet of keys with each key having 56 bits. There are three keying choices for data There are three keying choices for data encryption:

- 1) No keys are dependent on each other.
- 2) Keys 1 and 2 should be self-contained.
- 3) Each of the the keys should be the same.

Out of the above-mentioned standard, Triple DES follows third option. In spite of being 168 bits of key, length of key fall to 112 bits because of security falls. Triple DES is drawback compatible with regular DES.

3.1.3 Advanced Encryption Standard (AES)

Advance Encryption Standard (AES) is a standard use to encrypt sensitive data by implementing it in hardware and software. This standard was inodiated by U.S. government to protect the sensitive information [19]. It is a symmetric key algorithm which provides transparent analysis of the design where AES is been applied. AES is made up of three different block cyphers: AES-128, AES-192, and AES-256. Each cypher encrypts and decrypts a 128-bit data block utilising cryptographic keys of 128-bit, 192-bit, and 256-bit. AES specifies the number of transformations to be performed on data stored in an array. The number of rounds necessary for 128-bit keys is 10, 12 for 192-bit keys, and 14 for 256-bit keys, with the number of rounds determined by the key length.

3.1.4 Blowfish

Blowfish can be used instead of IDEA or DES. It's a symmetric block cypher with keys ranging from 32 to 448 bits in length. This property qualifies it for both export and domestic use. Blowfish was created by Bruce Schneier in 1993 as a free and fast alternative for conventional encryption algorithms. Since then, it has progressively gained popularity as a powerful encryption technique and has been extensively studied. Blowfish is unpatented and license free and it is available for free to all users.

3.2 Asymmetric Key Cryptography

The public-key encryption algorithm is also called as the asymmetric-key algorithm. Both systems, public and private keys, have a set of keys in asymmetric-key algorithms. Because of their intricate structure, asymmetric algorithms are slower than symmetric algorithms. Asymmetric-key algorithms like: RSA, ECC and Diffie Hellman.

3.2.1 Rivest Shamir Adelman (RSA)

The algorithm stands for Rivest Shamir Adleman algorithm. named after its inventors [20]. It falls under the class of public key cryptography (asymmetric key) dependent on number theory. Along with data encryption it is also designed for digital signature. In the proposed RSA algorithm, two very large prime numbers are selected for generating secret and public key. Involvement product of two large prime numbers for key generation creates complexity in guessing plain text from public key as well as cipher text.

$$De = 1 \text{ mod } ((p - 1)(q - 1)) \quad (1)$$

where p, q are two chosen large prime numbers and d, e are keys.

3.2.2 Elliptic Curve Cryptography (ECC)

Elliptic curve cryptography is one of the most widely used asymmetric cryptographic algorithm for resource-constraint devices [24]. ECC can be utilised in devices with little computing memory and processing, such as robots, IoT devices, and mobile phones, due to the small size of the keys. It is faster compared to other cryptographic algorithms because of it does not work on bilinear pairing functions. It is

computationally hard because it uses discrete logarithm problem (DLP). The security provided by other algorithms using a 1024-bit key is equivalent to the security provided by ECC using a 164-bit key. For real numbers, the elliptic curve equation is as follows:

$$y^2 = x^3 + ax + b, \text{ where } 4a^3 + 27b^2 \neq 0. \quad (2)$$

The dP (point multiplication) is given as the addition of P repeated d times using the point P upon that elliptic curve and the scalar d ($dP = P + P + P \dots + P$ (d times)). It uses smaller key size to encrypt and decrypt of the data.

4 Elliptic Curve Integrated Encryption Scheme (ECIES)

ECIES is a hybrid encryption system with an integrated encryption mechanism. It provides the attacker with semantic security. The public key from ECC is utilised to derive the symmetric encryption in ECIES [25]. Figure 2 shows how symmetric-key is generated in ECIES using AES. As a Private Key, the sender will produce a random key using the stated technique (d_A). On the elliptic curve, G will be a point. The public key (Q_A) would be produced using d_A and G.

$$Q_A = d_A \times G \quad (3)$$

Thus, the points on the elliptic curve are Q_A and G. Public Key (Q_A) will be sent by the sender to the receiver. Receiver will generate:

$$R = r \times G \quad (4)$$

$$S = r \times Q_A \quad (5)$$

where, r = random number generated.

In this step, a symmetric encryption will be generated and used to secure communications. The encrypted message will be sent to the sender along with R, which will allow the sender to compute the symmetric key that will be needed to decrypt the information.

$$S = d_A \times (r \times G) \quad (6)$$

$$S = r \times (d_A \times G) \quad (7)$$

$$S = r \times Q_A \quad (8)$$

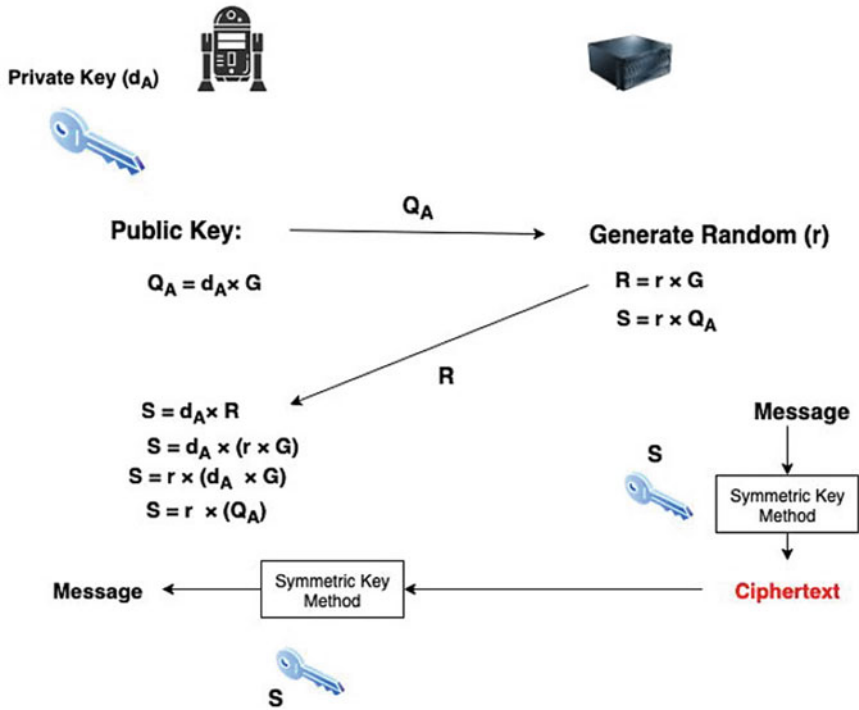


Fig. 2 ECIES using AES

Thus, S = symmetric key, which is same key that is generated by receiver. ECIES, makes use of the five functions:

- 1) **Key Agreement (KA):** This function is used for generating the secret key and is used by the parties involved in the communication.
- 2) **Key Derivation Function (KDF):** This function makes use of keying material and produces the set of keys and some other parameters involved.
- 3) **Encryption (ENC):** Encryption Algorithm (Symmetric).
- 4) **MAC:** Message Authentication Code (MAC) is utilised with respect to authenticate the integrity of the messages.
- 5) **Hash:** This is the Digest function which is used inside the Message Authentication Code function and Key Derivation functions.

4.1 Encryption Algorithm For ECIES

ALGORITHM 1: ALGORITHM FOR THE PROCESS OF ENCRYPTION.

Input: Public and private key of Alice i.e., u and U respectively, public key of Bob, v .
 Message that needs to be delivered to Bob, m .

Output: Cryptogram containing encrypted message.

- 1 Ephemeral **Key-pair Generation** by Alice: $U = u * G \rightarrow$ Alice (u, U)
- 2 Using **Key Agreement Function**, shared secret value is generated $\rightarrow u * V$
- 3 With the secret value and other optional parameters using **Key Derivation Function** $\rightarrow K_{ENC} || K_{MAC}$
- 4 Applying Encryption Algorithm on the message m along with $K_{ENC} \rightarrow$ Encrypted Message, c
- 5 **TAG production:** $c + K_{MAC} +$ other optional parameters + MAC functions $\rightarrow TAG$
- 6 Cryptogram is sent to the Bob \rightarrow Cryptogram($U || TAG || c$)

4.2 Decryption Algorithm for ECIES

ALGORITHM 2: ALGORITHM FOR THE PROCESS OF DECRYPTION.

Input: Cryptogram from Alice, Cryptogram($U || TAG || c$)

Output: Plain Text

- 1 **Shared secret value generation, $v * U$**
 $v * U == u * V \rightarrow$ Core concept of **Diffie-Hellman Procedure**.
- 2 Using **Key Derivation Function** along with shared secret value with the same optional parameter $\rightarrow K_{ENC} + K_{MAC}$ (same as Alice)
- 3 **TAG* is computed:** $K_{MAC} + c +$ same optional parameter $\rightarrow TAG*$
- 4 **if TAG* != TAG:**
- 5 | Cryptogram is **rejected** \rightarrow failure of MAC verification Procedure
- 6 **else:**
- 7 | Cryptogram is **accepted** \rightarrow deciphering continued
- 8 | Using **Symmetric Encryption Algorithm** along with K_{ENC} applied on c
- 9 | **Plain Text** is generated.
- 10 **end**

4.3 Notations

Table 1 defines the notations that are used throughout this paper.

5 Result and Discussion

Tables 2 and 3, displays the experimental results for the encryption algorithms AES, DES, 3DES, ECIES, Blowfish and RSA, which compares the three algorithms' AES, DES, 3DES, ECIES, Blowfish and RSA using the same text file for five experiments. Parameters for evaluation that are used to evaluate the encryption algorithm's performance: Encryption Time, Decryption Time, Encrypted File Size, Power Consumption, Speed, Security.

Table 1 Notation guide

u	Public key of Alice
U	Private key of Alice
v	Public key of Bob
V	Private key of Bob
m	Pure message
c	Encrypted message
G	Generator
k_{ENC}	Encryption key
k_{MAC}	MAC key

Table 2 Encryption time analysis of bag files

Data size	Encryption time (seconds)					
	DES	3DES	AES	Blowfish	RSA	ECIES
25 KB	0.417	0.328	0.521	0.132	0.59	0.632
50 KB	0.532	0.471	0.63	0.46	0.81	0.71
1 MB	0.831	0.801	0.69	0.51	1.24	0.79
2 MB	1.02	1.238	0.72	0.60	1.53	0.93
3 MB	1.32	1.62	0.81	0.72	2.01	1.02
4 MB	1.59	2.08	0.89	0.831	2.53	1.36
5 MB	1.88	2.327	0.97	0.91	2.94	1.1
6 MB	2.18	2.61	1.3	0.99	3.56	1.6
7 MB	2.52	3.12	1.72	1.32	3.92	1.92
8 MB	2.81	3.56	2.09	1.83	4.43	2.31

Table 3 Decryption time analysis of bag files

Data size	Decryption time (seconds)					
	DES	3DES	AES	Blowfish	RSA	ECIES
25 KB	0.4	0.38	0.373	0.121	0.32	0.231
50 KB	0.42	0.41	0.591	0.182	0.52	0.56
1 MB	0.62	0.58	0.62	0.33	0.823	0.63
2 MB	0.801	0.78	0.599	0.36	1.19	0.66
3 MB	0.92	0.88	0.62	0.63	1.61	0.71
4 MB	1.14	0.98	0.628	0.62	1.93	0.79
5 MB	1.29	1.13	0.641	0.69	2.16	0.81
6 MB	1.37	1.21	0.658	0.75	2.62	0.832
7 MB	1.46	1.29	0.679	0.83	3.01	0.854
8 MB	1.58	1.37	0.78	0.91	3.63	0.931

The computation time is the time it takes for an encryption algorithm to convert plain text to cypher text and vice-versa. The outcomes of the selected alternative encryption schemes are compared and analysed.

5.1 Encryption and Decryption Time

Tables 2 and 3 shows encryption and decryption time of DES, 3DES, AES, Blowfish, RSA and ECIES respectively.

Figure 3 shows the encryption time for different cryptographic algorithms. It is depicted from the figure that blowfish takes less time for encrypting the data i.e., it is faster. While, RSA takes more time for encrypting data.

Figure 4 shows decryption time for different cryptographic algorithms. It can be seen that encryption time for all algorithms is more than decryption time. Among all, decryption time for blowfish is less while for RSA decryption time is more. From the experimental analysis it can be seen that Blowfish is best choice in case of encryption and decryption time as it records shortest time among all algorithms. AES can be used in case if you need more confidentiality and integrity. Table 1 shows that ECIES uses small key sizes, it is fast and has low power consumption. In our work we are using SLAM (Simultaneous Localization and Mapping) for building maps. As robots has low processing power and memory, we will be using ECIES for encrypting the data.

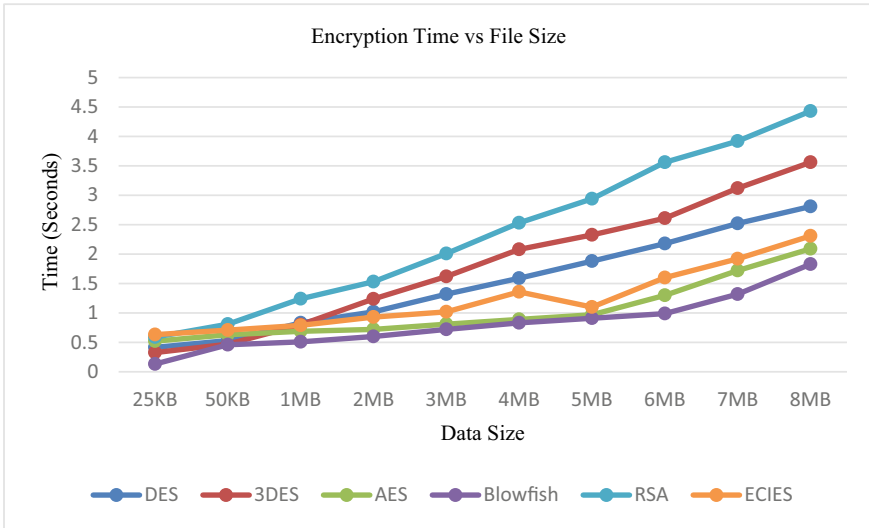


Fig. 3 Encryption time vs. file size

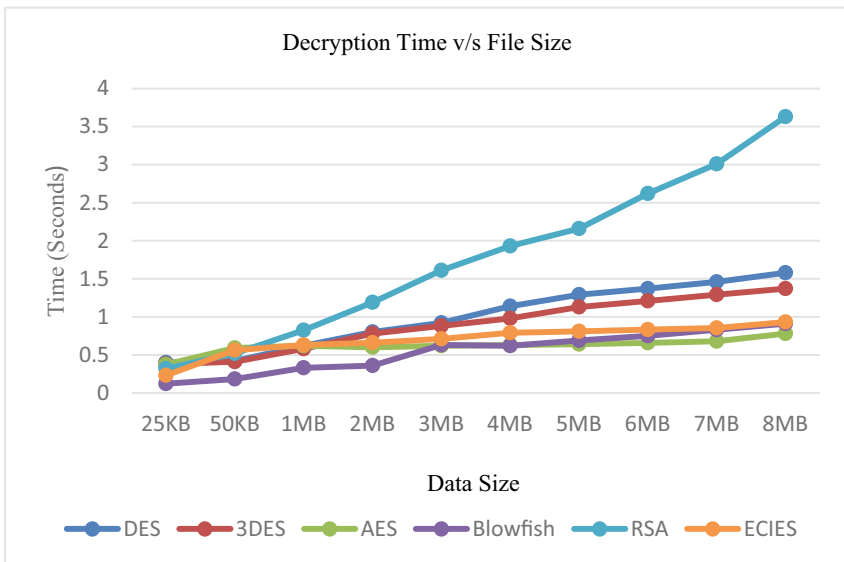


Fig. 4 Decryption time vs. file size

6 Conclusion and Future Scope

The fair comparisons between six regularly used algorithms and the simulated were offered in this research. For performance evaluation, the encryption methods AES, DES, 3DES, ECIES, Blowfish and RSA were chosen. In this paper, we have implemented different cryptographic algorithm in cloud-based multi-robot systems. The result shows that ECIES performs better in terms of encryption and decryption time compared to other algorithms in cloud robotics environment. The results of the comparative investigation revealed the capabilities of each algorithm. In the future, we will try to implement and test in physical cloud robotics system. Also, will try to search and test the lightweight cryptographic algorithms in cloud robotics network.

References

1. Putra SD, Yudhiprawira M, Sutikno S, Kurniawan Y, Ahmad AS (2019) Power analysis attack against encryption devices: a comprehensive analysis of AES, DES, and BC3. *TELKOMNIKA Telecommun. Comput. Electron. Control.* 17(3):1282–1289. <https://doi.org/10.12928/TELKOMNIKA.V17I3.9384>
2. Jain S, Doriya R (2022) Security framework to healthcare robots for secure sharing of healthcare data from cloud. *Int J Inf Technol* 1–11. <https://doi.org/10.1007/s41870-022-00997-8>
3. Hess E, Janssen N, Meyer B, Schütze T (2000) Information leakage attacks against smart card implementations of cryptographic algorithms and countermeasures—a survey. *Eurosmart Secur. Conf.* 130:55–64
4. Maniyath SR, Thanikaiselvan V (2020) An efficient image encryption using deep neural network and chaotic map. *Microprocess Microsyst* 77:103134. <https://doi.org/10.1016/j.micpro.2020.103134>
5. Chandra S, Paira S, Alam SS, Sanyal G (2014) A comparative survey of symmetric and asymmetric key cryptography. In: 2014 international conference on electronics and communication computer engineering, ICECCE 2014, pp 83–93. <https://doi.org/10.1109/ICECCE.2014.7086640>
6. Sharma DK, Singh NC, Noola DA, Doss AN, Sivakumar J (2021) A review on various cryptographic techniques & algorithms. *Mater Today Proc* 51:104–109. <https://doi.org/10.1016/j.matpr.2021.04.583>
7. Mutnuru S, Sah SK, Kumar SYP (2020) Selective encryption of image by number maze technique. *Int J Cryptogr Inf Secur* 10(2):1–10. <https://doi.org/10.5121/ijcis.2020.10201>
8. Jain S, Doriya R (2019) Security issues and solutions in cloud robotics: a survey. In: *Communication on Computer Information Science CCIS*, vol 922, pp 64–76. https://doi.org/10.1007/978-981-15-1718-1_6
9. Bhardwaj A, Som S (2016) Study of different cryptographic technique and challenges in future. In: 2016 1st International conference on innovation challenges cyber security, ICICCS 2016, no Iciccs, pp 208–212. <https://doi.org/10.1109/ICICCS.2016.7542353>
10. Hershey JE (1980) Implementation of mitre public key cryptographic system. *Electron Lett* 16(24):930–931. <https://doi.org/10.1049/el:19800663>
11. Chakraborty M, Jana B, Mandal T (2018) A secure cloud computing authentication using cryptography. In: 2018 international conference on emerging trends innovation engineering technology research, ICETIETR 2018, pp 1–4. <https://doi.org/10.1109/ICETIETR.2018.8529100>
12. Ramesh G (2012) A comparative study of six most common symmetric encryption algorithms across different platforms. *Int J Comput Appl* 46(13):6–9

13. Subbiah D (2021) Tribrid secure encryption technique to protect the data in the cloud
14. Patnaik S, Sunil A, Reddy R, Scholar M (2021) Hybrid cryptography algorithm for secure file storage in the cloud, XIV(X):25–29
15. Seth SM, Mishra R (2011) Comparative analysis of encryption algorithms for data communication. IJCST 2(2):292–294
16. Elminaam DSA, Kader HMA, Hadhoud MM (2010) Evaluating the performance of symmetric encryption algorithms. Int J Netw Secur 10(3):213–219
17. Xiao S, Yu ZJ, Deng YS (2020) Design and analysis of a novel chaos-based image encryption algorithm via switch control mechanism. Secur. Commun. Networks 2020:30–32. <https://doi.org/10.1155/2020/7913061>
18. Janakiraman S, Thenmozhi K, Rayappan JBB, Amirtharajan R (2018) Lightweight chaotic image encryption algorithm for real-time embedded system: implementation and analysis on 32-bit microcontroller. Microprocess Microsyst 56:1–12. <https://doi.org/10.1016/j.micpro.2017.10.013>
19. Ramesh A, Suruliandi A (2013) Performance analysis of encryption algorithms for information security. In: Proceedings of IEEE international conference on circuit, power computer technology, ICCPCT 2013, pp. 840–844, 2013, doi: <https://doi.org/10.1109/ICCPCT.2013.6528957>.
20. Bonde SY, Bhadade US (2018) Analysis of encryption algorithms (RSA, SRNN and 2 Key Pair) for information security. In: 2017 international conference on computer communication control automation, ICCUBEA 2017, pp 1–5. <https://doi.org/10.1109/ICCUBEA.2017.8463720>
21. Anas M, Imam R, Anwer F (2022) Elliptic curve cryptography in cloud security: a survey, pp 112–117. <https://doi.org/10.1109/confluence52989.2022.9734138>
22. Gurav YB, Patil BM (2022) Two-fold improved poor rich optimization algorithm based decentralized information flow control for cloud virtual machines: an algorithmic analysis, pp 417–425. <https://doi.org/10.1109/icssit53264.2022.9716462>
23. Chinnasamy P, Deepalakshmi P (2022) HCAC-EHR: hybrid cryptographic access control for secure EHR retrieval in healthcare cloud. J Ambient Intell Humaniz Comput 13(2):1001–1019. <https://doi.org/10.1007/s12652-021-02942-2>
24. Gabsi S, Kortli Y, Berouille V, Kieffer Y, Alasiry A, Hamdi B (2021) Novel ECC-Based RFID Mutual Authentication Protocol for Emerging IoT Applications. IEEE Access 9:130895–130913. <https://doi.org/10.1109/ACCESS.2021.3112554>
25. Gayoso Martínez V, Hernández Álvarez F, Hernández Encinas L, Sánchez Ávila C (2011) Analysis of ECIES and other cryptosystems based on elliptic curves

An IoT Based Environment Monitoring & Controlling System for Food Grain Warehouse



Vrinda Parkhi, Nishant Chavhan, Sharda Chandak, Bhushan Chaware, and Prasad Bongarde

Abstract The Global Hunger Index (GHI) -2021 puts India 100th out of 116 nations worldwide. Despite having enough food to serve its whole population, India is unable to provide an appropriate and consistent supply of meals to millions of its own people. Food grain waste is one of the factors contributing to this scenario. It also represents indirect waste of the water and electric power required to transform food grain into product via food processing enterprises, as well as energy and man-power used in agricultural operations. Furthermore, excessive food waste may rapidly undermine any country's economy. To conserve each grain of food, a warehouse monitoring system should be implemented to maintain food safety by reducing harvest loss and ensuring that it reaches needy people. This article describes the design and implementation of an IoT-based environment monitoring and controlling system for a food grain warehouse by employing different sensors to detect and monitor different parameters such as humidity, temperature, smoke, flame, and dangerous gases such as ammonia and methane. As a consequence, the appropriate environmental conditions will be provided to ensure the quality of food grains, and the information will be communicated to the authorities for further examination.

Keywords Internet of Things (IoT) · Food grain warehouse · Sensors · Methane · MQ-2 · MQ-4 · MQ-135 · ESP 8266

1 Introduction

India is the Sixth largest economy in the world and soon expected to reach \$5 trillion economy by the year 2025 but the results of the hunger statistics of our country are not encouraging. Food grain squandering is a predominant problem that immediately

V. Parkhi · N. Chavhan (✉) · S. Chandak · B. Chaware · P. Bongarde
Vishwakarma Institute of Technology, Pune 411037, India
e-mail: nishant.chavhan20@vit.edu

V. Parkhi
e-mail: vrinda.parkhi@vit.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
P. P. Joby et al. (eds.), *IoT Based Control Networks and Intelligent Systems*,
Lecture Notes in Networks and Systems 528,
https://doi.org/10.1007/978-981-19-5845-8_16

217

Global Hunger Index: India & Its Neighbours

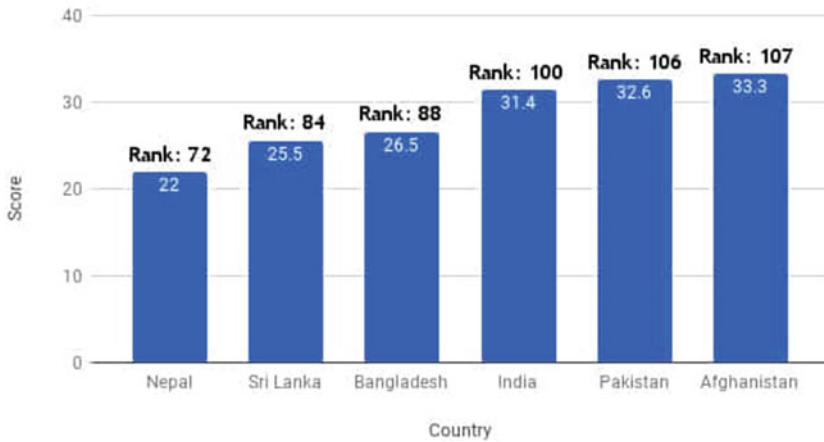


Fig. 1 Global hunger index [17]

requires an attention. According to the statistics issued by United Nations development program, in India up to 40% of food produced is wasted due to lack of proper storage facilities [16].

According to the statistics shown in Fig. 1 [17], India stands 100th place for safeguarding its population’s food safety across the world. Recent studies state that nearly 45.9% of children under five years of age are to be observed delayed growth [1].

This research work addresses the issue of food grain wastage in the warehouses. A novel environment monitoring and control system is proposed by utilizing various sensors to measure different levels of environmental parameters and control it with appropriate response actuators. This will help in providing a suitable environment for ensuring the quality of food grains and communicate the sensed information to the concerned authority.

Parameters such as temperature, humidity, atmospheric gases specifically methane gas as well as hazardous gases and fire flame are sensed by different sensors and act as indicators, which are visualized on both app and web server. By replacing or upgrading traditional warehousing system with the proposed IoT based Warehousing System, a huge amount of food grains can be saved. The other advantage of the proposed system is that, being a generalized system, it can further be enhanced to universal warehouses that store perishable goods other than food grains.

2 Literature Review

IoT based warehouse storage system mainly includes monitoring and controlling various parameters that affect the storage of food grain. Different sensors such as temperature sensor, humidity sensor, smoke sensor, fire sensor, and many more had been preceded by other researchers to establish a legitimate warehousing system. In addition to that, different hardware systems and software platforms are also proposed. In past, several researchers have worked on developing an efficient warehouse monitoring system, which is discussed further:

In 2013, S. Kaushik et al. [2], implemented ZigBee and Bluetooth based monitoring and controlling of food storage system. Sensor data received from sensor unit and authenticated using PIC microcontroller is stored in the database.

V. S. Suryawanshi et al. [3], have developed food grain storage monitoring and controlling system. The proposed system comprises of humidity and temperature sensors, which sense the parameter values and then update it to a real-time administrator dashboard through wireless device. Sensed data is being communicated to the personnel, who monitor the system by considering several environmental factors such as temperature, humidity, ammonia gas, etc. which affects the quality of food grain. An alarm signal is being generated and further an alerting message will also be sent to the registered system via a GSM module. The proposed storage model helps in the detection of environmental parameters in order to maintain quality of food grain.

Kamali, Kavya and Ramyadevi [4], have designed and prototyped a sensor-based monitoring system for warehouse maintenance. Inputs acquired by different sensors i.e., temperature, MQ-7 (Air quality measurement), humidity, motion, etc. are sent to Zigbee and then converted into digital signals using ADC followed by comparing the predefined parameters data values via algorithm. In addition to that, the system raises emergency alarm whenever it is required.

A. Srivastava et al. [5], introduced an assistive technology to protect food by proposing a food monitoring system. Input acquired by different sensors i.e., temperature, humidity, light and moisture is analyzed by a Raspberry Pi2 module and visualized on Tkinter GUI interface to send an alert message in emergency situation.

T. N. A. Kumar et al. [6], presented a system to control several environmental parameters inside the warehouse in order to monitor and control a warehousing system. The developed warehouse system is equipped with various sensors like temperature, humidity, smoke, load cell and LDR sensor analyzed by Renesas microcontroller and GSM in order to update the values to the web server.

Akila and Shalini [7], proposed an automated and mechanized system to control and monitor the quality of the food grains, which is supposed to be degrading via temperature, humidity and ammonia gas sensor, analyzed by Arduino Due R3 board and Wi-Fi transmitter in order to administer about how fresh the food grains are, to the concerned authority.

Ravi Kishore Kodali et al. [8], introduced a smart monitoring system for farm, processing plant, warehouse and market with the help of ESP8266 as a microcontroller. Presented system is equipped with temperature, humidity and MQ-135 gas sensor, in order to abate the levels of temperature and humidity in accordance with IoT based fans and cooling unit. Sensed data is being visualized on Node-RED as visual programming dashboard and on Amazon-Web Service (AWS) as cloud computing platform via GSM GPRS Wi-Fi transmitter. Concerned authority is able to monitor the mold/insect infestation so that immediate and possible action should be taken care of.

H. Nigam et al. [9, 12], proposed an aiding system which enables monitoring of indoor environmental framework mainly air. The proposed system in the paper majorly concerned upon monitoring of Air Quality Index (AQI) which comprises of CO, CO₂ and other environmental components of air. Node-Red Dashboard is being used in order to have a real time updated status for monitoring of AQI and hazardous gas level.

S. S. Sruthi et al. [10], proposed a system comprising of various sensors like Luminosity, Fire/Smoke, Door Status sensors, etc. to have monitoring system for cold storage which will be communicated to the user. Data sensed by sensors is visualized and stored on Things-speak web server. WAMP is used to host the website on the local machine.

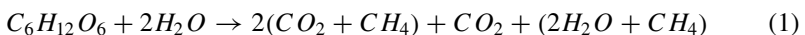
3 Methodology

3.1 System Architecture

The proposed system can be easily implemented at any food grain storage facility by changing the sensor parameters i.e., predefined thresholds as per the requirements of the environment.

In any food grain storage warehouse, humidity i.e., moisture content and temperature will be considered as the crucial factors, which directly contribute to the quality of food grain stored in warehouse.

If the warehouse is kept closed for an extended period of time with insufficient ventilation, several problems will arise, including rotting and degradation of food grains, as well as mold/insect infestation. Glucose in presence of moisture i.e., water molecules lead to fermentation and methanogenesis of glucose, where reacts to form Carbon dioxide (CO₂), Water (H₂O) and Methane (CH₄) and result in degrading the quality of food grain [11, 13]. During this reaction, H₂ is oxidized to H⁺ and CO₂ is reduced to CH₄. Overall reaction for fermentation and methanogenesis can be defined using Eq. (1) as:



Such problem can be conquered by controlling and maintaining favorable environment inside the warehouse. Further, it can be facilitated by scheduled visits of maintenance personnel to the warehouse. IoT assisted system with human presence makes this achievable.

Figure 2 shows block diagram of proposed system. The inputs acquired by the various sensors (DHT11, MQ-2, MQ-4, MQ-135 and IR Flame sensor) are then compared simultaneously with a predefined threshold data value, which is already embedded into the Arduino-code. The threshold values have been set by considering the required parameter values for developing an environment favorable to food grains [15]. If input values acquired by the sensor exceed a predefined threshold, then alarm signal is being generated via buzzer which is interfaced with the Node MCU ESP8266 Wi-Fi module as micro-controller in the proposed hardware assembly. Along with that, if the temperature value or humidity value or concentration of harmful gases or smoke exceeds predefined threshold value, then Brushless Fan and DC Motor (First one act as exhaust and second one as Ventilation) and Servo motor (Opening the warehouse window) will start operating as per the real time condition. In order to monitor remotely, all the sensed parameters are also communicated and authenticated to a remote server for monitoring and controlling. Hence proposed system enables bipartite advantage. Proposed system is able to monitor and control, if the in-person monitoring of sensed parameters via remote servers is not available in that case, still alarm signal is raised and immediate action to be initiated in order to de-escalate unnecessary wastage of food grain locally at warehouse.

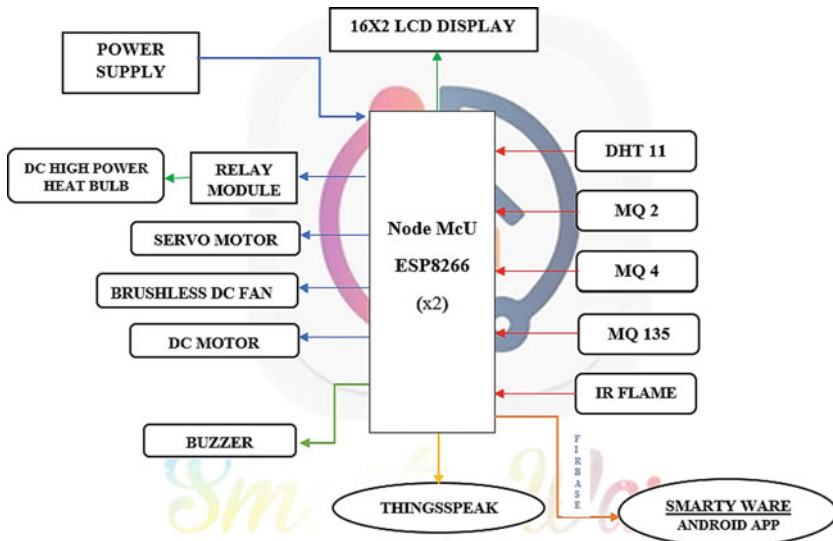


Fig. 2 Block diagram for proposed system

3.2 System Components

Multiple sensor and Developmental boards are listed as:

ESP8266 Node MCU. It is low-cost, self-contained Wi-Fi networking solution for continuous Internet connectivity. Sensed data in the proposed system will be available and updated on Things-speak web server and real time database- Firebase.

DHT11 Module Humidity Sensor. The temperature-humidity sensor known as DHT11 detects and controls temperature and humidity degrees in a single distinctive configuration. The sensor promises outstanding flexibility and excellent long-term reliability.

MQ-2 Smoke Sensor. MQ-2 is used to detect the concentration of smoke present in air. Hence it is known as Smoke Sensor. It usually senses the smoke by using the output voltage level. Greater the concentration of smoke, greater will be the voltage that it outputs and vice a versa. MQ-2 is wired to Node MCU to analyze the data it sends and accordingly buzzer is activated, alerting a respected authority for appropriate actions.

MQ-4 Methane Gas Sensor. MQ-4 is used in gas leakage detecting equipment used for home and industry, is suitable for detecting CH_4 , Natural gas. It can sense the noise of alcohol and cooking fumes and cigarette smoke.

MQ-135 Harmful Gas Sensor. MQ-135 sensor is used as it detects gasses NH_3 , NO_x , Alcohol, Benzene, Smoke, CO etc. As MQ-135 comprises of lower conductivity to clean air. As a result of it, conductivity increases with increase in concentration of harmful or hazardous gases present in air at that moment.

IR Flame Sensor. Flame sensor is used as it detects the presence of a fire source by enabling alarm in that response. It is interfaced with proposed system in order de-escalate the threat of fire.

4 Results and Discussions

Hardware integration using different sensors and it's testing is presented in this section. Figure 3 shows complete wiring of the system with its sensors, actuators and display.

When the system is powered and controller is embedded with the code, the parameter values under normal condition are observed as shown in Fig. 4. System is then exposed to a condition where parameter values cross its threshold. The display shows these values and driving of actuators is also observed.

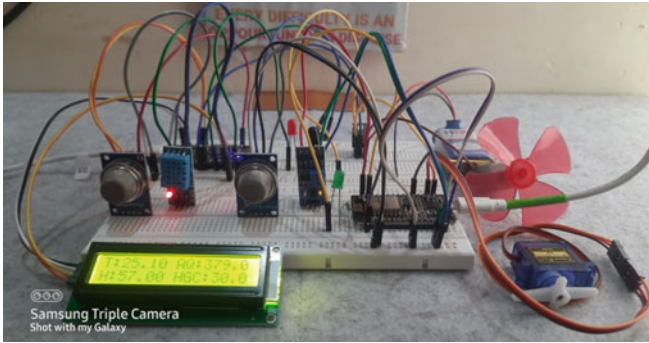


Fig. 3 Actual circuit for proposed system

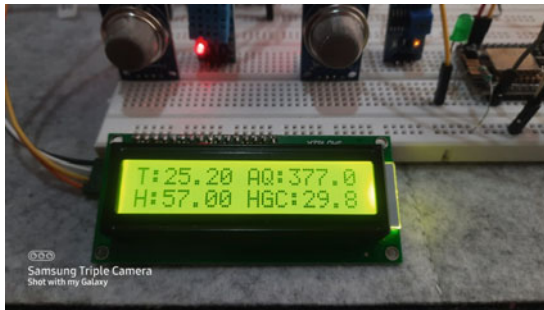


Fig. 4 Displaying several parameters sensed on LCD 16 x 2

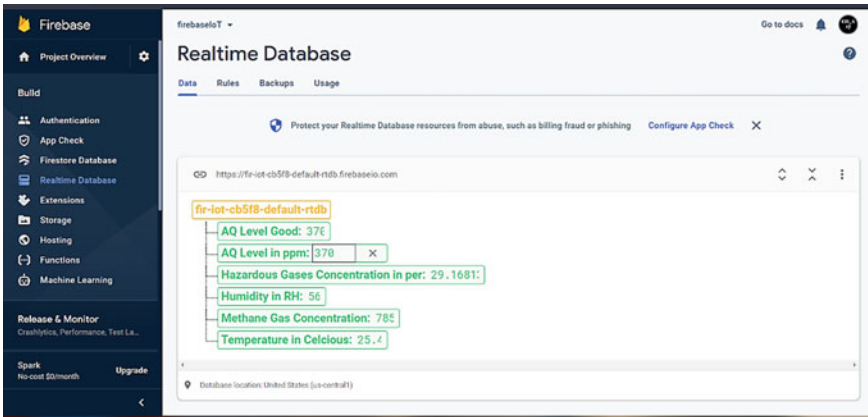


Fig. 5 Sensed realtime data on firebase used to communicate between Android application and proposed system



Fig. 6 Sensed data visualization on android application

Figure 5 shows the data being communicated from Node MCU to Firebase as Realtime database by embedding the Web API Key in the Arduino code and Android studio code, which is then further visualized over app.

The data communicated through Firebase is visualized on the android app named Smarty Ware as shown in Fig. 6. Air quality Level and Methane gas concentration is measured in ppm (parts per million) while Hazardous gases concentration is observed in percentage. Same sensor data is also visualized on Things-Speak Web server (Fig. 7) by embedding Web server Channel credentials in the code.

The system performance has been evaluated using performance parameters like the sensitivity of a sensors and reliability of the system, with latency of 1–2 s for establishing communication between Node MCU and Mobile application via Firebase server.

The proposed system is primarily based on client-side interface which will enable them only to visualize the sensed parameter remotely but not controlling several parameters remotely via server or admin interface. So far, proposed system is able to controlled sensed parameters autonomously in accordance with pre-programmed Arduino code embedded in system. Ammonia gas concentration cannot readily be sensed by proposed system. Till date, proposed system is not able to control quantitative and qualitative degradation of the food grains during storage via insect-pest, micro-organism, rodents and birds which limits food grain storage capacity in warehouse.

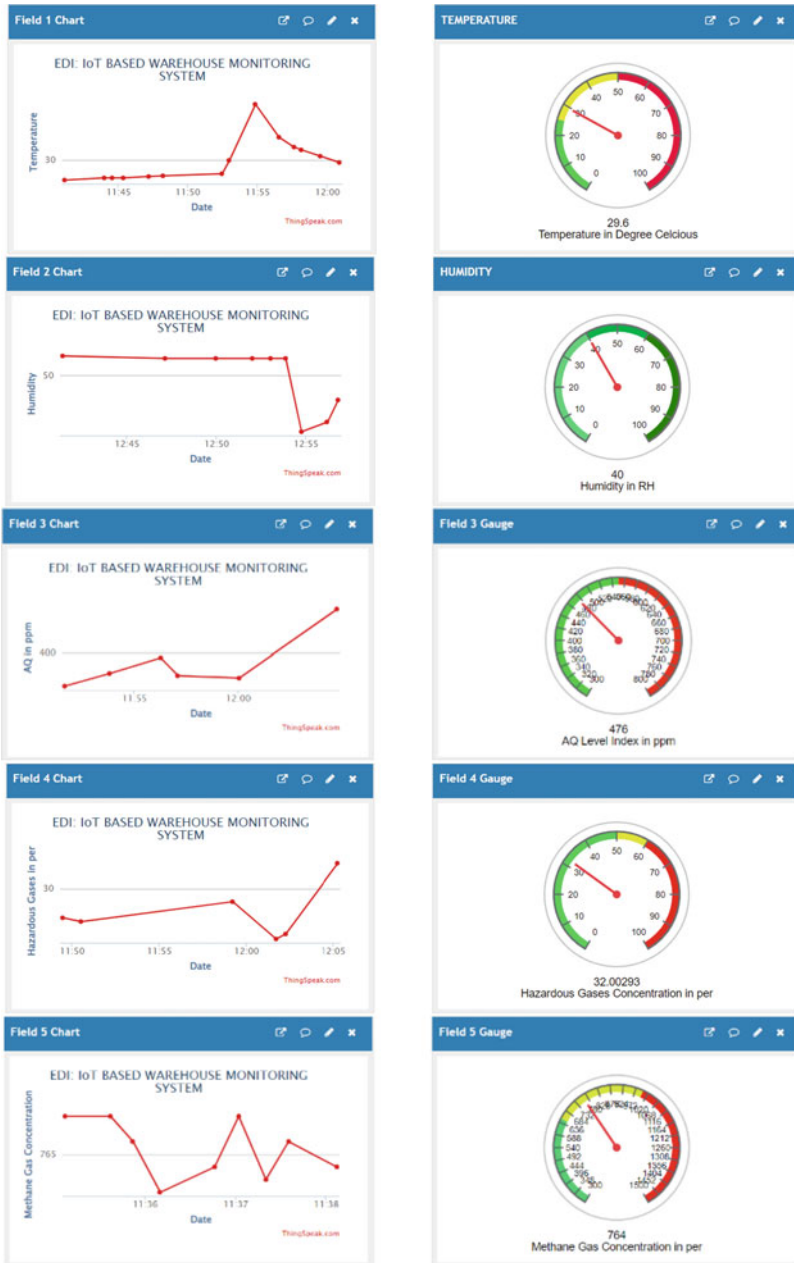


Fig. 7 Sensed data visualization on Things-Speak Web server

5 Conclusion

This proposed IoT based Warehouse system demonstrates the design and implementation of monitoring and controlling system used in the smart warehouse system. It will not only ensure a monitoring system, but also proposes a low-cost solution for maintenance or controlling of indoor warehouse environment with the assistance of IoT (Internet of Things). The proposed system elevates the performance of the warehouse management system in order to avoid food shortage and economic loss incurred thereby.

As IoT has crossed the threshold in every sector of advancement associated with science and technology, this technology will be easily implemented in agriculture field and accessible to farmers. It can further be enhanced with the upgradation of various sensors with high sensitivity, adaptability, and reliability, which can be helpful to modernize the warehouse even with a very efficient and affordable cost for farmers. Proposed system can also be equipped by enabling monitoring and cautioning system to the concerned authority for mold/insect infestation, so that immediate action such as pest control can be considered. The system can be more fruitful with the implementation of machine learning techniques by incorporating various neural networks [14].

Acknowledgements We would like to thank our mentor, for timely guidance and assistance during the course of this project. We are also thankful to our Institute and the Department of Multidisciplinary Engineering for giving this opportunity and freedom to work on this project.

References

1. Murarkar S, Gothankar J, Doke P, Pore P, Lalwani S, Dhumale G, Quraishi S, Patil R, Waghachavare V, Dhobale R, Rasote K, Palkar S, Malshe N (2020) Prevalence and determinants of undernutrition among under-five children residing in urban slums and rural area Maharashtra India: a community-based cross-sectional study. *BMC Public Health* 20(1):1559
2. Kaushik S, Singh C (2013) Monitoring and controlling in food storage system using wireless sensor networks based on zigbee & bluetooth modules. *Int J Multidisc Cryptol Inf Secur* 2(3):7–10
3. Suryawanshi VS, Kumbhar MS (2014) Real time monitoring & controlling system for food grain storage. *Int J Innov Res Sci Eng Technol* 3:734–738
4. Muthukumar S, Mary WS, Kamali K, Kavya S, Ramyadevi G (2018) Sensor based warehouse monitoring and control. In: *IEEE conference record # 42487*
5. Srivastava A, Gulati A (2016) iTrack: IoT framework for smart food monitoring system. *Int J Comput Appl* 148(12):1–4
6. Kumar TNA, Lalswamy B, Raghavendra Y, Usharani SG, Usharani S (2018) Intelligent food and grain storage management system for the warehouse and cold storage. *Int J Res Eng Sci Manag* 1(4):130–132
7. Akila A, Shalini P (2018) Food grain storage management system. *Int J Eng Technol* 7(231):170–173
8. Kodali RK, John J, Boppana L (2020) IoT monitoring system for grain storage. *IEEE Xplore*

9. Nigam H, Saini AK, Banerjee S, Kumar A (2017) Indoor environment air quality monitoring and its notification to building occupants. In: TENCON 2019, Kochi, India, 17–20 October 2019
10. Sruthi SS, Yasasweni DR, Swathi JN (2017) Cold storage traceability system. *Int J Res Appl Sci Eng Technol* 5:1086–1096
11. Lyu Z, Shao N, Akinyemi T, Whitman WB (2018) Methanogenesis. *Curr Biol* 28(13):R727–R732
12. Banerjee S, Saini AK, Nigam H, Vijay V (2020) IoT instrumented food and grain warehouse traceability system for farmers. In: International conference on artificial intelligence and signal processing (AISP)
13. Mahadevaswamy UB, Megha KM, Srivatsa AV, Jain AG, Rani BB (2021) Food grain storage monitoring system. *Int J Eng Res Technol (IJERT)*. ISSN: 2278–0181
14. Chen Joy Iong-Zong, Yeh Lu-Tsou (2021) Greenhouse protection against frost conditions in smart farming using IoT enabled artificial neural networks. *J Electron Inf* 2(4):228–232
15. Nanda SK, Vishwakarma RK, Bathla HVL, Rai A, Chandra P (2015) Harvest and post losses of major crops and livestock produce in India. Ludhiana: ICAR-All India Coordinated Research Project on Post-Harvest Technology
16. Food Wastage In India, And What You Can Do About It - CSR Journal (2018). <https://thecsrjournal.in/food-wastage-in-india-a-serious-concern/>. Accessed 29 May 2022
17. European NGOs of Concern Worldwide and Welthungerhilfe - Hindustan Times. <https://www.hindustantimes.com/india-news/india-worse-than-nepal-bangladesh-in-tackling-hunger-down-3-ranks-at-global-level-report/story-a2SyCOUEiNwRgMMI4cqZYN.html?> Accessed 02 Apr 2022

Numerical Analysis and ANN Modeling of the Intercooled, Reheat and Regenerative Gas Turbine Cycle



Milind S. Patil, Shyamkumar D. Kalpande, Sanjay P. Shekhawat, and Chandrashekhar D. Mohod

Abstract Continuous rise in population is the cause for increase in demand factor of the power plant. However, running of thermal power plant results into the release of objectionable pollutants. One way to have a control on these pollutants is the use of renewable energy sources like solar and wind. However, these sources are intermittently available and hence it is important that the quick backup and balancing is needed when renewable source is not available. In these situations, use of gas turbine is a good option. This paper considers the numerical analysis of the combined intercooled, reheat and regenerative cycle. Objective of this work was to develop ANN model, understand the various parameters, their effect on gas turbine performance and optimize the cycle parameters for maximum efficiency. Analysis shows that for turbine inlet temperature 1200 K and compressor inlet temperature 293 K the optimum pressure ratio is 9 and thermal efficiency is 32%. Analysis of the gas turbine cycle was performed for different pressure ratio, turbine inlet temperature, compressor inlet temperature, and heat exchanger effectiveness. Data obtained from the analysis was then used for the development of the prediction model using artificial neural network. The developed model has root mean square error equal to 0.00018 and the regression coefficient of the trained network is 0.99. This way the developed mathematical model was validated, and it has a good predictability.

Keywords Brayton cycle · Reheating · Regeneration · Neural network

1 Introduction

Day by day living standard of the people is increasing and also there is a rapid increase in the world population. This has affected energy consumption rate and the

M. S. Patil (✉) · S. D. Kalpande · C. D. Mohod
Guru Gobind Singh College of Engineering and Research Centre, Nashik, Maharashtra, India
e-mail: mspiso2012@yahoo.com

S. P. Shekhawat
Shram Sadhana Bombay Trusts College of Engineering and Technology, Jalgaon, Maharashtra, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
P. P. Joby et al. (eds.), *IoT Based Control Networks and Intelligent Systems*,
Lecture Notes in Networks and Systems 528,
https://doi.org/10.1007/978-981-19-5845-8_17

229

demand factor is increasing continuously. Power plants are struggling to manage the energy balance and peak load requirements. With increasing demand, the fossil fuel sources are depleting rapidly and in the twenty-first century 85% of energy is from the fossil fuels only [1]. In the power generation sectors gas turbine is the better choice because of their low CO₂, Nox emission, higher efficiency and smaller period of installation [2]. In the year 1701 John Barber invented the first gas turbine where he used a reciprocating type of compressor [3] and first simple gas turbine cycle was introduced in the year 1903 with turbine inlet temperature (TIT) 400 °C [4, 5]. Materials that can withstand with high temperature are then available in late 1950 and then in the year 1961, 18 MW turbine was developed with TIT 788 °C and thermal efficiency 25.4% [6]. Westinghouse, Mitsubishi Heavy Industries and Fiat Avio jointly developed the F class combined cycle gas turbine that has 1260 °C TIT and thermal efficiency 51.7% [6]. Rao et al. reported hybrid system IC-GT with high pressure Oxide Fuel Cell (SOFC) and SOFC with HAT (Humid Air Turbine) that has an efficiency of the 75% [7]. Use of regeneration i.e., heating the air before combustion is employed for improving gas turbine thermal efficiency. Operating gas turbine exit temperature is higher compare with the temperature of the air after compression. In a regenerative cycle exhaust gas, leaving the turbine heats air coming from the compressor. Temperature of the heated air depends on the effectiveness of the heat exchanger. Regenerator with high effectiveness has more potential of heat energy saving during the combustion process. Generally during the regeneration, the pressure of air is drop by 2% and due to heat loss and thermal resistance the effectiveness is less than 1 around 0.85 [8]. For higher regeneration pressure ratio and maximum gas temperature effectiveness of heat exchanger will be high. Another way to improve the gas turbine efficiency is increasing the specific output by expansion in two stages and the intermediate heating known as reheating. Reheating increases the heat energy required and reduces the thermal efficiency. However, if the reheat pressure ratio is maintained as 0.2 to 0.22 higher efficiency will be achieved. In reheating process maximum cycle temperature increases that and hence reduction in efficiency may not be severe. Due to increase in cycle maximum temperature and high exhaust gas temperature regeneration process can be employed in combination with reheat cycle and efficiency loss can be recovered. During compression of the air in the compressor its temperature increases and hence the specific volume increases. This will cause the higher work of compression. Employing a heat exchanger also known as intercooler results in the reduction in intermediate temperature and hence the work of compression [9–11].

Many thermal analysis of gas turbine cycle major assumptions are taken, and the effect of various losses and kinetic energy was neglected. Also, there were no other research reported that considers the artificial neural network modeling for predicting the performance of gas turbine. Hence the objective of this paper is to consider the following points and analyze the performance of gas turbine.

1. Due to higher fluid velocities kinetic energy changes are considered.
2. Pressure loss due to fluid friction and loss of energy is consider for analysis of combustion chamber, inter-cooler and regenerator.

3. Heat exchangers could not have 100% effectiveness. Also, for a very high effectiveness tube length of heat exchanger increases that increases cost. Hence regenerator effectiveness is considered for better economy.
4. Transmission efficiency is considered to accommodate the power consumed by the auxiliary systems, bearing friction and windage friction in losses.
5. Variable specific heat is considered as it is the function of the temperature and chemical composition.
6. Air fuel ratio and combustion efficiency is considered for the correctness of the thermal analysis.

Results of this thermal model was then used to develop the artificial neural network (ANN) for predicting the performance of gas turbine. This ANN model can used as a reference for understanding the performance without much loss of time also serves a good learning tool.

2 Cycle Description and Thermal Model

2.1 Thermodynamic Cycle

Figure 1 represents the layout the reheat and regeneration gas turbine plant. Two-stage compressor with an intercooler is coupled with two-stage turbine with a re-heater. Part of the turbine power is used to drive the compressor and remaining is available as output power. Air is compressed first in LP compressor to an intermediate pressure which is equal to the minimum work required for the compression assuming the perfect inter-cooling. In a perfect inter-cooling process, an intercooler is installed between the stage of the compressor and it has effectiveness equal to 1 and air from the low pressure compressor is cooled at constant pressure to its initial intake temperature. Air is then compressed in second stage compressor. Isentropic efficiency of compressor is taken as 0.85 and pressure drop of 0.1 bar is considered in the intercooler. A regenerator is a heat exchanger used between the compressor and high-pressure combustion chamber. Air is heated in the regenerator using exhaust gas from the low-pressure turbine. Heat energy is then added at constant pressure in the combustion chamber to a maximum temperature considering the metallurgical limits. After adding heat energy in combustion chamber, it is expanded in the high-pressure turbine to an intermediate pressure. After expansion in high pressure turbine air is reheated to a maximum temperature of 1000 K in the re-heater (low pressure combustion chamber). This heated air is then expanded in the second stage low pressure turbine. For this analysis isentropic efficiency of the turbine is taken as 0.85, combustion efficiency as 0.8 and a pressure drop of 0.1 bar is considered. An engineering equation solver program was used for the numerical analysis of this represented system.

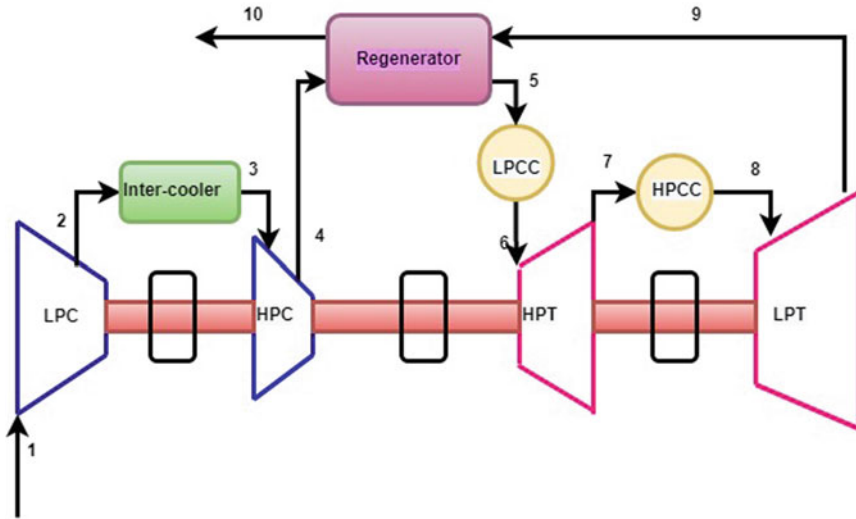


Fig. 1 Thermodynamic cycle

Assumptions made for the analysis are as explained below.

1. Two stage expansion with reheating of gas to maximum temperature of 1000 K
2. For air specific heat ratio is taken as 1.4 and for gas it is 1.33
3. Velocity of air at compressor inlet is taken as constant and effect of KE is considered during the analysis
4. Specific energies (energy per unit mass) are considered for the analysis
5. Atmospheric pressure is taken as 100 kPa
6. Pressure drop in the re-heater, intercooler, regenerator and combustion chamber is taken as 100 kPa
7. Compression and expansion processes are irreversible with an isentropic efficiency of 0.8
8. Inter-cooling is perfect
9. Reheat pressure ratio is constant
10. Fuel has a heating value of 40 MJ/kg (Normally the fuel used has a heating value of 40 to 44 MJ/kg, For this analysis lower heating value is considered).

2.2 Thermal Model

Specific heat of the air varies with the temperature, and it is calculate by fitting the equation from the data set represented by Chappel, M. S., and Cockshutt [12]. Equation 1 represents the specific heat of air for temperature range of 200 to 800 K and Eq. 2 is used for a temperature range of 800 to 2200 K.

$$c_{pa} = (1.0189 \times 10^3) - (0.13784 \times T_a) + (1.9843 \times 10^{-4} T_a^2) + (4.2399 \times 10^{-7} T_a^3) - (3.7632 \times 10^{-10} T_a^4) \quad (1)$$

$$c_{pa} = (7.9865 \times 10^2) + (0.5339 \times T_a) - (2.2882 \times 10^{-4} T_a^2) + (3.7421 \times 10^{-8} T_a^3) \quad (2)$$

Specific heat of a gas is calculated according to the Eqs. 3 and 4 as

$$c_{pg} = c_{pa} + \left[\frac{f}{1+f} \right] B_T \quad (3)$$

For a temperature range of 200 to 800 K the factor B_T is calculated according to Eq. 4 and 800 to 2000 K Eq. 5 was used.

$$B_T = (-3.59494 \times 10^2) + (4.5164 T_g) + (2.8116 \times 10^{-3} T_g^2) - (2.1709 \times 10^{-5} T_g^3) + (2.8689 \times 10^{-8} T_g^4) - (1.2263 \times 10^{-11} T_g^5) \quad (4)$$

$$B_T = (1.0888 \times 10^3) - (0.1416 T_g) + (1.196 \times 10^{-3} T_g^2) - (1.2401 \times 10^{-6} T_g^3) + (3.0669 \times 10^{-10} T_g^4) - (2.6117 \times 10^{-14} T_g^5) \quad (5)$$

Since gas turbine uses rotary compressor velocity of the air at inlet to the compressor is high and hence changes in kinetic energy and the stagnation temperature are calculated using equation

$$T_1 = T_a + \left(\frac{C_1^2}{2c_{pa}} \right) \quad (6)$$

Air is cooled between the stages of the low pressure and high-pressure compressor. Inter-cooling is assumed as perfect and the intermediate pressure of the two-stage compression corresponds to requirement of minimum work of compression. Process of compression is irreversible and Eqs. 7 and 8 are used for isentropic efficiency.

$$\eta_{LP\text{compressor}} = \frac{T_2' - T_1}{T_2 - T_1} \quad (7)$$

$$\eta_{HP\text{compressor}} = \frac{T_4' - T_3}{T_4 - T_3} \quad (8)$$

Compressor exit temperature T_2 is calculated by the Eq. 9

$$T_2 = T_1 + \left(\frac{T_1}{\eta_c} \right) \left[\left(\frac{P_2}{P_1} \right)^{\frac{\gamma_a - 1}{\gamma_a}} \right] \quad (9)$$

The turbine and compressors are mounted on the common shaft. Part of the work is used to drive the compressor and the remaining is the output power of the plant. Air after leaving the low-pressure compressor enters in the intercooler where it is cooled to its initial temperature as the inter-cooling is assumed as a perfect. 0.1 bar pressure was taken during the flow of air through the intercooler. Air after leaving the high-pressure compressor it enters in the regenerator (heat exchanger) where it absorbs the heat energy from the hot gases coming out from the high-pressure turbine. Since air is heated using heat recovery amount of specific heat energy needed to be added in the combustion chamber will reduce and thus there is a saving in the fuel required for the combustion process. Air temperature leaving the regenerator is estimated by the Eq. 10.

$$\varepsilon = \frac{T_5 - T_4}{T_9 - T_4} \quad (10)$$

Heat is then supplied in the high-pressure combustion chamber by burning hydrocarbon fuel. The specific heat supplied in the HP and LP combustion chamber is determined by the Eq. 11.

$$q_s = c_{pg}(T_6 - T_5) + c_{pg}(T_8 - T_7) \quad (11)$$

Required air fuel ratio is then calculated by applying energy balance to the HP and LP combustion chamber. Efficiency of the combustion is taken as 0.8 and the heating value of the is taken as 42 MJ/kg. Figures 2 and 3 shows the details of the energy interactions with the HP and LP combustion chambers. Equations 12 and 13 are then used to estimate the air fuel ratio required.

$$\eta_{cc} CV = \frac{m_a}{m_f} [(1 + m_f) c_{pg} T_6 - c_{pa} T_5] \quad (12)$$

$$(m_a + m_f) c_{pg} T_7 + \eta_{cc} m'_f CV = (m_a + m_f + m'_f) c_{pg} T_8 \quad (13)$$

Net specific work output is the difference between the total turbine work and compressor work which is estimated by the Eq. 14 and thermal efficiency is estimated by the Eq. 15.

$$w_{net} = (m_a + m_f) c_{pg} (T_6 - T_7) + (m_a + m_f + m'_f) c_{pg} (T_8 - T_9) - m_a c_{pa} (T_2 - T_1) - m_a c_{pa} (T_4 - T_3) \quad (14)$$

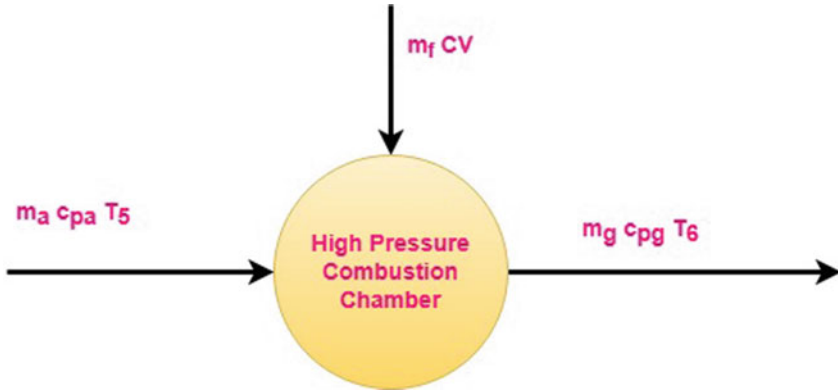


Fig. 2 Energy balance for high pressure combustion chamber

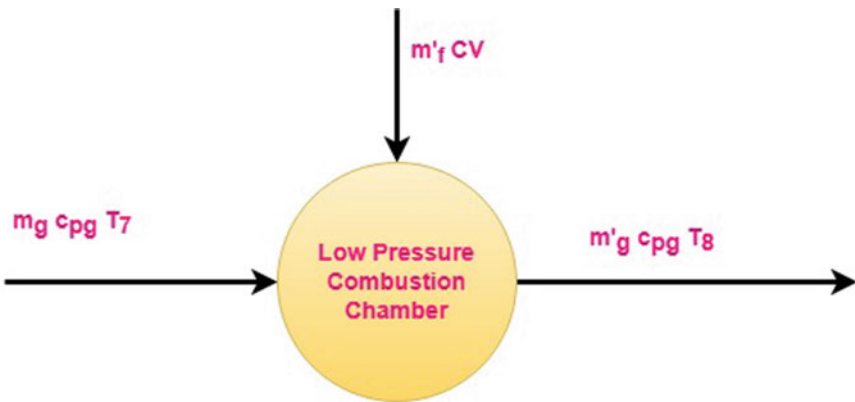


Fig. 3 Energy balance for low pressure combustion chamber

$$\eta_{thermal} = \frac{w_{net}}{q_s} \tag{15}$$

All equations discussed above are solved simultaneously using a computer program engineering equation solver (EES). Program was executed for each of the variables like turbine inlet temperature (TIT), pressure ratio, effectiveness of regenerator and the compressor inlet temperature.

3 Results and Discussion

3.1 Effect of Pressure Ratio and Turbine Inlet Temperature (TIT)

The effect of pressure ratio and the turbine inlet temperature (TIT) on thermal efficiency is shown in the Fig. 4. For any pressure ratio, increase in turbine inlet temperature increase thermal efficiency. It is observed that for every turbine inlet temperature efficiency is maximum at one pressure ratio known as optimum pressure ratio. After this pressure ratio efficiency decreases with decrease in specific turbine output. Rate of decrease in thermal efficiency for low TIT is high compare with high TIT. For a turbine inlet temperature of 700 K optimum pressure ratio is 6 and for the range of temperature 800 to 1200 K the optimum pressure ratio is 9. Hence it is economical to operate the turbine at a pressure ratio of 9. With a pressure ratio 6 and maximum temperature 700 K thermal efficiency is 17.54% and with a pressure ratio 9, temperature of 1200 K thermal efficiency is 32.83%. Thus, with increase in temperature, efficiency increases by 87%. At TIT temperatures of 1200 K thermal efficiency curve gets flatten slowly hence increasing temperature would be advantageous however TIT is limited by turbine blade material. With today’s research advances in materials maximum temperature of 1500 K is possible to improve the efficiency further [6].

Figure 5 represents the effect of TIT and pressure ratio on the specific heat supplied. Specific heat and net specific turbine work increases with increase in turbine inlet temperature. However, the rate of rise in specific turbine output is more only up

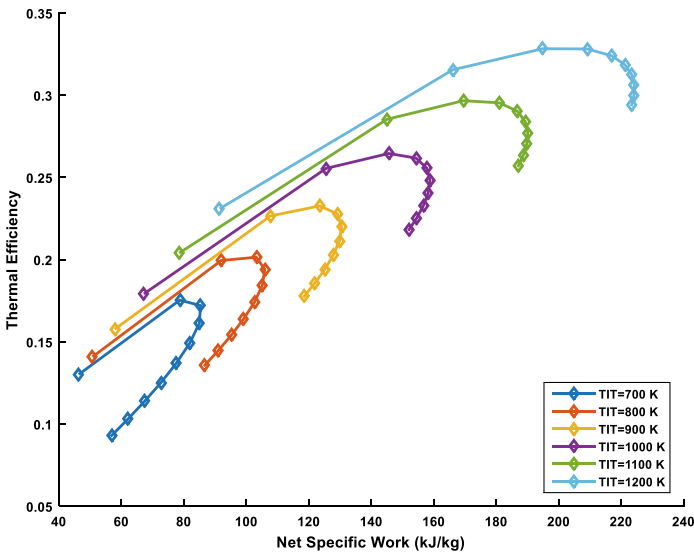


Fig. 4 Variation in thermal efficiency with pressure ratio and TIT

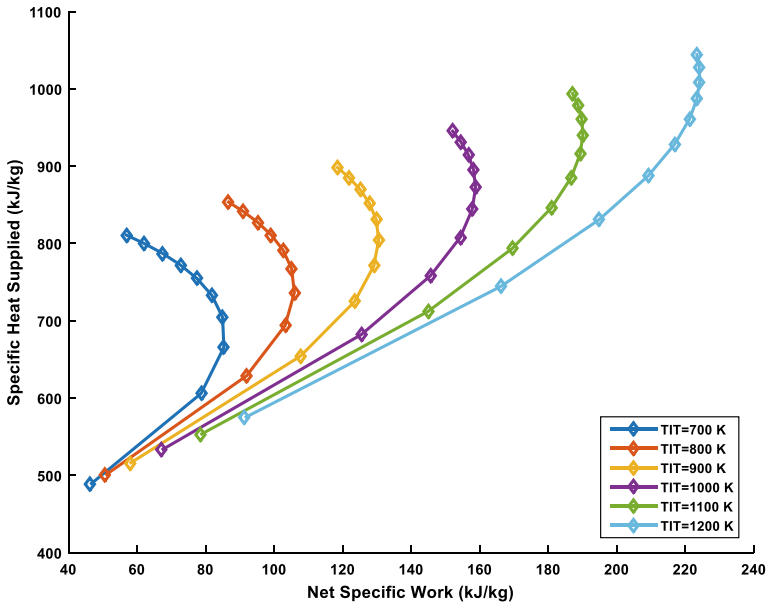


Fig. 5 Variation in heat supplied and turbine output with pressure ratio and TIT

to the optimum pressure ratio. Till pressure ratio of 9 specific work output increases and then the rate of increase in specific heat supplied is more than rate of increase in turbine work output and hence turbine efficiency decreases. At a TIT of 1200 K as pressure ratio increases from 3 to 9, net work increases by 54% and specific heat increases by 49%. However, if pressure ratio increases from 3 to 12 overall turbine work increase by 6% and specific heat increases by 13% and hence thermal efficiency decreases.

Effect of turbine inlet temperature and pressure ratio is represented in the Fig. 6. Air fuel ratio decreases up to optimum pressure ratio for any turbine inlet temperature. For TIT of 700 K air fuel ratio decrease by 19% as pressure ratio increases from 3 to 6 (optimum pressure ratio) and for TIT 1200 K air fuel ratio decreased by 31% as pressure ratio increases from 3 to 9 (optimum pressure ratio). Thus, if the turbine is operated at high temperature (within the temperature limits of the material) and the optimum pressure ratio there is higher potential of saving in the fuel and fuel handling cost. This saving in fuel ultimately results into the lower cost of electricity production per kWh. As the pressure ratio increases beyond the optimum pressure ratio slowly curve approaches the straight vertical line means the net specific turbine output will remain constant with smaller decrease in the air fuel ratio. Thus, operating the cycle beyond optimum pressure ratio is not economical in respect of thermal efficiency, fuel consumption and net specific turbine output.

Effect of pressure ratio and turbine inlet temperature on the specific fuel consumption is represented in the Fig. 7. For a TIT range 700 K to 900 K fuel consumption decreases and at 700 K fuel consumption decreases faster than 900 K. Up to the

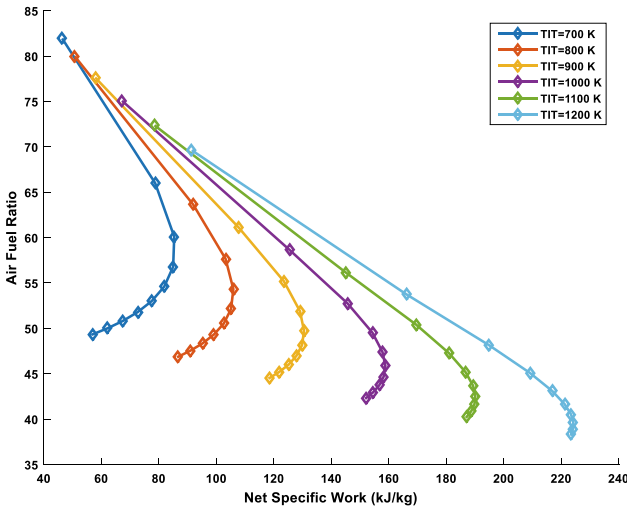


Fig. 6 Variation in air fuel ratio with pressure ratio and TIT

optimum pressure ratio of 6 with TIT 700 K specific fuel consumption is decreased by 32% and then increase rapidly at higher pressure ratio. After an optimum pressure ratio at a lower TIT of 700 K SFC increases with a very small change in the net specific work output of the turbine. It is observed that for a higher TIT, after optimum pressure ratio rate of decrease in specific work output is low. For a higher TIT as pressure ratio increases beyond optimum pressure ratio, specific work output decreases and fuel consumption increases. Higher value of specific fuel consumption is an indicator of increase in size of the plant and hence turbines are required to be operated at the optimum pressure ratio with maximum TIT.

3.2 Effect of Pressure Ratio and Compressor Inlet Temperature (CIT)

Figure 8 represents the effect of compressor inlet temperature for a given TIT of 700 K at various pressure ratio. It is observed that thermal efficiency decreases with increase in compressor inlet temperature (CIT). For every inlet temperature at a certain pressure ratio thermal efficiency is higher. For a CIT of 273 to 278 K the optimum pressure ratio is 9 and it is about 6 for a temperature range of 283 to 303 K. Rate of decrease in thermal efficiency at higher CIT is more than the lower CIT. At a CIT of 273 K maximum thermal efficiency is 0.2123 and for CIT 303 K maximum thermal efficiency is 0.1842. Beyond the optimum pressure ratio 9 and at CIT of 273 K specific work output is decreased by 12% and whereas at CIT of 303 K and

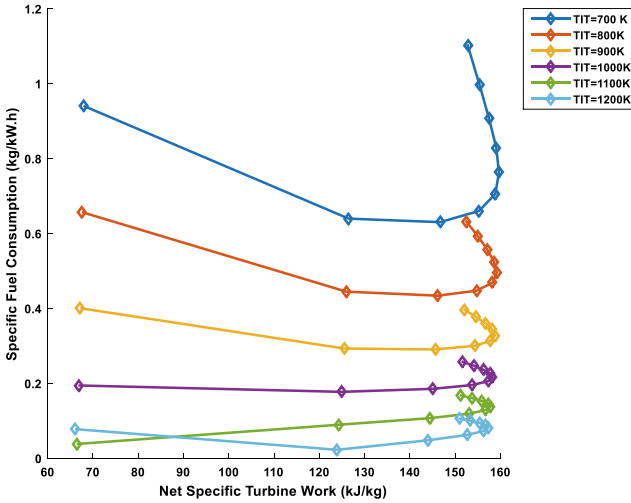


Fig. 7 Variation in fuel consumption with pressure ratio and TIT

optimum pressure ratio 6 net specific work output of a turbine is decreased by 20%. Thus, the rate of decrease in thermal efficiency for high CIT is more than lower CIT.

Figure 9 represents the effect of compressor specific work required. It is observed that at higher temperature of 303 K net specific work of compression increases rapidly

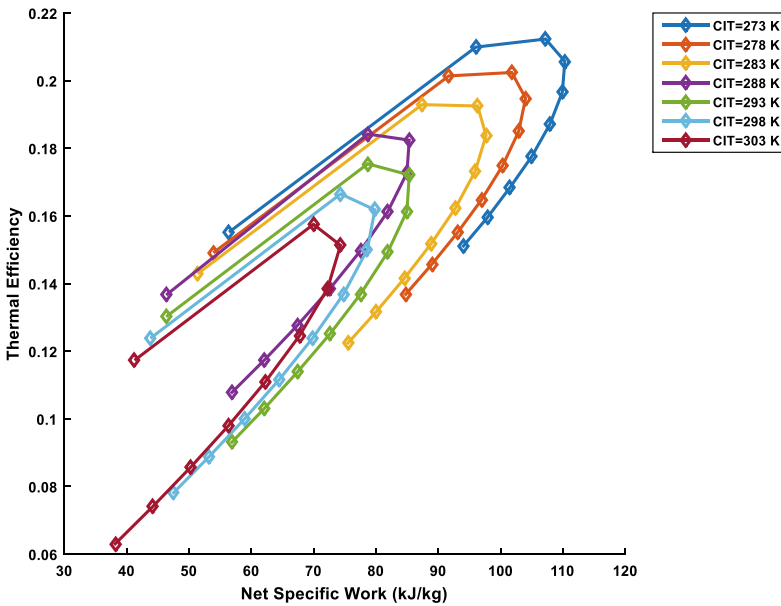


Fig. 8 Variation in thermal efficiency

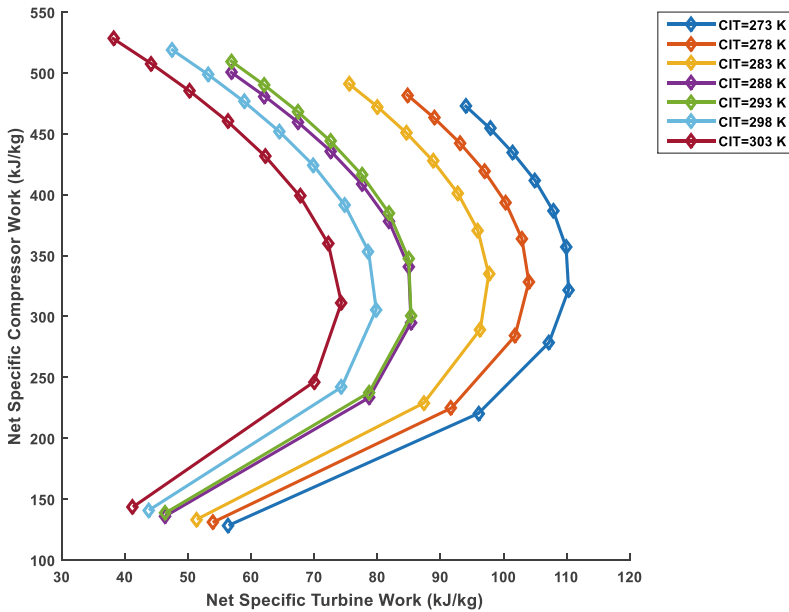


Fig. 9 Variation in specific fuel consumption with pressure ratio and TIT

and the net specific work output of the turbine decreases. This is why the thermal efficiency decreases at higher CIT.

Effect of CIT on total specific heat supplied in the combustion chamber is represented in the Fig. 10. Increase in CIT decreases the specific heat supplied at for pressure ratio. For CIT of 273 K at a pressure ratio of 3 specific heat supplied is 363.6 kJ/kg whereas at a CIT of 303 K at the same pressure ratio heat supplied is 357.7 kJ/kg, thus there is decrease in heat supplied by 1.61%. Till the optimum pressure ratio rise in net specific turbine output is more than the net rise in compressor input and there after turbine output decreases and the compressor output increases. Hence the turbine efficiency decreases with increase in CIT.

Figure 11 represents the effect of the CIT on the specific fuel consumption, it is observed that specific fuel consumption is low for a lower CIT. For a CIT of 273 K to 298 K specific fuel consumption decreases upto the pressure ratio of 9 and then again increases. Hence, the optimum value of pressure ratio is 9 at which fuel consumption is minimum. With increase in pressure ratio SFC decreases, and the net specific turbine output increases up to the optimum pressure ratio. After the optimum pressure ratio specific work output decreases with increase in SFC. For CIT of 273 K up to the optimum pressure ratio 9 SFC decreases by 35% and at a CIT of 303 K up to the optimum pressure ratio 6 SFC decreases by 31%. The rate of rise in SFC is very high for the higher CIT than at lower CIT. This can be observed from the slope of the line as shown in Figs. 11 and 12.

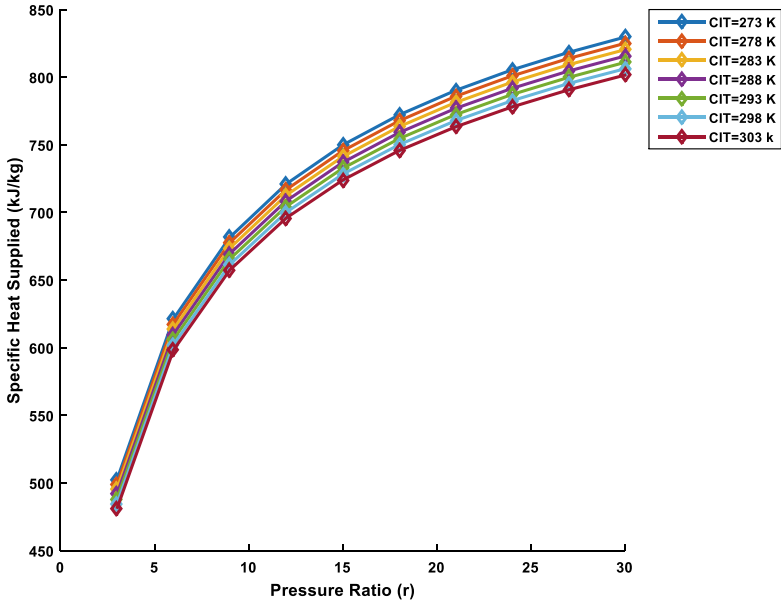


Fig. 10 Variation in specific heat supplied with pressure ratio and TIT

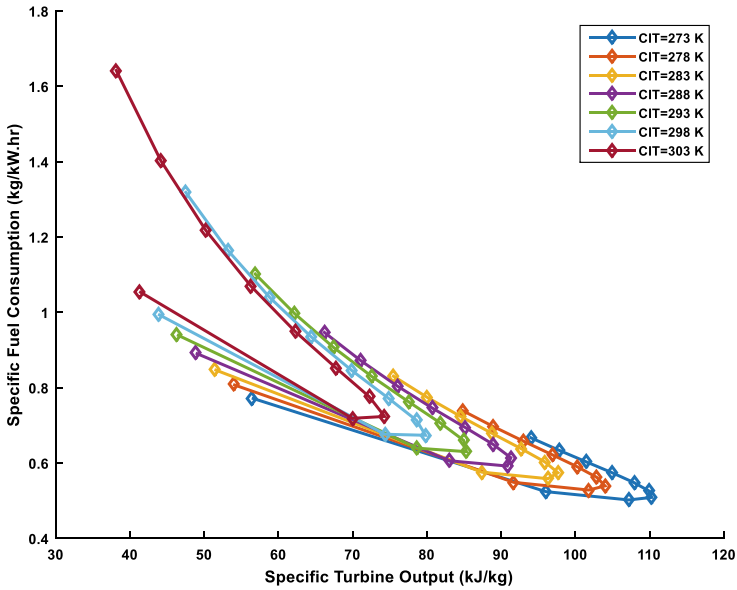


Fig. 11 Specific fuel consumption and turbine output at various TIT

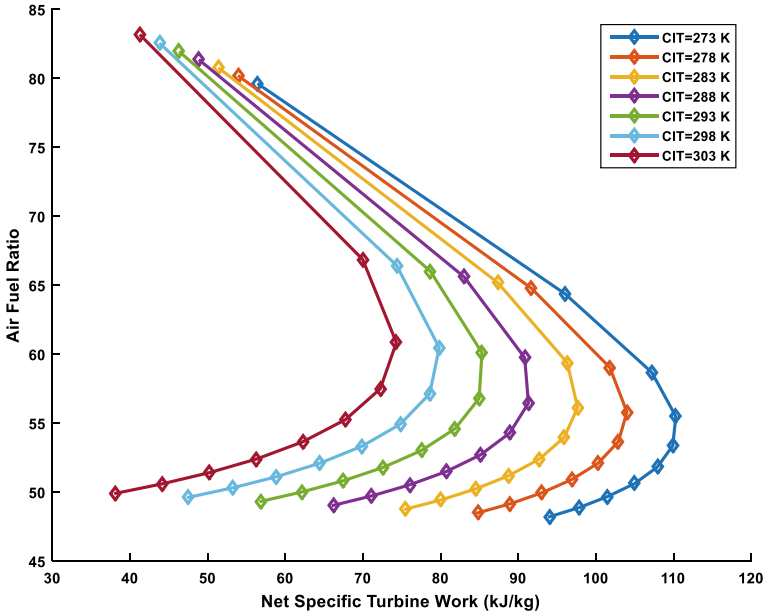


Fig. 12 Variation in air fuel ratio with pressure ratio and TIT

3.3 Effect of Effectiveness of the Heat Exchanger

Effect of effectiveness of heat exchanger on specific work output and thermal efficiency for a maximum temperature of 1000 K is shown in the Figs. 13 and 14. It is observed that with increase in pressure ratio thermal efficiency increases first and then decreases. Thus, at a certain pressure ratio, thermal efficiency is maximum known as optimum pressure ratio. For a pressure ratio of 12 and regenerator effectiveness 0.5 maximum thermal efficiency is 23.65% whereas for an effectiveness of 0.8 thermal efficiency is 28.36%. For an effectiveness of 1 thermal efficiency is 35%. At higher pressure ratio, increase in effectiveness of regenerator results in decrease of thermal efficiency. After the optimum value of pressure ratio, specific output of the turbine decreases, and its rate of decrease is more at higher value of the effectiveness and hence thermal efficiency decreases rapidly.

Also, it is observed from the Fig. 15 that for the lower pressure ratio, specific heat recovery was highest, and it increases with increase in effectiveness of heat exchanger. For the same pressure ratio specific heat recovered with the use of regenerator increases linearly with increase in effectiveness. As effectiveness of heat exchanger changes from 0.5 to 1 specific heat recovery at lower pressure ratio 3 is 50% whereas at a pressure ratio 30 it is about 48%.

Figure 16 represents the effect on specific fuel consumption with variation of effectiveness and the pressure ratio for a constant maximum TIT of 1000 K and CIT

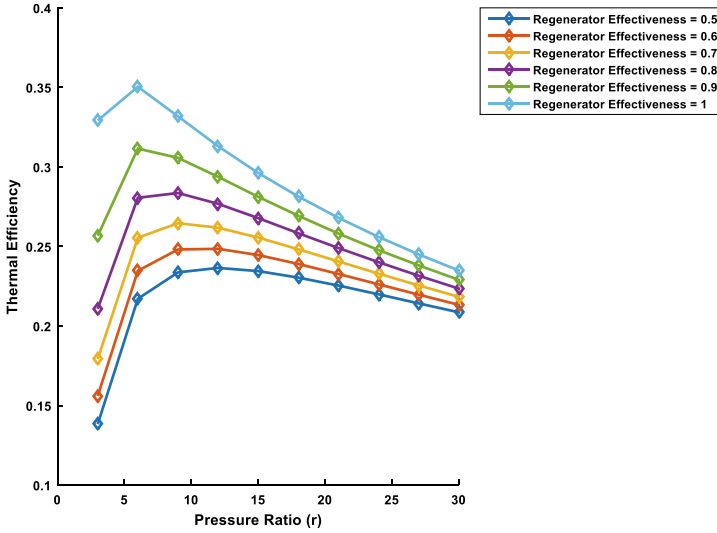


Fig. 13 Variation in thermal efficiency with regenerator effectiveness

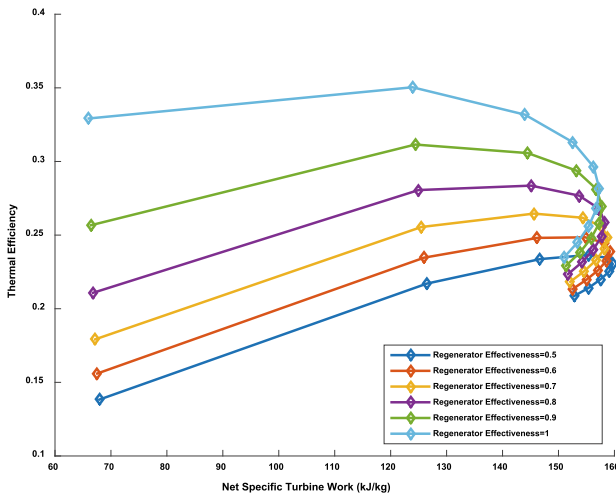


Fig. 14 Thermal efficiency variation with regenerator effectiveness

293 K. As pressure ratio increases from 3 to 6 fuel consumption decreases and after the pressure ratio of 6 fuel consumption again increases.

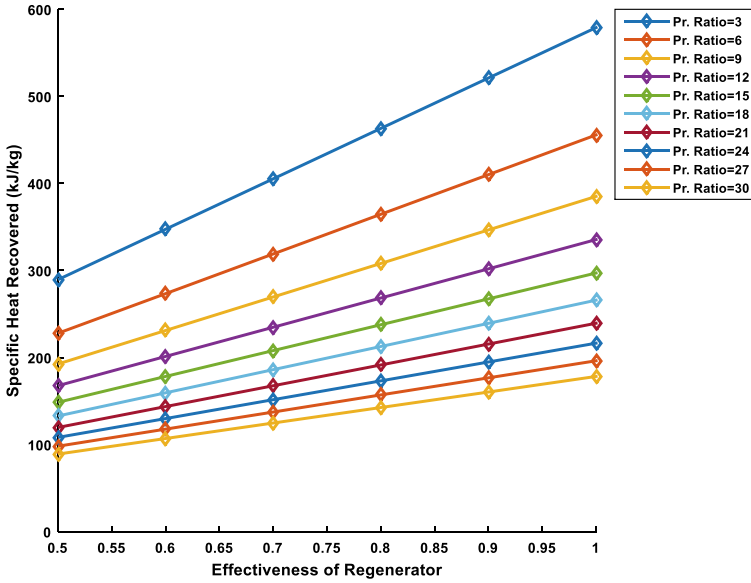


Fig. 15 Variation in specific heat recovery with pressure ratio and regenerative effectiveness

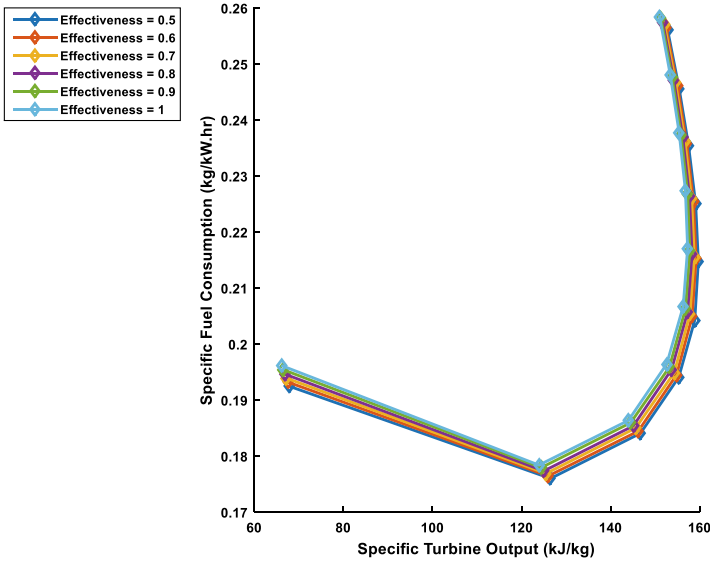


Fig. 16 Effect of regenerator effectiveness on specific fuel consumption

3.4 Development of ANN Prediction Model

Scientific method of gas turbine blade fault was proposed by Zhuo et al. in [21] using Elman neural network with hidden output feedback. Use of multi network ANN model using back propagation was proposed by Mohammadreza Tahan et al. [22] for gas turbine performance. Gas turbine system fault detection and diagnosis was proposed by Yu Zhang et al. [23]. Sina Tayarani-Bathaie S. and Khorasani K. in [24] studied the fault detection for aircraft gas engines using neural network. Many studies using ANN was observed mainly for fault detection. Few research reports are observed for the performance prediction of gas turbine using ANN. Hence, this study focused on ANN model development for the prediction of the gas turbine performance.

Using an ANN is a novel approach for the modeling of the gas turbine. Many complex problems can be easily solved by ANN because of its self-learning ability. Artificial neural network is also known as neural network that is useful for pattern recognition and trend analysis. Neural network technology is the study that is analogues the human brain and nervous system. ANN modeling includes the steps like system analysis, data collection, data processing, network architecture, selection of training set, testing set, and validation set etc. During the computation ANN consist of highly interconnected neurons that processes with an external input. These neurons are arranged as input, hidden and output layer as shown in Fig. 17. For a complex system the number of neurons required are more. ANN performs the iterative procedure and learns the relation between input and output thus it has a self-training approach.

Every input parameter has its own weight, and these are determined and adjusted during the training process. Thus, the selection of the input and output parameters is the most important phase of ANN modeling, and the accuracy of the output parameters is verified using the sensitivity analysis [13]. Following are the steps adopted for the development of the ANN modeling.

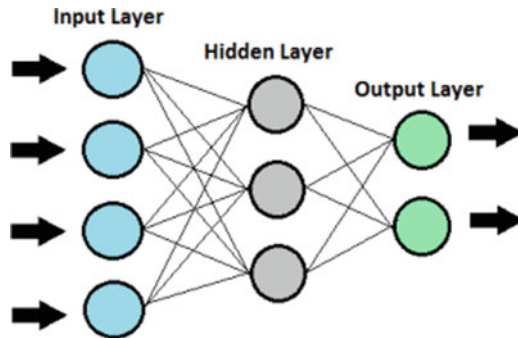


Fig. 17 Structure of ANN

1. Study of the system – To decide the input and output structure of the ANN this step is important. Gas turbine system was studied for the reheating and regeneration system. Different input parameters were varied, and the performance of the system was studied.
2. Data collection – Gas turbine numerical model was developed, and the model was evaluated using the engineering equation solver program [14]. Care has been taken while developing the numerical model to consider the actual operating conditions and the various pressure losses and the efficiencies of the components. Data was then generated by solving the numerical model with variation in the important parameters. The target data for the study was considered as the thermal efficiency of the gas turbine [15].
3. Network Architecture – Neural Network Architecture was developed using MATLAB software. Two different networks are generally employed known as feed forward and feed backward. In a feed forward network there is no feedback from the neurons whereas in feedback word output depends on the current input to the network and the previous input thus there is feedback by every neurons [16].
4. Network training and validation – Two different training approaches are used known as supervised training and unsupervised training. In supervised training the input and target outputs are known. Whereas in an unsupervised training target output is unknown. This model uses supervised training approach [17].
5. This model uses a batch training which can be easily handled with the MATLAB toolbox.
6. Data obtained is the divided into three subsets as training, testing and validation. 70% data set is used for training and 15% data is used for testing and validation each. Training data is used for computation of the gradient and weight updation. Validation data set was used for the verification of the developed model. Test set data was used after the training and validation to test the model.
7. The training objective in ANN is to reduce the error in order to improve accuracy of the model. Accuracy of the developed model was measured by considering the MSE (mean square error) and RSME (root mean square error as defined in the following equations [18–20]).

$$MSE = \frac{1}{n_d} \sum_{i=1}^{n_d} \left(\frac{y_{mi} - y_i}{y_{mi}} \right)^2$$

$$RMSE = \sqrt{\frac{1}{n_d} \sum_{i=1}^{n_d} \left(\frac{y_{mi} - y_i}{y_{mi}} \right)^2}$$

8. To find the best-suited model the code generated was run in the MATLAB and different ANN structures were trained with the use of training, validation and testing data set and different values of the hidden layers. Finally, the model

MSE was considered for the selection. Best model was selected which has the minimum MSE.

3.5 Results of the ANN Prediction Model

Figure 18 represents the results of best selected model based on its performance. Training, validation and testing performance is represented. It is observed that the model has the minimum value of the MSE at the epoch number 2 and equals to 0.00018488. The training was continued for 8 more number of iterations, and it was observed that the MSE value was nearly constant for testing and validation thus the model has good performance with low value of mean square error. Figure 19 represents the performance of developed ANN model which has a mu value 0.00001. Since the mu value is very low the model is effective for prediction of turbine performance.

Regression coefficient R^2 indicates the relation between the model output and the targeted output as represented in the Fig. 20. It is the statistical measure that indicates how well the model fits the approximation. The value of R^2 is less than one. Closer value indicates the good predictability. The regression value for the training, testing and validation is 0.97551, 0.90042 and 0.98084. Further overall R value 0.9679. Since R value for overall network, testing and validation closely matches the model has a good performance and the predictability with a mean square error of 0.00018488.

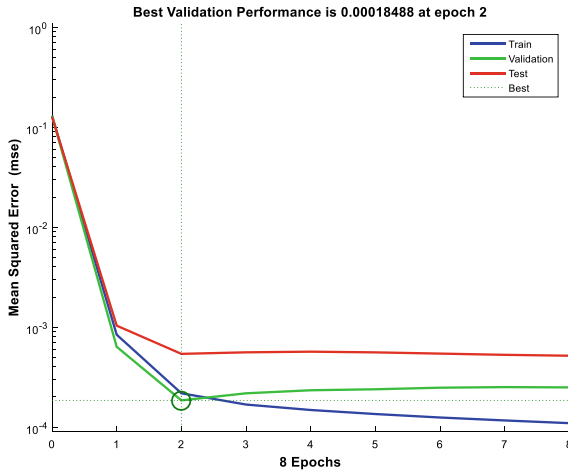


Fig. 18 Performance of ANN Model

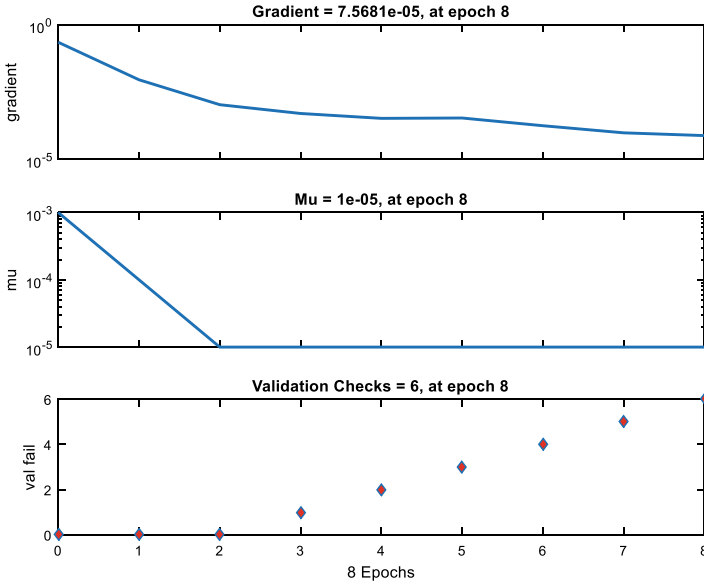


Fig. 19 Performance of ANN Model

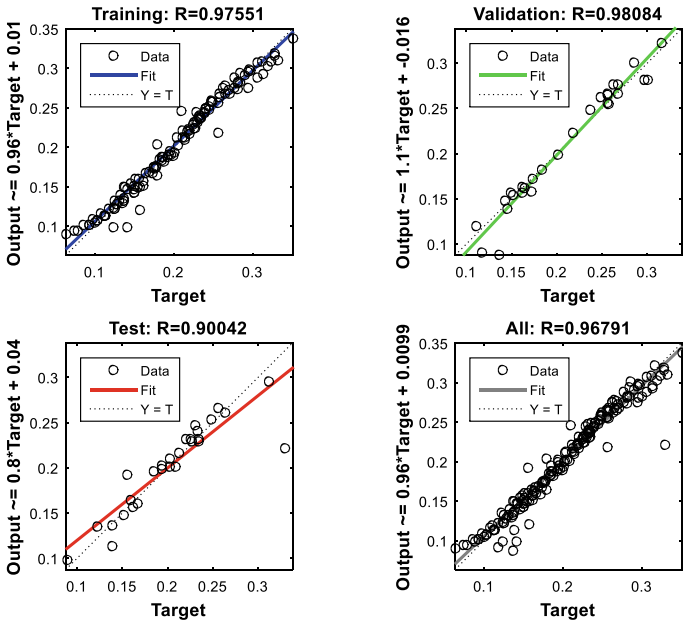


Fig. 20 Regression plot for the developed ANN Model

4 Conclusion

Numerical analysis was conducted for the combined intercooled, reheat and regenerative cycle. Turbine was analyzed for different pressure ratio, compressor inlet temperature, turbine inlet temperature and effectiveness of heat exchanger. Following are the conclusions of this study.

- It is observed that for a turbine inlet temperature of 700 K optimum pressure ratio was 6 and for the range of temperature 800 to 1200 K the optimum pressure ratio was 9.
- Running the turbine at high temperatures 1200 K and at optimum pressure ratio of 9 thermal efficiency curve gets flatten slowly hence increasing temperature would be advantageous however TIT is limited by turbine blade material.
- For TIT of 1200 K as pressure ratio increases from 3 to 9 net work increases by 54% and specific heat increases by 49%.
- For a TIT 1200 K air fuel ratio decreased by 31% as pressure ratio increases from 3 to 9
- Up to the optimum pressure ratio of 6 with TIT 700 K specific fuel consumption is decreased by 32% and then increase rapidly at higher pressure ratio.
- Rate of decrease in thermal efficiency at higher CIT is more than the lower CIT. At a CIT of 273 K maximum thermal efficiency is 0.2123 and for CIT 303 K maximum thermal efficiency is 0.1842.

This work was carried to investigate the methodology of modeling and simulation of gas turbine using artificial neural network. This approach is useful for predicting the performance before designing and manufacturing the gas turbine. Data obtained from the analysis was then used for the development of the prediction model using artificial neural network. The developed model has root mean square error equal to 0.00018 and the regression coefficient of the trained network is 0.99. This way the developed mathematical model was validated, and it has a good predictability.

References

1. Bhargava RK (2006) Global energy market – Past, present, and future. ASME Paper No. GT2006–91322
2. Termuehlen H (2001) 100 years of power plant development, ASME Press, New York, ISBN: 0791801594
3. Sawyer RT (1947) The modern gas turbine, 2nd edn. Prentice-Hall, New York
4. Bakken LE, Jordal K, Syverud E, Veer T (2004) Centenary of the first gas turbine to net power output: A tribute to Aegidius Elling. ASME Paper No. GT2004-53211
5. Lukas H (1986) Survey of alternative gas turbine engine and cycle design. EPRI Report, AP-4450, Research Project 2620-2
6. Scalzo AJ, Bannister RL, DeCorso M, Howard GS (1994) Evolution of heavy-duty power generation and industrial combustion turbines in the United States. ASME Paper No. 94-GT-488
7. Rao AD, Samuelsen GS, Robson FL, Geisbrecht RA (2002) Power plant system configurations for the 21st century. ASME Paper 2002-GT-30671

8. Rahman MM, Ibrahim TK, Taib MY, Noor MM, Kadirgama K, Bakar RA (2010) Thermal analysis of open-cycle regenerator gas-turbine power-plant. *World Acad Sci Eng Technol* 68:801–806
9. Mahmood FG, Mahdi DD (2009) A new approach for enhancing performance of a gas turbine (case study: Khangiran Refinery). *Appl Energy* 86:2750–2759
10. Mahmoudi SM, Zare V, Ranjbar F, Farshi L (2009) Energy and exergy analysis of simple and regenerative gas turbines inlet air cooling using absorption refrigeration. *J Appl Sci* 9(13):2399–2407
11. Cohen H, Rogers GFC, Saravanamuttoo HHH (1996) Gas turbine theory. In: 4th Edition, Longman Group Ltd. England, pp 37–62. ISBN 0–582–23632-0
12. Chappel MS, Cockshutt EP (1974) Gas turbine cycle calculations: thermodynamic data tables for air and combustion products for three systems of units, Ottawa, NRC No. 14300
13. Caudill M (1989) Neural network primer: part I. *AI Expert* 2(12):46–52
14. Klein A, Alvarado FL (2004) EES-Engineering Equation Solver. Version 6.648 ND, F-Chart Software, Middleton
15. Beale MH, Hagan MT, Demuth HB (2011) Neural Network Toolbox™ User’s Guide, R2011b ed. MathWorks, Natick, p 404
16. Curtiss PS, Massie DD (2001) Neural network fundamentals for scientists and engineers. www.cibse.org/pdfs/neural.pdf
17. Cybenko G (1989) Approximation by superpositions of a sigmoidal function. *Math 195 Control Signals Syst* 2(4):303–314
18. Karlik B, Olgac AV (2010) Performance analysis of various activation functions in generalized MLP architectures of neural networks. *Int J Artif Intell Expert Syst (IJAE)* 1(4):111–122
19. Debes K, Koenig A, Gross HM (2010) Transfer functions in artificial neural networks: a simulation-based tutorial. Department of Neuroinformatics and Cognitive Robotics, Technical University Ilmenau, Ilmenau
20. Rajesh SR (2021) Gas leakage detection in pipeline by SVM classifier with automatic eddy current based defect recognition method. *J Ubiq Comput Commun Technol (UCCT)* 3(03):196–212
21. Zhuo P, Zhu Y, Wu W, Shu J, Xia T (2018) Real-time fault diagnosis for gas turbine blade based on output-hidden feedback elman neural network. *J Shanghai Jiaotong Univ (Sci)* 23:95–102
22. Tahan M, Tsoutsanis E, Muhammad M, Karim ZAA (2017) Performance based health monitoring, diagnostics and prognostics for condition-based maintenance of gas turbines: a review. *Appl Energy* 198:122–144
23. Zhang Y, Bingham C, Garlick M, Gallimore M (2017) Applied fault detection and diagnosis for industrial gas turbine systems. *Int J Autom Comput* 14(4):463–473
24. Sina Tayarani-Bathaie S, Khorasani K (2015) Fault detection and isolation of gas turbine engines using a bank of neural networks. *J Process Control* 36:22–41

Implementation of IoT Enabled Home Automation System



Dikshan Shah

Abstract For people suffering from quadriplegia, paraplegia, or other types of physical disabilities, the quality of life diminishes as the day progress into months. This research proposes a cumulative and effective solution for monitoring and controlling home appliances using IOT Stack Level 1. This research can communicate with home appliances using Bluetooth through low power consumption protocols with minimal costs and storage issues. This Home automation approach is controlled by an android application developed by MIT App Inventor, which communicates with Bluetooth signals and conveys the same to the relay board. An extra facet of this research implementation is the socket range. The socket can control all appliances ranging from the air conditioner, washing machine, phone charger, and coffee maker, to name a few. The Bluetooth communicates with the corresponding relays using the application made on the phone. Here, the user can move and control the appliances directly using their voice or clicking a button. It allows the user to communicate and manage any device using the relay board. So we suggested a climbable and price-efficient Home Automation System for the physically challenged.

Keywords Home automation · Arduino Nano · Bluetooth · Internet of Things

1 Introduction

The quality of one's life can be largely excelled by the ease of living facilities and the following tendencies. Freedom to move, operate, and do their daily chores largely affects one's mindset. The less fortunate ones often find themselves in tribulation when asking for help, even while doing small chores like switching on the light or Fan. It is uncanny and inhumane not to consider a world where the physically disabled can be made. The initial phase of this research was to strategize a way to help and benefit the disabled solely. For people with less than 25% disability, this

D. Shah (✉)

Vanita Vishram Women's University, Surat, Gujarat, India

e-mail: dikshan.shah@vwwusurat.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

251

P. P. Joby et al. (eds.), *IoT Based Control Networks and Intelligent Systems*,

Lecture Notes in Networks and Systems 528,

https://doi.org/10.1007/978-981-19-5845-8_18

approach can also be a boon. Freedom to work and operate is largely limited for the individuals whose daily needs are dependent on assistance and help.

While technology moves further, we planned to increase the HMI (Human to Machine Interaction) by solely voice commands cheaply and efficiently. With HMI moving a step ahead and merging with the Internet, we were introduced to the Internet of Things, popularly known as IoT. This research aims to connect any possible electrical appliances with Bluetooth and access them from the range of your home. The major problem of hosting on the Internet using Wi-Fi is security. While security in IoT poses a major threat to the integrity of the concept, Bluetooth-controlled modules will help prevent this. The IoT applications are not just limited to one particular field [8]. There is a tremendous improvement in the existing technology from small-scale implementations to large-scale integrations.

Our suggested research approach is an open-ended application that can primarily operate any electrical appliances in your home. It can also control fans, lights, bulbs, and sockets and is not limited to multimedia applications [12]. A conversation with the physically disabled revealed that system control appliances, but each has its system and interfaces. Hence, the goal was to build a comprehensive system and cohesive and tightly coupled with one interface. Making it cheap was the aim, but building it together was the motivation.

A home appliance can be defined as a device or an instrument that works on electrical energy and performs a particular function [4]. Automation in home appliances reveals the basic control to switch it on or off. One can use it remotely or close to the device [13]. A range of Bluetooth is approximately within the entire home unless you're living in Buckingham Palace [11]. Imagine entering a room and controlling every appliance present with your voice/button. Not only will the disabled be benefited, but older people who live alone will be hugely comforted with this living style. The suggested approach can be controlled with just your android phone (with the configuration version of 4.2 or higher) and Bluetooth. Therefore, people who have low network connectivity will use this system with complete ease. This paper will describe the approach taken and the methodologies used in developing this automated system.

2 Functional Requirements

The major functional requirements and their uses are given below:

2.1 *Internet of Things*

IoT is the base for creating and communicating with a thing that predominantly doesn't use network help to work. Wikipedia defines IoT as the interconnection of various objects, introducing a fangled communication between objects and humans.

Although it seems quite far-fetched to communicate with a non-living thing as if it were human, IOT has made it possible to introduce huge advancements in technology, healthcare, better living, security, mining, and agriculture, to name a few [9]. Introduction to this varied sector has brought in a humongous wave of technological advancements that will develop day by day.

2.2 System Architecture

The IOT architecture plays a major role in yielding, designing, and operating the entire system. With its high flexibility in information communication, the various projects serve a dominant use in mechanisms and functionalities. In our research implementation, we have used stack level-1 incorporated with an Arduino Nano ATmega328P MCU board as in Fig. 1 and a Bluetooth HC05 sensor as in Fig. 2 helps to pair up your device [10]. The HC05 module can also be configured with a password or pin to enhance security. Building a dynamic approach can bring forth a new revolution era when standardizing research models.

3 Proposed Model

The users who are using the existing system might add that the proposed implementation is a notable change to the world of home automation, given its flexibility and cost. Because it doesn't require Internet Connectivity, and no extra Wi-Fi configuration is needed. Moreover, it can also be used by people who are not much techno-savvy [13]. The range of people who can use this and how they can use it is mentioned in Table 1.

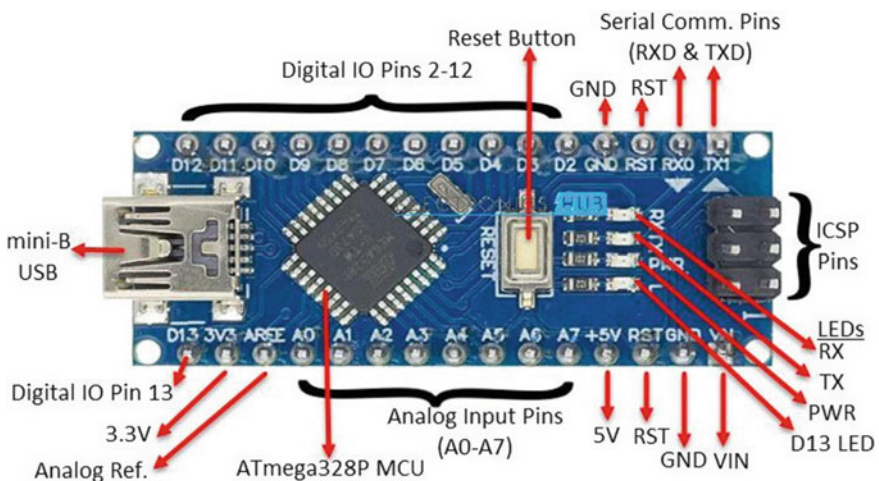


Fig. 1 Arduino architecture [14]

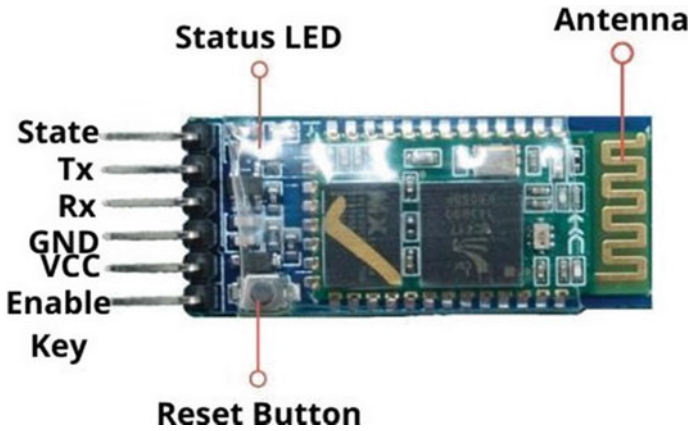


Fig. 2 Layout of Bluetooth board [15]

Table 1 List of disabled people using IOT

Users and the Use cases	Flexibility and Operation
Physical disability ~ Dexterity	Voice-controlled, Remote help
Up to 20% disability	Voice, touch button
Cognitive	Touch Button
Hearing	Voice or Touch
Visual impairment	Voice-controlled guidance
Elderly	Voice/Touch depending on the user

Our main objective was to develop a cost-effective and comprehensive one-stop solution to support the disabled/elderly. The proposed research architecture provides wireless flexibility along with portability. One can plug in the board to the main mount or have individual ones for each part of the room. Whatever be the case, the solution aims to benefit society by delivering and interconnecting various home appliances and a route to communicate between them. It reduces the deployment cost, as all one needs is to buy the pre-programmed Nano board and the Bluetooth module, and a relay for communication. The block diagram of the proposed system is mentioned in the below Fig. 3.

The Bluetooth module asks for a connection request as soon as the circuit is switched on. Upon pairing with the device, one has to open the app developed using MIT App Inventor. Once connected, the existing conditions of the circuit are sensed and updated on the application runtime. Then, every command on the app is signaled through the Bluetooth module, which is connected to the Nano board. The Nano then signals the relay to work accordingly. So ON/OFF module is switched ON/OFF according to the command given through the application.

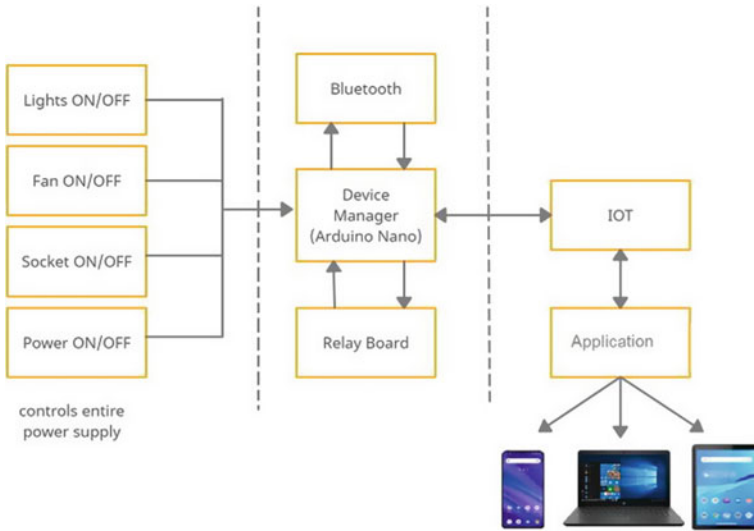


Fig. 3 Architecture of the proposed system

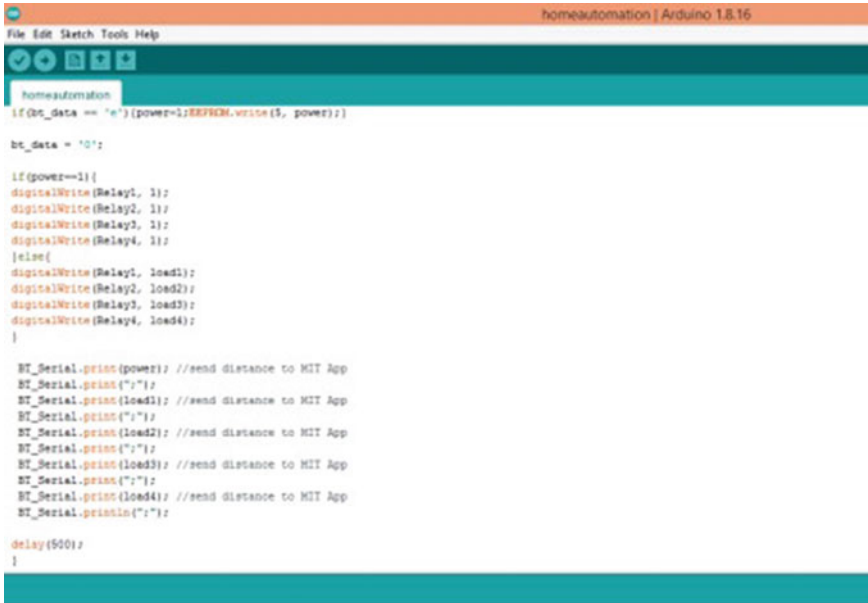
The relay board is connected to various appliances, for which the user completely decides. The proposed implementation can be further be modified and enhanced by plugging in a fire detection system that alerts the users and the nearby brigade and stations.

4 Implementation Design

The design implementation is categorized into five various processes explained below:

4.1 Arduino Nano Board

The Arduino Nano is a small pen drive-sized microcontroller board, a breadboard-friendly device based on ATmega328 [5]. It requires some device drivers to work with, especially COM32 ports. To program the same, install the Arduino programming environment as in Fig. 4. Go to Tools -> Board -> Arduino Nano. Select Processor-> ATmega328. Plugin with the USB cable and make sure that you have selected the appropriate port. Then write your sketch and upload it to configure your board.



```
homeautomation | Arduino 1.8.16
File Edit Sketch Tools Help

homeautomation
if(bt_data == "e"){power=1;EEPROM.write(5, power);}

bt_data = "0";

if(power==1){
digitalWrite(Relay1, 1);
digitalWrite(Relay2, 1);
digitalWrite(Relay3, 1);
digitalWrite(Relay4, 1);
}else{
digitalWrite(Relay1, load1);
digitalWrite(Relay2, load2);
digitalWrite(Relay3, load3);
digitalWrite(Relay4, load4);
}

BT_Serial.print(power); //send distance to MIT App
BT_Serial.print("");
BT_Serial.print(load1); //send distance to MIT App
BT_Serial.print("");
BT_Serial.print(load2); //send distance to MIT App
BT_Serial.print("");
BT_Serial.print(load3); //send distance to MIT App
BT_Serial.print("");
BT_Serial.print(load4); //send distance to MIT App
BT_Serial.println("");

delay(500);
}
```

Fig. 4 Arduino sketch

4.2 Bluetooth HC05 Module

The Bluetooth HC05 module Fig. 5 doesn't require any special coding as it is just a hot plug-and-play device. While it is true to say that these devices are already pre-programmed, a few formalities need to be considered first. The Bluetooth module has four major pins, and connecting each pin with its appropriate representative in the Arduino Nano board is the main job. A voltage higher than 5 V can burn the Bluetooth, and a sample connection is given below.

4.3 Relay Board Module

The relay board is a programmable computer board that serves in controlling the input-output supply by controlling the voltage serving it. It consists of switches that trigger the inflow and outflow of current.

The relay board can consist of different channels where a channel means one single device. So a four-channel relay board can control up to four devices.

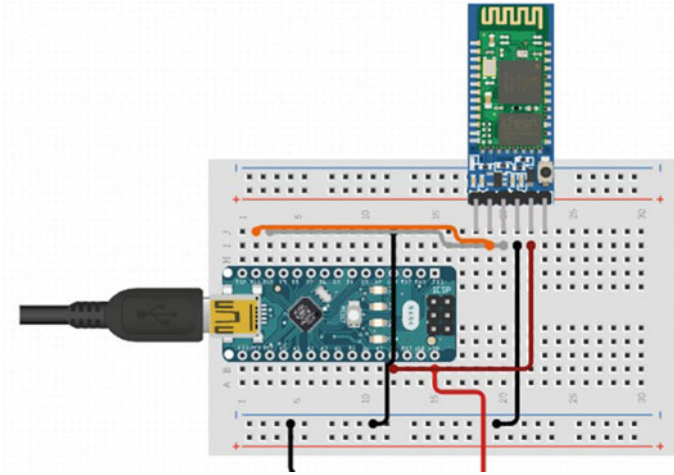


Fig. 5 Connecting Arduino and Bluetooth on a breadboard

Relay boards can give real-time control to the home appliances by connecting with the Arduino Nano. With the Nano signaling of the relay to control the appliance, implementation is complete. Smart Relay has a feature to control Home appliances by mobile Bluetooth. A circuit diagram is given in the below Fig. 6.



Fig. 6 Circuit diagram of the proposed implemented system

5 Proposed IOT Architecture

The proposed model architecture of IOT Level-1 is used for home automation that is given in the Fig. 7 below. Every IOT device follows the OSI model for reference and communication.

The physical layer in the OSI model is responsible for the actual communication between two networks. Therefore, it is also solely responsible for all the device communication in the circuit. All the devices to be authenticated and connected are present in the physical layer. The data link layer is responsible for delivery, protocols, and MAC control. The connected sensors and devices, in an IoT ecosystem, should be able to flawlessly communicate with other devices through the Gateway. In our model, the data link layer also serves as the protocol controller by establishing a gateway. This layer also controls the event-based trigger connection between IoT Gateway, communication protocols, and the device manager. The communication protocol governs and controls the system.

The next layer is the network or the transport layer, where the actual process to process or host to host transfer takes place. We have two hosts: The Bluetooth

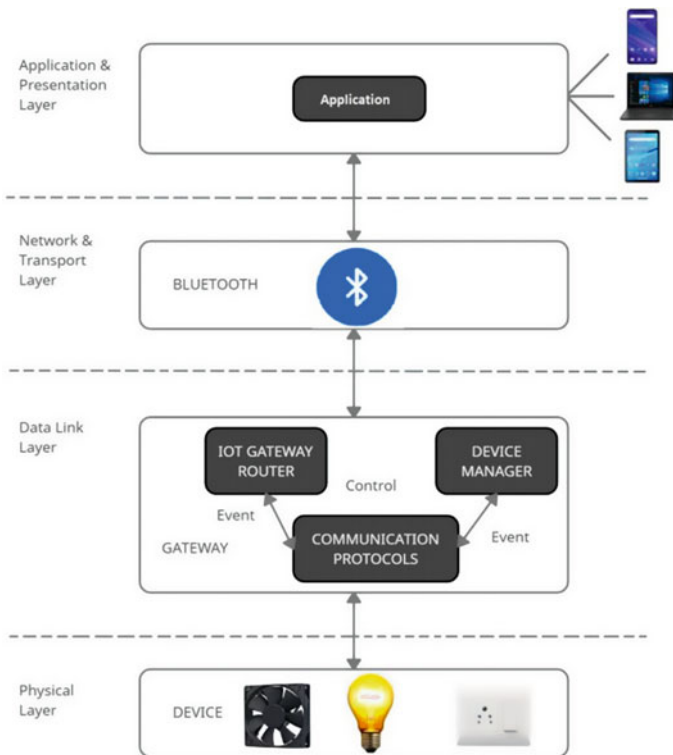


Fig. 7 Proposed IOT architecture for home automation

transmitter and the receiver. The HC05 module is the Bluetooth transmitter that transmits Bluetooth signals received on your phone or the Bluetooth receiver. Even this module acts as the receiver, so two-way transmission or connection will be established in the transport layer. And the last layer is the application layer, where an actual application or the user interaction occurs. The phone acts as a transmitter when the application communicates with a circuit.

6 User Interaction

User interface or User Interaction is the main layer where the user will act or see. Hence, this layer must be made user-friendly and easy to use. We have implemented it with android as it is easy to use and gives many built-in modules and applications to work with it. We have used MIT App Inventor for the same. MIT App Inventor is an online development platform that anyone can control to resolve real-world problems. It also offers a web-based WYSIWYG (What you see is what you get) IDE for developing mobile phone applications directing the Android and iOS operating systems. It uses a block-based programming language built on Google Blockly. When the PressAndSpeakButton is clicked the SpeechRecognizer event is called and is ready for you to speak. The BeforeGettingText event will be activated before the speech has been accepted and identified. Then the Label will display no text on the screen. The AfterGettingText event will be activated once the speech has been received and recognized. Then the Label will display the text on the screen. The workflow and the sequence flow of the app inventor as shown in Fig. 8 are easy to use.

7 Implementation Details

We have selected the Arduino environment. It will begin with the project configures and boot the operating system selected [6]. It helps when you need a clean buffer and is often recommended when configuring Arduino projects. The first thing that needs to be done is to install all the necessary ports. When you plug the USB in your laptop/computing device, ensure that the device detects your Arduino board [7]. The sketch window is as shown in the below Fig. 9.

After the initial configuration, type in your code. The algorithm used in our implementation is as follows:

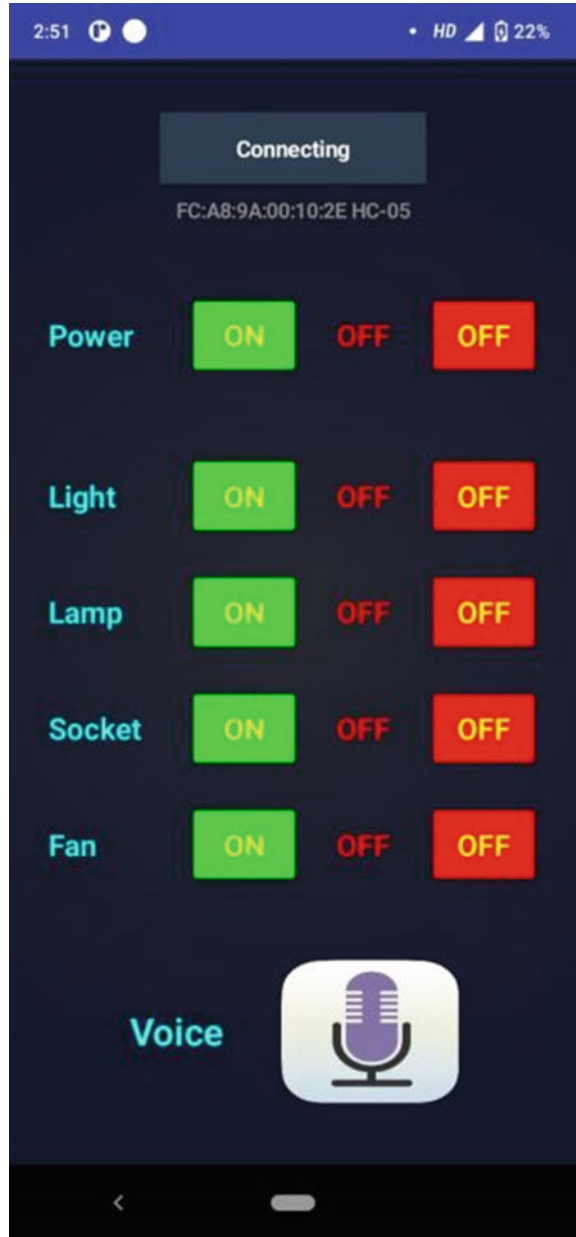
Step 1: *Input User Speech using the application.*

Step 2: *Interpret the user's speech and number it accordingly. Power is numbered as A/a, and Fan is numbered as B/b, Light is numbered as C/c, etc.*

Step 3: *Keep the on/off numbering separate.*

Step 4: *Initially, read the pin and set it at 9600. Import the EEPROM driver. Read from EEPROM and store it in different variables.*

Fig. 8 The Android application to manage IOT device





```
sketch_oct24a
void setup() {
  // put your setup code here, to run once:
}

void loop() {
  // put your main code here, to run repeatedly:
}
```

Fig. 9 Model implementation—Sketch window

Step 5: Switch it according to user input and send data to the MIT App.

Step 6: If power is 1(ON), switch all the other devices as 1(ON) else, keep it 0 (OFF).

These specifications can control any part of your home. If required, a room per room dimension can also be mapped and controlled. Earlier models like Ayad G. Ismaeel’s Home Automation model as in Fig. 10 were only modeled for specific disabled and non-disabled people only [1], and its cost was about 30\$, so it became relatively expensive. Mohd H. Abd Wahab’s model [3] was developed only for physically challenged and older people. However, it didn’t aim to create a comprehensive solution, and it was also relatively costly. Models like Shih-C Chen, Chung-M [2], and Wu were only for people with severe disabilities. We have developed a quick, easy, and low-cost model for people with low to severe disabilities with costs up to 4\$ to implement and use.

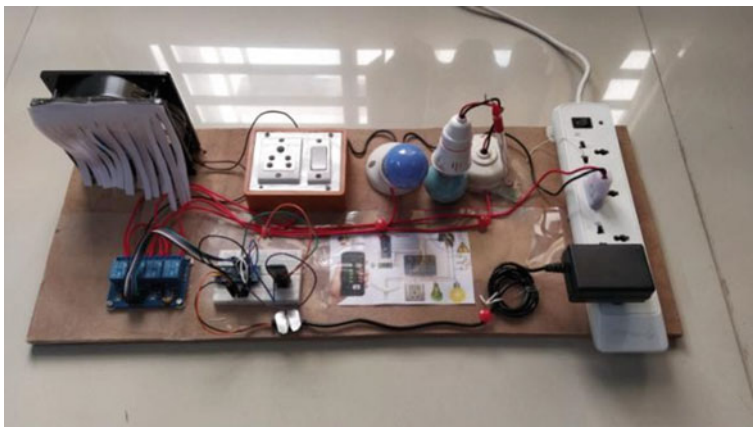


Fig. 10 Implemented IOT enabled Home automation model

8 Conclusion and Future Work

In conclusion, our approach has aimed to create a model that gives you the best designs at the minimum cost possible with efficient outputs. It aims to provide the unfortunate and the elderly with a system that will revolutionize the independence of each disabled person. Our model uses Arduino and Bluetooth to control the home appliances and power them on with your Android phone, and one can even use the voice to command the power appliances. The socket gives the user a wide range of flexibility to work with all appliances. Therefore, commands like ‘make my toast’ or ‘make my coffee’ will be useful. The benefit of the model is the visual consistency that a person can speak that command, functional consistency a person can use the buttons to operate, and, most importantly, a person can do things.

In the future, we would like to enhance the functionalities in this system by creating a timer-controlled on/off switch, a door locking and authentication functionality, and a fire sensor integration for emergencies. Even Smart Meter and Smart plugs can also be developed for Home automation systems to maintain electricity consumption [16]. The microcontroller-based wireless motor control unit was also designed as a Home Automation device [17]. Any device in our home can be converted into digital form.

References

1. Ismaeel AG, Kamal MQ (2017) Worldwide auto-Mobi: Arduino IoT home automation system for IR devices. In: International conference on current research in computer science and information technology (ICCCIT), IEEE Xplor, Iraq, pp 52–57, April 2017
2. Chen S-C, Wu C-M et al (2017) Smart home control for the people with severe disabilities. In: International conference on applied system innovation (ICASI), IEEE Xplore, pp. 503–506. <https://doi.org/10.1109/ICASI.2017.7988465>
3. Wahab MHA (2016) IoT-based home automation system for people with disabilities. In: 5th international conference on reliability, Infocom technologies and optimisation (ICRITO), IEEE Xplore, India. <https://doi.org/10.1109/ICRITO.2016.7784923>
4. Gunge VS, Yalagi PS (2016) Smart home automation: a literature review. In: National seminar on recent trends in data mining-RTDM 2016
5. Jain S, Vaibhav A, Goyal L (2014) Raspberry Pi-based interactive home automation system through E-mail. In: 2014 International conference on optimization, reliability, and information technology (ICROIT). IEEE
6. Lamine H, Abid H (2014) Remote control of domestic equipment from an Android application based on Raspberry Pi card. In: IEEE transaction 15th international conference on sciences and techniques of automatic control and computer engineering – STA 2014, Hammamet, Tunisia, 21–23 December 2014
7. Chowdhury, MdN, Nooman MdN, Sarker S (2013) Access control of door and home security by Raspberry Pi through internet. Int J Sci Eng Res 4
8. Zhang W, Qu B (2013) Security architecture of the Internet of Things oriented to perceptual layer. Int J Comput 2:2–13
9. Kelly SDT, Suryadevara NK, Mukhopadhyay SC (2013) Towards the implementation of IoT for environmental condition monitoring in homes. Sensors J 13(10):3846–3853
10. Assaf MH et al (2012) Sensor-based home automation and security system. In: 2012 IEEE international instrumentation and measurement technology conference (I2MTC). IEEE

11. Al-Kuwari A-M, Ortega-Sanchez C, Sharif A, Potdar V (2011) User-Friendly Smart Home Infrastructure: BeeHouse. In: IEEE 5th international conference on digital ecosystems and technologies, 31 May–3 June 2011, Daejeon, Korea
12. Atzori L, Iera A, Morabito G (2010) The internet of things: a survey. *Comput Netw* 54(15):2787–2805
13. Han J, Yun J, Jang J, Park K-R (2010) Userfriendly home automation based on the 3D virtual world. *IEEE Trans Consum Electron* 56(3), 1843–1847
14. Al-Ali A-R, Al-Rousan M (2004) Java-based home automation system. *IEEE Trans Consum Electron* 50(2):498–504
15. Teja R (2021) Layout of Arduino Nano Board, 12 January 2021. <https://www.electronicshub.org/arduino-nano-pinout/>
16. <https://www.roboelements.com/product/hc05-bluetooth-transceiver-module-with-ttl-output/>. Layout of Bluetooth Board, 12 February 2021. <https://www.roboelements.com/product/hc05-Bluetooth-transceiver-module-with-ttl-output/>
17. Hamdan YB (2021) Smart home environment future challenges and issues-a survey. *J Electron* 3(01):239–246
18. Vinothkanna R (2020) Design and analysis of motor control system for wireless automation. *J Electron* 2(03):162–167

Scheduled Line of Symmetry Solar Tracker with MPT and IoT



A. B. Gurulakshmi, Sanjeev Sharma, N. Manoj, Nikhil A. Bhinge, H. M. Santhosh, and O. M. Yogesh

Abstract The enormous ability to generate the sustainable and eco-friendly energy has increased because, there have been energy shortages in small and developing countries due to the shortage and depletion of non-renewable energy resources in the nature, and hence there is a need for the usage for the renewable energy which will help to reduce the emission of greenhouse gases, global warming and help in the climate change which are the major threats in the future. Among all renewable energies, the solar energy is one of the biggest and eco-friendly energy generators to the world. However, since many years solar energy is being used but not in a complete effective way. The Photovoltaic Solar Systems (PSS) suffer from insufficient sunrays due to the orientation angle i.e., the sunrays are not able to be acquired by the solar panel after a certain period of time because, the solar panel is fixed in one direction and it cannot be moved as the sun rotates, due to which maximum utilization of energy does not occur. To overcome this challenge, a solar tracker has been designed which tracks the exact polar co-ordinates using the existing data and instructs the solar panel to rotate in the direction of sunlight so that the maximum energy can generate the electricity. Moreover, the Maximum Power Tracking (MPT) controller has been used to achieve the maximum energy storage without wasting the generated energy from the solar panel.

Keywords Maximum Power Tracking (MPT) · Photovoltaic Solar System (PSS) · Internet of Things (IoT) · Line of symmetry

1 Introduction

In the current scenario, the articulation of fossils fuels is at its peak. The infusing of non-renewable energy sources has caused pollution and hence cause a major threat to the environment [1]. In the recent years, the usage of non-renewable energy sources has been more and hence it is getting depleted day by day. As per the latest news,

A. B. Gurulakshmi (✉) · S. Sharma · N. Manoj · N. A. Bhinge · H. M. Santhosh · O. M. Yogesh
Department of ECE, New Horizon College of Engineering, Bangalore, India
e-mail: gurulakshmiab@gmail.com

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
P. P. Joby et al. (eds.), *IoT Based Control Networks and Intelligent Systems*,
Lecture Notes in Networks and Systems 528,
https://doi.org/10.1007/978-981-19-5845-8_19

265

there have been power storages in many parts of the country in 2021, due to shortage of coal and other non-renewable energy sources. In order to avoid the environment degradation, the renewable energy sources came into existence, where renewable source of energy is freely available in nature which does not harm the environment and the renewable energy sources are sustainable and never get replenished in the future [2–4]. India has a vast green energy source like solar, wind and hydro etc. which can be effectively used to build a sustainable environment around us.

As per survey, India's potential in renewable energy generation is great with 40% electricity used for domestic purpose. India is 3rd largest producers and consumer of electricity. Green energy sources like solar, wind, and tidal are generated but solar energy is found to be the easiest and efficient way. The energy sources can be used for a large scale as well as small scale production. The harvested solar energy can be substituted as a major source to fulfil the needs of the people. The solar energy is produced from the sun rays and it can be converted to electrical energy with the help of photovoltaic cells, which converts the solar energy to electrical energy [4, 5]. During the conversion of solar energy to electricity PV cells, there are some challenges such as the solar rays falls on the PV cells but only for a certain period of time. So, in order to make the solar rays fall on the PV cells at all intervals of time during the day till the sunsets, the implementation of scheduled line of symmetry solar tracker using Maximum Power Tracking (MPT) is done, where the angle of the PSS cells is adjusted based on the sun's movement from east to west. The solar panel is moved with respect to the sun trajectory using the live co-ordinates from the cloud server, where the solar rays later are converted to electric energy by photovoltaic cells present in the solar panel [6]. If the solar panel is fixed, it will not be efficient as compared to the rotating one. The implementation of scheduled line of symmetry solar tracker will have the latitude and longitude co-ordinates, where azimuthal angle is calculated, where the panel moves in such a way that the solar rays exactly falls on the solar panel and hence the generation of electric power will be high and hence more efficient [7–9].

Scheduled line of symmetry solar tracker efficiency can be improved by using MPT. MPT is used to estimate the maximum power in varying conditions of a day. The different conditions can be morning, afternoon, and evening [10]. The sunlight varies as it keeps moving from east to west, and hence the solar radiation also varies accordingly. MPT achieves the efficient power generation by the solar panel depending on solar panel temperature and load's characteristics, where power is generated by varying the load characteristics. The system is optimized when the maximum power is transferred by varying the load characteristics, and that point is called as maximum power point [10–12]. The optimal circuit is designed such that there will be an optimal load to the photovoltaic cells, and the voltage can be converted suitable to the specific device or appliance.

2 Existing System

2.1 *Optimized Single-Axis Schedule Solar Tracker in Different Weather Conditions*

Nurzhigit Kuttybay et al., 2020 [6], proposed a model for the effective use of the solar panel tracker. In normal solar tracker, the solar panel is fixed but the model designed is a single axis solar tracker which will rotate accordingly to the direction of the sun due to which the efficiency of the solar panels increases and gives more energy for the storage and utilization.

- Limitations: In this model the MPPT (Maximum Power Point Tracking) system is not used due to which there is a slight loss in the efficiency.

2.2 *Coordinated Power Management and Control of Renewable Energy Sources Based Smart Grid*

Siva Ganesh et al., 2021 [3] developed a model to study the effective energy management system for a 4wire 1 MW smart grid system composed of three solar plants and three wind farms with a battery bank.

- Limitations: This model shows how to generate and store electricity but fails when it comes to the effective storage of electricity.

3 Proposed System

A system has been proposed to align a solar panel which receives maximum solar radiation. The solar coordinates are calculated at all momentary time instants and the solar panel is rotated to increase the effective utilization of Photovoltaic Solar System (PSS).

Figure 1 shows the solar tracker system that has seven main functional units, namely microcontroller, MPT controller, line of symmetric rotating solar panel, battery, router, server and appliances. The cloud server is used to store the data of latitude and longitude of the given region, then the angle at which the solar panel is to be rotated is calculated using the co-ordinates which will be further sent to the microcontroller. The router acts as the bridge between the microcontroller and cloud server. Microcontroller is the control unit and gives the PWM signal to the servo motor. To rotate the solar panel, the information obtained by microcontroller from the cloud server is used. The servo motor is attached to the microcontroller and turns the solar panel. The energy generated from the solar panel will be stored in the battery, and MPT controller is used between the battery and solar panel to

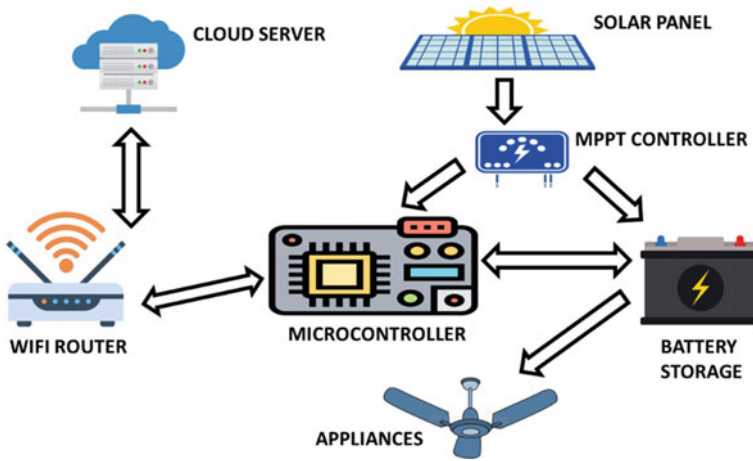


Fig. 1 Block diagram and pictorial representation of scheduled line of symmetry solar tracker

ensure no wastage of energy and also takes care of the energy fluctuations occurring, and always gives the maximum power to the battery. MPT controller is DC to DC converter which has high power efficiency in conversion. Incremental conductance is the simplest method for obtaining maximum power point of solar panel. The energy generated can be used by various electrical appliances.

3.1 Calculation of Azimuthal Angle

The polar coordinate of sun is calculated with respect to the location to determine the angle at which the solar panel should be rotated to receive maximum sunlight as shown in Fig. 2.

Declination Angle (δ) is the angle between the equator and line drawn from the centres of Sun and Earth axes, and it is given by,

$$\delta = 23.45^\circ \times \sin \left[\frac{d + 284}{365} \times 360^\circ \right] \text{ degrees} \tag{1}$$

where,

δ = Declination angle

d = day number of the year, (If February 10 is the date, then d = 41)

Solar Time Period-Local (STP-L) is the time calculated according to the position and observation of the sun from one specific location, and is given by,

$$STP - L = LTZ + (4 \text{ min/degree}) \times (LSTM - LP) + E \tag{2}$$

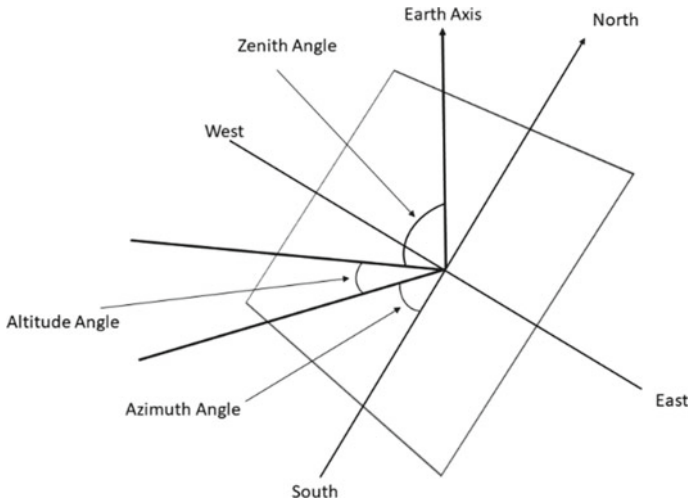


Fig. 2 Solar coordinates

LTZ = Local Time Zone

LSTM = Longitude of Standard Time Meridian

LP = Longitude at the specific Position

E = Equation of time

Local Longitude of Standard Time Meridian (LSTM) is the longitude displacement by 15° from prime meridian axis, and is represented as,

$$LSTM = 15^\circ \times \left(\frac{LP}{15^\circ} \right) \tag{3}$$

Equation of time (E) is the measure of the difference between local solar time and mean solar time, and it can be written as,

$$E = 9.87 \times \sin(2D) - 7.53 \times \cos D - 1.5 \times \sin D \tag{4}$$

$$D = 360^\circ \times \left[\frac{d-81}{365} \right]$$

Solar Hour angle (H) is the angle difference between the earth axis and the zenith.

$$H = \frac{(No. of minutes from midnight of that day, LST)}{4 \text{ min per degree}} \tag{5}$$

where, H = Hour angle

Hour angle will be negative in morning and positive in the afternoon.

Solar Altitude angle ($\alpha 1$) is the angular height of the sun in sky, facing the Zenith angle (θz). It is given by

$$\cos(\theta z) = \sin(\alpha 1) = \cos(l)\cos(\delta)\cos(H) + \sin(l)\sin(\delta) \tag{6}$$

where, $l = \text{latitude } [0 \text{ to } 90^\circ]$

Azimuth angle is the angle between the sun rays and earth axis from the south direction. It is given by,

$$\cos(\beta_1) = \frac{\sin(\alpha_1) \times \sin(l) - \sin(\delta)}{\cos(\alpha_1) \times \cos(l)} \tag{7}$$

3.2 Circuit Diagram

Figure 3 shows the circuit diagram of scheduled line of symmetry solar tracker. ESP8266 (Node-MCU) is used as the Wi-fi microcontroller. The ESP8266 receives azimuthal angle at specific interval of time from the server and gives the PWM signal to servo motor, which is connected to the solar panel to rotate. The amount of energy generated from the solar panel will be around 10 W. Raspberry Pie uses the Mosquitto server for data transfer. Raspberry Pie and Node MCU are connected to the local router. Arduino board is connected to Node MCU and then it is connected to the servo motor.

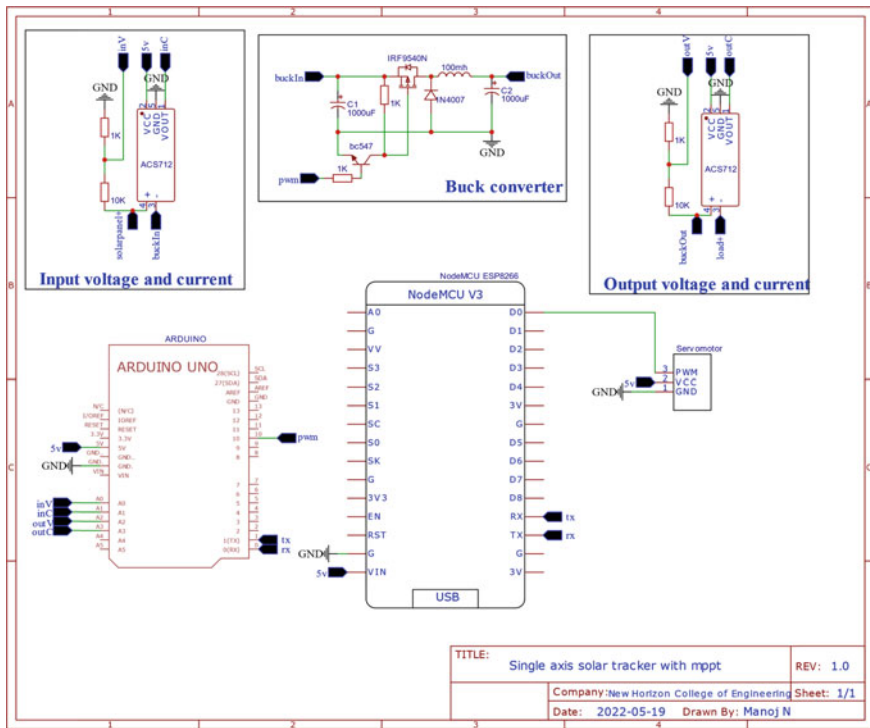
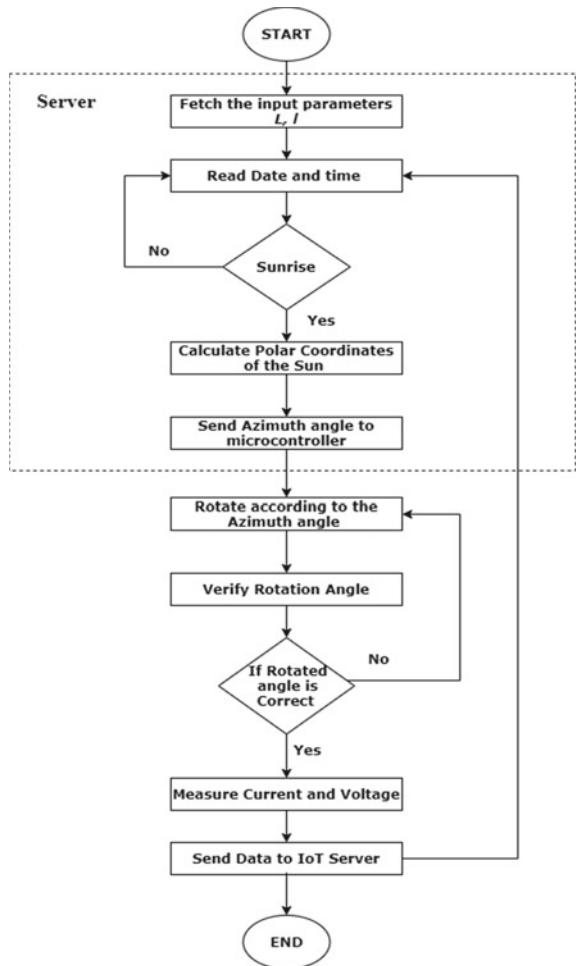


Fig. 3 Circuit diagram of scheduled line of symmetry solar tracker

3.3 Flow Chart

Figure 4 shows the flow chart of the operation of scheduled line of symmetry solar tracker depending upon the astronomical calculation. Longitude, latitude, time and date are considered as the input parameters which will be stored in the server. The spherical co-ordinates of the sun are sent and the azimuthal angle is calculated and sent to Node MCU and servo motor rotates the solar panel. Further, the generated electrical power is observed and recorded in the server for the data storage and future analysis.

Fig. 4 Flow chart of proposed methodology



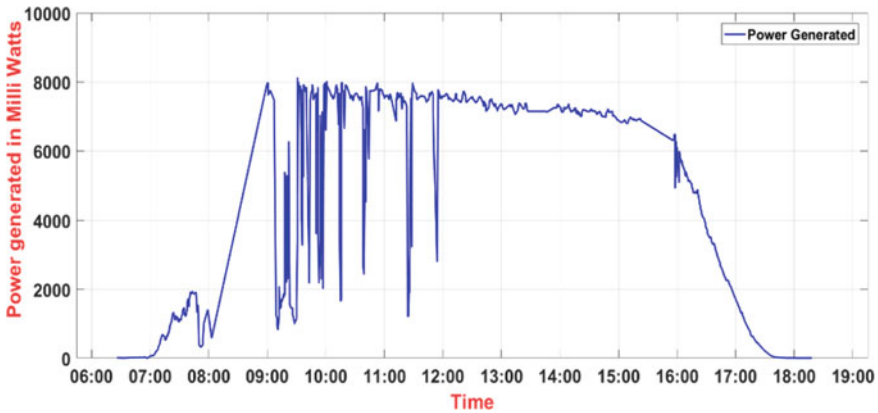


Fig. 5 Graph showing constant power gain by scheduled line of symmetry solar tracker

4 Results and Discussion

In the fixed solar panel system, the solar panel orientation cannot be changed as it is fixed to certain angle according to solar radiation. Due the fixed position of solar panel the efficiency of power generation will reduce. So, to increase the generated power efficiency, the scheduled line of symmetry solar tracker is used.

On the other hand, the generated power cannot be used completely because the load resistance will differ for each application and there is a loss of power. To overcome the power loss issue, maximum power point tracking system is applied to the line of symmetry rotating solar panel. About 85% of the generated energy is efficiently delivered to the appliance. At any given load resistance, the power delivered from the system will be maximum.

Figure 5 shows the amount of solar power generated from 6 am to 6 pm by the line of symmetry rotating solar panel. The average maximum power generated is 8 W. From the above graph, it is observed that constant power has been generated over a time period of 9 am to 4 pm. Using IoT, the visualization of power generation becomes easier and the solar power data can be stored for future analysis. Hence, by using line of symmetry rotating solar tracker the uniform energy can be generated.

Figure 6 is the prototype of line of symmetry rotating solar panel. It consists of 10 W Solar Panel, 12 V Battery, Arduino Board, Node MCU, Current sensors, Buck Converter and Servo Motor. The screenshot of the Node red is shown in the Fig. 7.

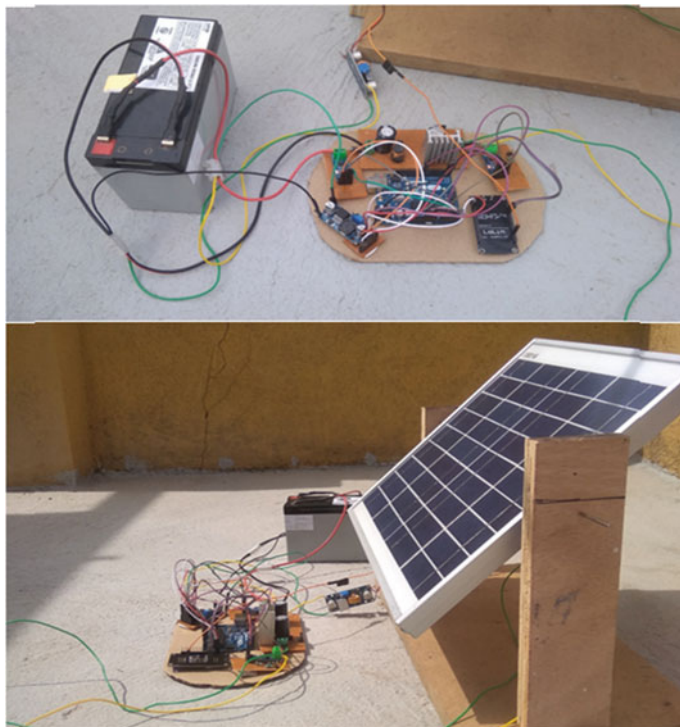


Fig. 6 Prototype of scheduled line of symmetry solar tracker

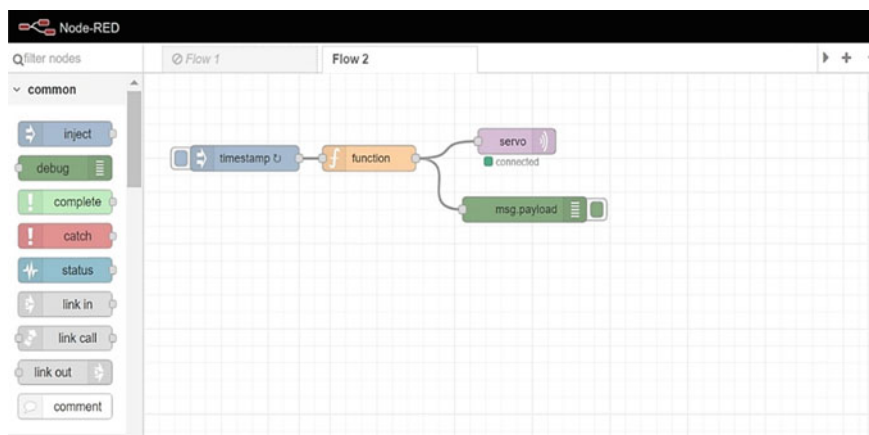


Fig. 7 Screenshot of Node-red

5 Conclusion

The conventional energy resources are depleting in the present years (such as, coal shortage and environmental pollution). So, renewable sources of energy like solar, wind, and tidal which are abundantly available in nature for free as well as which will not get depleted easily can be used. Therefore, this project uses solar energy for power generation, which is not exhaustible. Solar panels absorb the solar radiation emitted by the sun and the solar energy is converted to electric power. Moreover, a scheduled line of symmetry solar tracker with MPT systems has been designed, where the solar panel rotates according to the azimuthal angle calculated by solar co-ordinates, so that the solar radiation falls on the panel that is exactly perpendicular to the sun. Furthermore, different types of MPT algorithms which will be suitable for tracking and which has highest efficiency has been analysed. It is concluded that incremental conductance algorithm which is more efficient and simple for hardware and software implementation suits this project. The incremental conductance algorithm depends on the P-V curve slope, which is dependent on solar radiation and load resistance.

References

1. Prabha SJ, Nisha KCR (2019) A transformer-less current source inverter for grid-connected SPV system. In: IEEE international WIE conference on electrical and computer engineering (WIECON-ECE), pp 1–5
2. Rodrigues S et al (2016) Economic feasibility analysis of small scale PV systems in different countries. *Sol Energy* 131:81–95
3. Malla SG (2021) Coordinated power management and control of renewable energy sources based smart grid. *Int J Emerg Electr Power Syst* 000010151520210113
4. Nsengiyumva W, Chen SG, Hu L, Chen X (2018) Recent advancements and challenges in solar tracking systems (STS). *Renew Sustain Energy Rev* 81:250–279
5. Nisha KCR (2016) PV powered generalized multicell switched-inductor embedded quasi Z-source inverter using MSP-430 controller. In: International conference on circuit power and computing technologies (ICCPCT), pp 1–6
6. Kuttybay N (2020) Optimized single-axis schedule solar tracker in different weather conditions. *Energies* 13:5226
7. Pillai DS, Ram JP, Ghias AMYM, Mahmud MA, Rajasekar N (2020) An accurate, shade detection-based hybrid maximum power point tracking approach for PV systems. *IEEE Trans Power Electron* 35(6):6594–6608
8. Nisha KCR, Basavaraj TN (2013) DC link embedded impedance source inverter for photovoltaic system. In: International conference on circuits, power and computing technologies (ICCPCT), pp 418–423
9. Ranganathan R, Mikhael W, Kutkut N, Batarseh I (2011) Adaptive sun tracking algorithm for incident energy maximization and efficiency improvement of PV panels. *Renew Energy* 36:2623–2626
10. Nisha KCR, Jain S (2015) Solar powered switched capacitor reduced source DC-link impedance source inverter system. *Int J Appl Eng Res* 33071–33077
11. Rajesh T, Gunapriya B, Sabarimuthu M, Karthikkumar S, Raja R, Karthik M (2021) Frequency control of PV-connected micro grid system using fuzzy logic controller. *Mater Today Proc* 45(2)
12. Priyabrata A, Pankaj RKR, Asis M (2015) Optimal renewable energy project selection: a multi-criteria optimization technique approach. *Glob J Pure Appl Math* 11(5):3319–3329

A New Method for Secure Transmission of Medical Images Using Wavelet Transform and Steganography



S. Jayanth, Y. Sushwanth, Poornima Mohan, N. Tejesh, and M. Pavan

Abstract Protection of data is the major concern as digitization has become a new meta and technology is reaching its peaks. The proposed method in the paper is useful for secure transmission of any secret images including medical images. We first encrypted our image then embedded cover. The scrambled data embedded cover image can be used for secure transmission of data. We obtained satisfactory outputs in implementation. We have used python for coding purpose.

Keywords Compression · DWT (Discrete Wavelet Transform) · Steganography · Bit plane manipulation

1 Introduction

Medical data is considered as most important data because they contain sensitive data about the patient and this data falling in wrong hands may lead to some critical situations that may put the patient's personal life and health in risk. There can be situations in which such patient specific information has to be transmitted over public channels for example to a caretaker.

So, in this work we designed and implemented a technique for secure transmission of medical images over any network. In our method we employ DWT (Discrete Wavelet Transform) for encryption of the medical image. After encryption, we have

S. Jayanth (✉) · Y. Sushwanth · P. Mohan · N. Tejesh · M. Pavan
Department of Electronics and Communication Engineering, Amrita Vishwa Vidyapeetham,
Kollam, Amritapuri, India
e-mail: jayanthsamudayapalepu@gmail.com

Y. Sushwanth
e-mail: sushwanth2580@gmail.com

P. Mohan
e-mail: poornimamohan@am.amrita.edu

N. Tejesh
e-mail: nalluritejesh@am.students.amrita.edu

M. Pavan
e-mail: mpavankumar@am.students.amrita.edu

used steganography and scrambling. The correct way of reversing the scrambled data will only be known to the intended receiver.

2 Related Work

In [1], the authors proposed a secure and efficient way of transmitting medical images using Discrete wavelet transform encryption which can work on JPEG2000 images. Paper [2] uses multiple layers of security for medical images in which first there is an encryption and the result of encryption is then embedded into a carrier image to obtain a stego image. In reference [12], a new approach for medical image encryption using the information provided by edgemap, is proposed. The method can be used for selective encryption of portions in the image or encryption of whole image. Reference [13] focuses on secure transmission of medical images over internet. The method proposed in [13] uses RC6, ECDSA and SHA256 for key generation, encryption and hashing. [14] provides a comparative study of RSA and AES algorithms in the context of medical image security. Authors of [15] have used a combination of RSA and AES algorithms for medical image security.

Reference [6] describes a novel encryption method for medical images which can be applied on color and gray scale images. The input image is divided into smaller blocks, which are then scrambled and to diffuse the scrambled image, keys are generated using chaotic maps. The algorithm is proved to outperform many existing ones. [7] proposes a new method for medical image encryption. The image is first high pass filtered and a combination of modified Arnold's cat map and Advanced Encryption Standard algorithm is used for encryption of information. The proposed approach has good encryption capability and it is claimed to have lesser computational cost.

The method proposed in [8] uses two permutation methods for medical image security. The method is proved to have better efficiency. The approach proposed in [9] does double compression using DCT and arithmetic coding on medical images and finally the data is encrypted using chaotic sequences. The method proposed in [10] uses Discrete Haar Wavelet transform for compression and chaos based DNA cryptography for encryption of medical images. The proposed algorithm in [11], is a combination of a data hiding technique and an encryption method which can provide required security for medical images.

In reference [3], the authors have analyzed and tested various Wavelets in the context of compression of samples of sounds generated from different musical instruments referring to SNR and distortion and came to the conclusion that db2 wavelet at a level 4 transformation is producing the best compression ratio with less degradation. In paper [4], a 9/7 filter is used to make changes in DWT to decrease the computational complexities and increase the performance. Paper [5] describes a unique way of using steganography for hiding image in a carrier as well as image compression for SAR images. They have used compression methods which compresses the data to 50 or 75%.

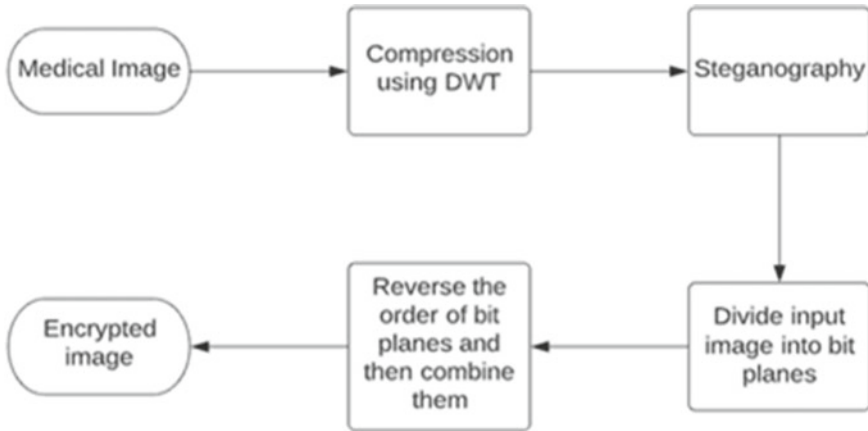


Fig. 1 Proposed method’s block diagram

Reference [16] describes a scheme for watermarking in the context of image security, which employs CNN. Reference [17] focuses on enhancement of robustness in medical image watermarking using hybrid transform.

3 Proposed Work

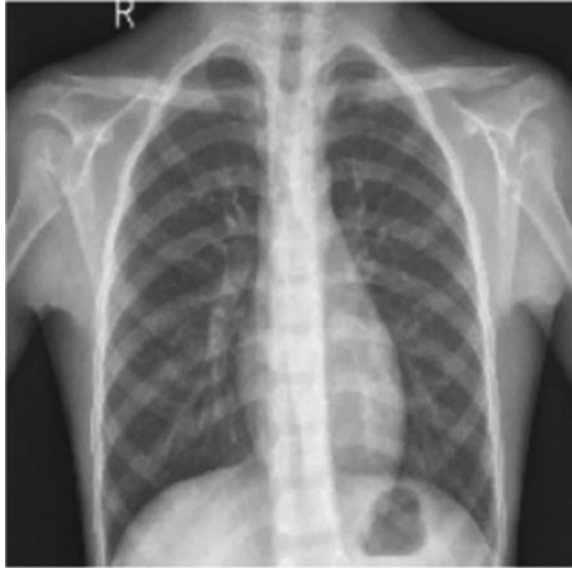
Figure 1 shows the detailed flow graph of how the method works. First, we will take a gray scale medical image and apply DWT to it. We get a set of coefficients as the result.

Then we embed the non-zero DWT coefficients into the carrier image using LSB (least significant bits) Steganography. After that we reversed the order of the bit planes resulting in a scrambled image which can be transmitted through any given network and to the intended receiver the locations of non zero DWT coefficients can be separately shared which can help him/her in reconstructing the image.

4 Materials

4.1 Discrete Wavelet Transform (DWT)

DWT transforms the given input to a series of coefficients which indicates the presence or absence of different frequency bands in the input. The Resulting coefficients after the transformation have values which are mostly zeros or close to zeros and absence of which doesn’t effect the output image when we reconstruct and in this

Fig. 2 Original image

way we can achieve compression too without losing the information that would affect the quality of reconstructed image.

In our project we used the Daubechies wavelet. It is a set of independent functions defining DWT.

Figure 2 is the X-Ray image for which DWT is computed, and Fig. 3 is the image reconstructed using 10% of the DWT coefficients. The result is not satisfactory as lot of information that is significant is being lost.

Similarly, Fig. 4 is the image reconstructed using 50% of the DWT coefficients and result as we can see is satisfactory since there is no loss of information that is significant.

4.2 Least Significant Bits (LSB) Steganography

The process of LSB steganography is that selected least significant bits corresponding to every pixel intensities in the cover image is replaced by data bits in the data to be hidden in series to obtain the output. Depending on the amount of data to be hidden, we can replace one or a set of LSBs in the input image.

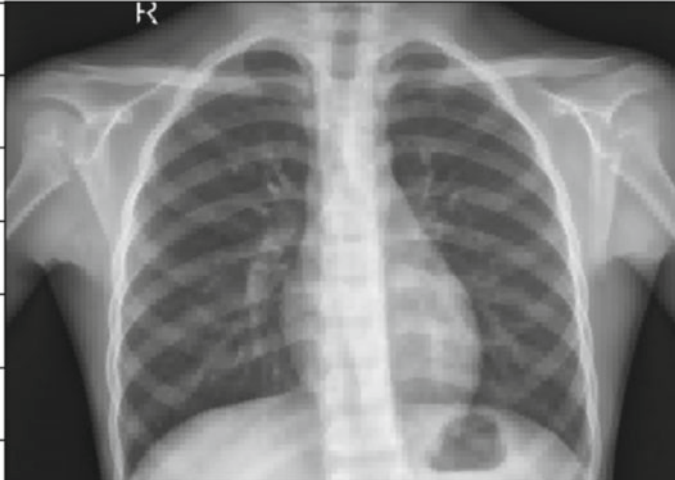


Fig. 3 Reconstructed image using 10% of DWT coefficients

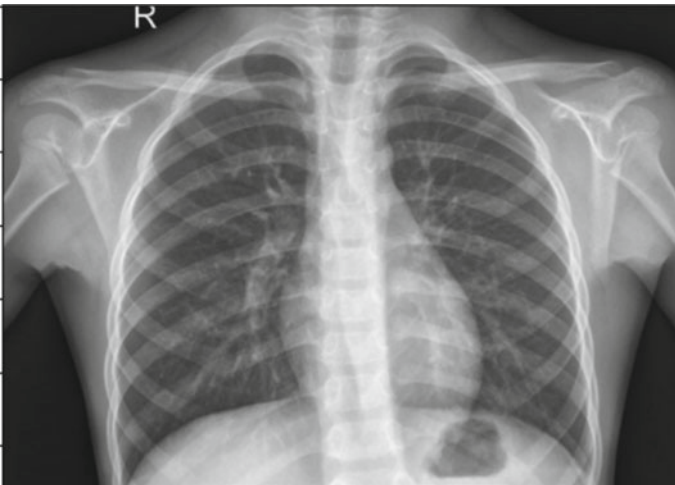


Fig. 4 Reconstructed image using 50% of DWT coefficients

4.3 Bit-Plane Reversal

When we consider a gray scaled 8-bit image, each pixel is encoded as 1 byte in a computer, for example 0 -> 00,000,000 and 255 -> 11,111,111. Bits in the extreme left are known as most significant bits (MSB), any changes made to this bits changes the contents of image and bits that are at the extreme right are known as least significant bits (LSB) and changes to this bits will not effect the contents of the image.

6	7	6	6	7
0	0	0	1	2
1	1	1	2	3
4	5	5	4	2
6	6	6	7	7

Fig. 5 Three-bit data in represented in pixel value form

Fig. 6 Three-bit data in represented in binary form

110	111	110	110	111
000	000	000	001	010
001	001	001	010	011
100	101	101	100	010
110	110	110	111	111

1	1	1	1	1
0	0	0	0	0
0	0	0	0	0
1	1	1	1	0
1	1	1	1	1

MSB plane

1	1	1	1	1
0	0	0	0	1
0	0	0	1	1
0	0	0	0	1
1	1	1	1	1

Centre bit plane

0	1	0	0	1
0	0	0	1	0
1	1	1	0	1
0	1	1	0	0
0	0	0	1	1

LSB plane

Fig. 7 Splitting the bit planes

We will have a MSB plane, an LSB plane and 6 other bit planes for an 8 bit image. If we reverse the bit planes we can see that the LSB plane would become MSB plane. We built a scrambled image by reversing the bit planes in the image.

Lets take an example of 3 bit image, first we need to represent the grayscale image in its pixel value form as shown in Fig. 5, then we need to convert the image into binary form as shown in Fig. 6, then we used python bit slicing technique to split the bit planes as shown in Fig. 7 and used array manipulation technique in python to reverse the order of bit planes.

5 Results and Conclusion

We took an X ray image shown in Fig. 8, calculated it's DWT coefficients and embedded the quantized non zero coefficients in the LSBs of cover image to get stego image. When we apply discrete wavelet transform to an image of size $m \times m$, it would result in the coefficient matrix of size $m \times m$. As per the standard steganography method we have to embed the X-ray image bits in the LSB of the cover photo. If we consider an 8-bit image of size 200×200 .

The total number of bits we have to embed for perfect reconstruction is 320,000 bits. But now after applying DWT to the image the coefficients can be quantified and thresholded to cut down the insignificant coefficient. After this the coefficients will



Fig. 8 X ray image

be converted to binary format. Now we can see the difference in the number of bits which are embedded.

For, Threshold = 80% (i.e. allowing 80 percentage of coefficients)

Number of bits which are to be embedded: 276,000

PSNR b/w the initial x-ray and the x-ray received at the receiver: 40.05 db

For, Threshold = 70% (i.e. allowing 70 percentage of coefficients)

Number of bits which are to be embedded: 239,750.

PSNR b/w the initial x-ray and the x-ray received at the receiver: 39.55 db.

For, Threshold = 60% (i.e. allowing 60 percentage of coefficients)

Number of bits which are to be embedded: 199,080.

PSNR b/w the initial x-ray and the x-ray received at the receiver: 38.75 db

Later the stego image is bitplane reversed and that can be used for secure transmission of data. The cover image, stego image and bitplane reversed stego image are shown in Figs. 9, 10 and 11.

At the receiver end, the data can be reconstructed by doing bitplane reversal of the image received, extraction of LSBs and inverse DWT. The reconstructed image

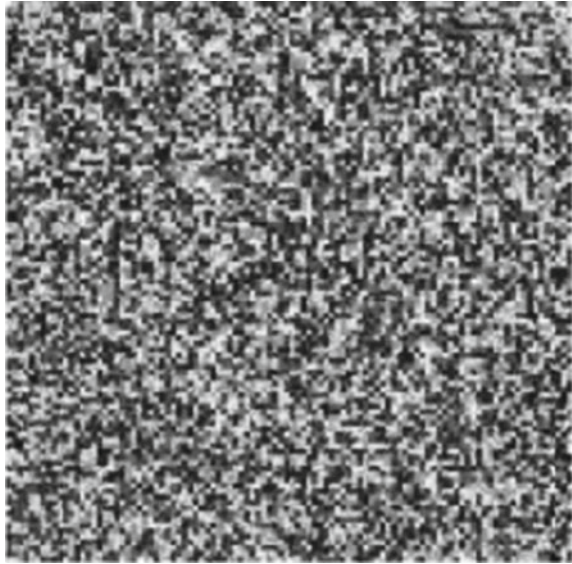


Fig. 9 Cover image



Fig. 10 Stego image

Fig. 11 Bit plane reversed stego image



is shown in Fig. 12. As a performance measure, we have calculated the PSNR value. We got 40.05 dB as the PSNR value. In this work we used JPEG medical images.



Fig. 12 Reconstructed X ray image

References

1. Abdmouleh SMK, Khalfallah A, Bouhleb MS (2017) A novel selective encryption DWT-based algorithm for medical images. In: 14th international conference on computer graphics, imaging and visualization, pp 79–84. <https://doi.org/10.1109/CGiV.2017.10>
2. Priyadarshini A, Umamaheswari R, Jayapandian N, Priyananci S (2021) Securing medical images using encryption and LSB steganography. In: 2021 international conference on advances in electrical, computing, communication and sustainable technologies (ICAECT), pp 1–5. <https://doi.org/10.1109/ICAECT49130.2021.9392396>
3. Sharma R, Kesarwani A, Mathur PM (2014) A novel compression algorithm using DWT. In: 2014 annual IEEE India conference (INDICON), pp 1–4. <https://doi.org/10.1109/INDICON.2014.7030363>
4. Rajasekhar V, Vaishnavi V, Koushik J, Thamarai M (2014) An efficient image compression technique using discrete wavelet transform (DWT). In: 2014 international conference on electronics and communication systems (ICECS), pp 1–4. <https://doi.org/10.1109/ECS.2014.6892826>
5. Jayachandran M, Manikandan J (2010) SAR image compression using steganography. In: 2010 international conference on advances in computer engineering, pp 203–206. <https://doi.org/10.1109/ACE.2010.15>
6. Kamal ST, Hosny KM, Elgindy TM, Darwish MM, Fouda MM (2021) A new image encryption algorithm for grey and color medical image. *IEEE Access* 9:37855–37865. <https://doi.org/10.1109/ACCESS.2021.3063237>
7. Shalaby MAW, Saleh MT, Elmahdy HN (2020) Enhanced Arnold's cat map-AES encryption technique for medical images. In: 2020 2nd novel intelligent and leading emerging sciences conference (NILES), pp 288–295. <https://doi.org/10.1109/NILES50944.2020.9257876>
8. Hasan MK et al (2021) Lightweight encryption technique to enhance medical image security on internet of medical things applications. *IEEE Access* 9:47731–47742. <https://doi.org/10.1109/ACCESS.2021.3061710>
9. Afandi TMK, Fandiantoro DH, Endroyono, Purnama IKE (2021) Medical images compression and encryption using DCT, arithmetic encoding and chaos-based encryption. In: 2021 international seminar on intelligent technology and its applications (ISITIA), pp 1–5. <https://doi.org/10.1109/ISITIA52817.2021.9502246>
10. Akkasaligar PT, Biradar S (2020) Medical image compression and encryption using chaos based DNA cryptography. In: 2020 IEEE Bangalore humanitarian technology conference (B-HTC), pp 1–5. <https://doi.org/10.1109/B-HTC50970.2020.9297928>
11. Abdel-Nabi H, Al-Haj A (2017) Medical imaging security using partial encryption and histogram shifting watermarking. In: 2017 8th international conference on information technology (ICIT), pp 802–807. <https://doi.org/10.1109/ICITECH.2017.8079950>
12. Zhou Y, Panetta K, Aagaian S (2009) A lossless encryption method for medical images using edge maps. In: 2009 annual international conference of the IEEE engineering in medicine and biology society, pp 3707–3710. <https://doi.org/10.1109/IEMBS.2009.5334799>
13. Nagarajan SM, Deverajan GG, Alshehri KUTMMD, Alkhalaf S (2021) Secure data transmission in internet of medical things using RES-256 algorithm. *IEEE Trans Ind Inform.* <https://doi.org/10.1109/TII.2021.3126119>
14. Kumar BJS, Roshni Raj VK, Nair A (2017) Comparative study on AES and RSA algorithm for medical images. In: 2017 international conference on communication and signal processing (ICCSP), pp 0501–0504. <https://doi.org/10.1109/ICCSP.2017.8286408>
15. Kumar BJS, Nair A, Raj VKR (2017) Hybridization of RSA and AES algorithms for authentication and confidentiality of medical images. In: 2017 international conference on communication and signal processing (ICCSP), pp 1057–1060. <https://doi.org/10.1109/ICCSP.2017.8286536>

16. Dhaya R (2021) Light weight CNN based robust image watermarking scheme for security. *J Inf Technol Digit World* 3(2):118–132
17. Manoharan JS (2013) A hybrid transform for robustness enhancement of watermarking in medical images. *Int J Imag Robot* 9(1):61–72

Deep Learning Based Hand Sign Recognition in the Context of Indian Greetings and Gestures



Rohan Saxena, Romy Garg, Bhoomi Gupta, and Narinder Kaur

Abstract Communication through sign language can be orchestrated in a variety of ways. There are certain words of the spoken language that can be directly represented and interpreted through simple gestures. Words like नमस्ते (Namaste) ॐ, बढ़िया (good) ॐ, बेकार (bad) ॐ etc. have designated signs or gestures. However, there are certain words that don't have predefined signs. Prior to the development of deep learning methodologies and algorithms, research in the field of sign language interpretation and translation was few. The most typical method for interpretation is to extract features from coordinated movements using Image Processing Algorithms, then use Convolutional Neural Networks to learn these features and increase utility. Advances in deep learning have led to the creation of Object Detection Algorithms that, when used in conjunction with neural networks, can identify all types of objects. You Only Look Once (YOLO) is one such algorithm that excels in identifying custom objects. It's utilized in conjunction with Darknet, a neural network architecture. People who use sign language frequently need to rely on a translation to get their message across to someone who does not understand sign language. Dependency on a translator can create issues and potentially render the person incapable of acting independently. The creation of a system that can help people use sign language without depending on another person can really help them be independent and ignite the confidence to present themselves to the world without any fear. Thus in a country like India, it is very important that we develop a hand gesture detection system that can identify Indian gestures.

Keywords Deep Learning · You Only Look Once · Indian Sign Language · DarkNet-53

R. Saxena · R. Garg · B. Gupta (✉) · N. Kaur
Maharaja Agrasen Institute of Technology, Rohini, Delhi 110086, India
e-mail: bhoomigupta@mait.ac.in

N. Kaur
e-mail: narinderkaur@mait.ac.in

1 Introduction

Deaf and mute persons use Sign Language to communicate. It allows people to share their views, ideas, and other information with the rest of the globe. The vocabulary, grammar, and lexicons of sign language are well-defined. British Sign Language (BSL), American Sign Language (ASL), Japanese Sign Language (JSL), Indian Sign Language (ISL), and other varieties of sign language exist based on geography and context of spoken language. The focus of this study is on Indian Sign Language (ISL), which is based on previous research in ASL.

1.1 Sign Language Translation

Gesture recognition is a small part of a high-level system (Vogler and Metaxas 2003) [1]. This larger system is composed of several sequential steps, each having its own associated fields of research. Each individual step lays out the foundation for its subsequent steps (Vogler and Metaxas 2003). Therefore, while each step can be explored independently, any advancements in one field, such as Object Detection (C. Zhang et al., 2006) [2] (Song, Chen, Huang, Hua, and Yan 2011) [3], Feature Extraction (Ren, He, Girshick, and Sun, 2015) [4], etc. can have a severe impact in all other remaining fields of research. The analogous field of research for this larger system is known as Sign Language Translation (Vogler & Metaxas, 2003).

1.2 Deep Learning

Deep Learning is basically a neural network that ideally has 3 or more layers. The emphasis is on duplicating the cognitive and decision-making abilities of the human mind, neurons to be specific, and creating a system that can work at the same level and capacity. It is a subgroup of Machine Learning. These neural networks try to copy the functioning of the human brain by training themselves (manually or automatically) using enormous amounts of dataset. The human brain is capable of observing intricate details about an object and learning from these subtle nuances to create a sense of understanding regarding the differences between multiple objects. Inference gathered and conclusions made regarding these objects are then stored in the memory. By using information stored in the memory, the brain can now detect the same object even in different settings. The objective of deep learning is to mimic a similar mechanism, create a network that can learn from the information provided to it and serve a purpose using all learned details. (Deng and Yu 2014) [5].

Multiple layers are recommended so that the machine can make more accurate predictions. Even though a single neural network is usable, it is not accurate enough.

Deep Learning (AI) has gained huge traction lately due to the fact that it eliminates human intervention making the process fast and precise.

1.3 Convolutional Neural Networks

CNN is one of the most advanced and suitable technologies that could have been used for hand gesture recognition. As the convolutional layer is where the majority of the computation takes place, it is the most important component of a CNN. It has three parts: input data, a filter, and a features map.

Multiple convolutional layers form the architecture of convolutional neural networks. This architecture is very similar to the pattern of neurons in the human brain and therefore the name “Neural Network”.

The algorithm of CNN works as follows: An image is taken as an input, it is accorded weights and biases relative to its quality, and then it is used as a dataset according to the requirements of a particular project.

1.4 Object Detection

One of the most prominent and flexible study areas in the field of computer vision (C. Zhang et al., 2006) is object detection.. It has become the building block for several real-world applications such as Object Tracking, Autonomous Driving, Face Detection, Video Surveillance, etc. Object Detection is the task of detecting a custom object in some form of graphical media such as an image, video, etc. (Wikimedia, 2008).

These images or videos can either contain multiple objects or a few objects at multiple locations. The task is not limited to just listing the different objects that are there. It also entails supplying information about the object’s location in the image. This is usually accomplished by providing the coordinates of the object in the image. Additional information may also include a bounding box that specifies the location as well as the probability with which the object was detected.

1.5 Neural Network Approaches

The traditional Convolutional Neural Network (CNN) uses a sliding window to search for an object in every possible position in the image. However, it is inefficient. This is due to the simple fact that different objects can exist in different sizes in the image and thus running the network with a predetermined sliding window size was slowing down the network exponentially.

RCNN's were one of the first attempts at improving traditional convolutional neural networks for the task of object detection. RCNN's improved the detection process by creating regions using a selective search algorithm. These regions were then passed through a CNN to classify the object in the image. Bounding boxes that create these regions are then modified using regression techniques (Girshick, Donahue, Darrell, & Malik, 2014). RCNN's were a great innovation when first introduced in the realm of object detection. However, they were unable to generate real time results. They are also largely dependent on the selective search algorithm (Uijlings, Van De Sande, Gevers, & Smeulders, 2013). It is observed that there is no process of learning at the stage of region creation (bounding box generation), thus there is a high probability of bad regions being generated.

Faster RCNN (Von Zitzewitz, 2017) (Ren et al., 2015) (Girshick et al., 2014) are an advancement from conventional RCNN's. While both methods follow a similar approach, Fast RCNN feeds the entire image as the input to the CNN as opposed to individual regions in conventional RCNN's. The images are used to generate convolutional feature maps. A separate network is used to locate regions of interest from the feature map. The regions of interest propose areas in the image where it is likely to detect an object. A Pooling layer serves the purpose of resizing regions, which are then fed to a hidden layer of the network that predicts the appropriate class of the detected objects. Faster RCNN are good at detection and provide improved accuracy over RCNN's but lack real-time application. In addition, they make use of a two-step network that is complex and may require individual training for each network making it computationally expensive.

You Only Look Once, YOLO (Redmon etc., 2016), Single Shot Detectors (Liu etc., 2016), etc. are a few examples of the approaches that outperform the networks listed above. For the purpose of this research, we will not dive into the details of any other framework other than YOLO and its different types.

1.6 You Only Look Once

The YOLOv3 architecture follows the same principles as its predecessors. The difference arises in its ability to provide a faster and more reliable prediction. This is achieved by making subtle changes to the existing YOLO9000 architecture. The most notable change is the addition of more layers is the convolutional neural network. The new feature extractor is created to act as a hybrid that encapsulates residual network architectures onto the YOLO framework. The network is larger than Darknet-19 but has several significant shortcut connections. The new network is termed as Darknet-53. Figure 2 (Redmon & Farhadi, 2018) shows the structure of Darknet-53. The system is said to perform as efficiently as other approaches and has the added advantage of being faster.

	Type	Filters	Size	Output
	Convolutional	32	3 × 3	256 × 256
	Convolutional	64	3 × 3 / 2	128 × 128
1×	Convolutional	32	1 × 1	
	Convolutional	64	3 × 3	
	Residual			128 × 128
	Convolutional	128	3 × 3 / 2	64 × 64
2×	Convolutional	64	1 × 1	
	Convolutional	128	3 × 3	
	Residual			64 × 64
	Convolutional	256	3 × 3 / 2	32 × 32
8×	Convolutional	128	1 × 1	
	Convolutional	256	3 × 3	
	Residual			32 × 32
	Convolutional	512	3 × 3 / 2	16 × 16
8×	Convolutional	256	1 × 1	
	Convolutional	512	3 × 3	
	Residual			16 × 16
	Convolutional	1024	3 × 3 / 2	8 × 8
4×	Convolutional	512	1 × 1	
	Convolutional	1024	3 × 3	
	Residual			8 × 8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Fig. 1 Structure of DarkNet-53

2 Related Work

This is an extension on the work done by us in minor project work as cited [6]. In this paper, we have covered the improvement in YOLO technology over the last 5 to 10 years. Also we have looked at some papers that deal with sign language translators. As we have taken inspiration from research done on American Sign Language (ASL) [7] to incorporate our learnings from Indian Sign Language (ISL). Thus it becomes necessary to give an account of research done on both the languages with major emphasis being on ISL.

In their study work, Sirshendu H., Sankhadeep C., V. Santhi, Nilanjan D., Amira S. Ashour, Valentina Emilia Balas, and Fuqian Shi(2017) [8] discuss three strategies for recognising Indian Sign Language motions. They created the NN-GA, NN-EA, and NN-PSO approaches by combining neural networks (NN) with genetic algorithms (GA), evolutionary algorithms (EA), and particle swarm optimization (PSO).

J. L. Raheja, A. Mishra, and A. Chaudhary (2016) [9] attempted to detect Indian Sign Language motions in real time using dynamic hand gesture recognition algorithms. To record motions, they employed high-resolution videos. For pre-processing, the video was transformed to HSV color space, and then gesture segmentation was performed using skin pixels. To construct the dataset and detect the gestures, the Support Vector Machine (SVM) was employed.

Joyeeta Singha, Karen Das (2013) [10] have tried to classify Indian Sign Language gestures using Euclidean distance based classification technique with weighted EigenValue. Their approach has four stages, according to their study paper: skin filtering, hand cropping, feature extraction, and classification.

Divya Deora, Nikesh Bajaj (2012) [11] Their article outlines a framework for a human-computer interaction that can comprehend Indian Sign Language motions (ISL). One of the key findings of it is that the complexity of recognizing gestures increases as both hands are involved. Therefore, segregation becomes difficult.

H. Muthu Mariappan, V. Gomathi (2019) [12] have employed machine learning algorithms for Indian Sign Language gesture recognition. The skin segmentation function of openCV is used to identify and track Regions of Interest (RoI). Hand gestures are trained and predicted using fuzzy c-means clustering machine learning methods.

Using the YOLO (You Only Look Once) method, Bhavadharshini M, Josephine Rachael J, Kamali M, and Sankar S. (2021) [13] attempted to create a system that can recognise American Sign Language motions in real time. The programme starts with data collecting and then moves on to pre-processing gestures before tracing hand movement with a combinatorial algorithm.

Rajesh KM, Dasri Hema Teja Anirudh B, Ghali S. Venkata Sai Yashwanth A. (2021) [14] have performed image classification, image localisation and image detection with the aid of YOLOv3. They were able to identify multiple objects in a single frame by employing multiple neural networks.

V. Adithya, R. Rajesh (2020) [15] has created a video dataset of various gestures in Indian Sign Language. The classification of gestures was done using both support vector machines and deep learning. Hand gestures of people aged 22–26 were used for collecting the data. This has helped in creating a database which is much needed for any research work.

Anup N., Jay Shankar Prasad, Soumik M., Pavan C. and G. C. Nandy (2010) [16] created a database using pre-existing data of ISL. Because of its attraction for illumination and orientation invariance, they employed direction histogram for categorization. Euclidean distance and KNN were also used.

3 Technologies

3.1 Python

Python is a programming language that is flexible, simple and has reliable tools required to create modern software that distinguishes it from other programming languages. Python is appropriate for machine learning as it is consistent and anchored on simplicity. Python is the greatest programming language for machine learning because of its independent platform and popularity among programmers.

Python libraries used are: CV2, Numpy, Mediapipe, Tensorflow.

4 Methodology

4.1 Algorithm

Following steps were involved while designing the algorithm:

- Import necessary packages
- Initialize mediapipe
- Load the gesture recognizer model
- Load class names
- Initialize the webcam
- Read each frame from the webcam
- Flip the frame vertically
- Get hand landmark prediction
- Print (result)
- Post process the result
- Print (id, lm)
- Drawing landmarks on frames
- Predict gesture
- Print (prediction)
- Show the prediction on the frame
- Show the final output
- Release the webcam and destroy all active windows

We have tried to recognize greetings and gestures of Indian Sign Language by following the above mentioned algorithm.



Fig. 2 Elaborating the results compiled

4.2 Dataset

This dataset contains around 2400 photos of ten distinct gestures. Each directory contains around 200 photos for training purposes and approximately 40 images for testing reasons. This dataset is mostly used for recognition of hand gestures.

4.3 Architecture

The YOLO algorithm focuses on only looking at an input image once. The network is attempting to extract features from the image in order to produce a corresponding feature map that can be used to identify items in a given image when it is being inspected. The network builds confidence in predicting items and their classes in an image as the feature map is filled with more information about the subtle subtleties of each object. A traditional Convolutional Neural Network is used for feature extraction.

The network consists of several convolution layers that help in transforming the input image. The architecture (network pipeline) is shown in Fig. 1 (Redmon et al., 2016). At the end of the pipeline are two fully connected network layers that are responsible for predicting bounding box coordinates and object class probabilities. The architecture can be modified as per user requirement. This is usually performed by altering the number of identified objects, the number of training batches, the bounding box anchors, and so on.

5 Result

Our model recognized various signs through hand gestures, but it couldn't recognize when the object angle changed. It recognized various symbols like बढिया(good) ☺, बेकार(bad) ☹, Call me, Fist, Peace, Rock and Smile with good accuracy.

The accuracy for नमस्ते(Namaste) 🙏 was not upto the mark, the algorithm was giving output based upon single hand recognition. The single hand was also recognised as Namaste.

6 Conclusion

The study focused on creating and proposing a model that could accurately and precisely predict the occurrence of an Indian sign language gestures like नमस्ते(Namaste) 🙏, बढिया(good) ☺, बेकार(bad) ☹ etc. using the You Only Look Once (YOLOv3) Algorithm.

7 Future Scope

We will feed more data to our model to make it detect Indian Sign Language movements with both hands, as our model is unable to do so.

There is still opportunity for improvement in the dataset. The precision of the algorithms will improve as the dataset is larger. As a result, more data will undoubtedly improve the model's accuracy in the future, as well as provide a method for recognising Indian Sign Language (ISL) [17].

References

1. Vogler C, Metaxas D (2004) Handshapes and movements: multiple-channel american sign language recognition. In: Camurri A, Volpe G (eds) *Gesture-based communication in human-computer interaction*. GW 2003. Lecture notes in computer science, vol 2915. Springer, Heidelberg. https://doi.org/10.1007/978-3-540-24598-8_23
2. Zhang Y, Kiselewich SJ, Bauson WA, Hammoud R (2006) Robust moving object detection at distance in the visible spectrum and beyond using a moving camera. In: *2006 conference on computer vision and pattern recognition workshop (CVPRW'06)*, pp 131–131. <https://doi.org/10.1109/CVPRW.2006.174>.
3. Song Z, Chen Q, Huang Z, Hua Y, Yan S (2011) Contextualizing object detection and classification. In: *CVPR 2011*, pp 1585–1592. <https://doi.org/10.1109/CVPR.2011.5995330>
4. Ren S, He K, Girshick R, Sun J (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. In: *Advances in neural information processing systems*, vol 28. <https://proceedings.neurips.cc/paper/2015/hash/14bfa6bb14875e45bba028a21ed38046-Abstract.html>
5. Li D, Yu D (2014) Deep learning: methods and applications. *Found Trends Signal Process* 7(3–4):197–387. <https://doi.org/10.1561/20000000039>
6. Saxena R, Garg R, Gupta B, Kaur N (2022) Hand gesture recognition. SSRN: <https://ssrn.com/abstract=4027056> or <https://doi.org/10.2139/ssrn.4027056>
7. American Sign Language. <https://www.nidcd.nih.gov/health/american-sign-language>
8. Hore S, et al (2017) Indian sign language recognition using optimized neural networks. In: Balas V, Jain L, Zhao X (eds) *Information technology and intelligent transportation systems*. *Advances in intelligent systems and computing*, vol 455. Springer, Cham. https://doi.org/10.1007/978-3-319-38771-0_54
9. Raheja JL, Mishra A, Chaudhary A (2016) Indian sign language recognition using SVM. *Pattern Recogn Image Anal* 26:434–441. <https://doi.org/10.1134/S1054661816020164>
10. Singha J, Das K (2013) Indian sign language recognition using eigen value weighted Euclidean distance based classification technique. arXiv preprint [arXiv:1303.0634](https://arxiv.org/abs/1303.0634). <https://arxiv.org/abs/1303.0634>
11. Deora D, Bajaj N (2012) Indian sign language recognition. In: *2012 1st international conference on emerging technology trends in electronics, communication & networking*, pp 1–5. <https://doi.org/10.1109/ET2ECN.2012.6470093>
12. Mariappan HM, Gomathi V (2019) Real-time recognition of Indian sign language. In: *2019 international conference on computational intelligence in data science (ICCIDS)*, pp 1–6. <https://doi.org/10.1109/ICCIDS.2019.8862125>.
13. Bhavadharshini M, Racheal JJ, Kamali M, Sankar S (2021) *Advances in parallel computing technologies and applications*, vol 40, pp 159–166. <https://doi.org/10.3233/APC210136>

14. Megalingam RK, Babu DHTA, Sriram G, Avvari VSY (2021) Concurrent detection and identification of multiple objects using YOLO algorithm. In: 2021 XXIII symposium on image, signal processing and artificial vision (STSIVA), pp 1–6. <https://doi.org/10.1109/STSIVA53688.2021.9592012>
15. Adithya V, Rajesh R (2020) Hand gestures for emergency situations: a video dataset based on words from Indian sign language. Data in Brief 31, 106016. ISSN 2352–3409. <https://doi.org/10.1016/j.dib.2020.106016>
16. Nandy A, Prasad JS, Mondal S, Chakraborty P, Nandi GC (2010) recognition of isolated Indian sign language gesture in real time. In: Information processing and management. BAIP 2010. Communications in computer and information science, vol 70. Springer, Heidelberg. https://doi.org/10.1007/978-3-642-12214-9_18
17. Indian Technical Sign Language Dictionary. <https://indiansignlanguage.org>

Pedestrian Detection Using MobileNetV2 Based Mask R-CNN



Sonal Sahu, Satya Prakash Sahu, and Deepak Kumar Dewangan

Abstract Pedestrian detection system is one of the recent technological innovations to save human lives on the roadways. According to WHO, road accidents highly contribute to the increasing mortality. These traffic accidents can be avoided by utilizing an autonomous vehicle equipped with AI algorithms to identify the pedestrians efficiently. Advances in computer vision and deep learning techniques open up new research possibilities. This research study has introduced novel algorithms to combine two of the most efficient deep learning models. It is well-known that the Mask R-CNN is the most efficient deep learning model used for performing object detection in two stages. This process leverages high accuracy but it also include limitations such as low detection speed and high computational cost. To resolve this problem a lighter version of Mask R-CNN with MobileNetV2 architecture has been developed. In order to make Mask R-CNN light, some modifications have been done in Region Proposal Network (RPN) like a Convolution Operation is replaced by the Depthwise Separable Convolution Operation. To further speed up the process, MobileNetV2 architecture is used in the place of ResNet-101. MobileNetV2 uses inverted residual block and linear bottleneck to generate less number of parameters and reduced network computation to save time and speed up the process. The main goal of the proposed model is to detect the pedestrian with high accuracy and at high speed by consuming less computational cost without compromising robustness of the system. Further, the experimentation has been carried out with INRIA dataset and recorded a 98.9% detection rate, 0.87 mAP and 0.85 mIoU, which is far better than the standard Mask R-CNN with ResNet-101 architecture. The proposed model's weight is also 65% lighter than the conventional model, allowing it to operate faster and spend less time in interpretation. The performance of ResNet-101 based Mask

S. Sahu (✉) · S. P. Sahu

Department of Information Technology, National Institute of Technology, Raipur, India
e-mail: sonal.sahu96@gmail.com

S. P. Sahu

e-mail: spsahu.it@nitrr.ac.in

D. K. Dewangan

Department of Computer Science and Engineering, ITER, SOA University, Bhubaneswar, India
e-mail: deepakdewangan@soa.ac.in

R-CNN and MobileNetV2 based Mask R-CNN were compared in this study. This model can be expanded in future to yield more accurate findings and improve the ability to deal with emerging challenges in smart vehicles.

Keywords Mask R-CNN · MobileNetV2 · Pedestrian detection · Autonomous vehicle · Deep learning · Segmentation

1 Introduction

Autonomous vehicles have replaced human drivers with AI-based technologies which generate a lower error rate than the human brain. In general, we all know that the human brain has constraints, such as the fact that the human brain can only work for a fixed length of time without any time interval, whereas AI-based systems have no such restrictions until there is a power or hardware failure. Drivers may become fatigue, intoxicated, or distracted while driving, and make a poor judgment in a particular circumstance, resulting in a road accident. To reduce traffic accidents, researchers have designed and developed a pedestrian detection system, which never fails to make a proper judgment at the right time. The pedestrian detection system is considered as the backbone of self-driving AI cars or enhanced video surveillance systems. An AI based vehicle detection system needs a camera, RADAR or LiDAR to visualize the surroundings and capture it as input and pass it to the detection model for further operation. If the autonomous vehicle uses a camera, it records video of the surroundings and extracts the images from it and passes it to the detection algorithm for the further process. In the further process those pedestrians are detected who are present in that visible range of camera. If the pedestrian is detected by the system, it will alert the driver or give the appropriate instruction to ensure safe driving. LiDAR embedded systems use light and RADAR embedded systems use waves to detect the object and measure the distance between vehicle and pedestrian based on the time consumed by the light or wave to come back after the bouncing from objects. Nowadays, camera is used only when the weather is clear and visibility is good enough for the camera. If the visibility is not enough due to bad weather, the system will use RADAR or LiDAR. Visibility is the most difficult work for the pedestrian detection system. There are many solutions for it like if the visibility is not enough through the camera then the sensor works well in those situations. Sometimes the pedestrian body temperature also helps to detect it because it is different from others and the temperature detector sensor easily detects.

Computer Vision is the most important sub-domain of artificial intelligence, which helps to train the computer to understand the real world. Machines can detect and recognize objects as humans from digital images. The main goal of the computer vision technology is to perform object detection in a simple way with high speed and accuracy. Nowadays, researchers are working on algorithms that can detect the objects just like humans or better than human. Some of the real-time applications of computer vision are autonomous cars, face detection, and object detection. The

integration of deep learning models in computer vision can achieve better performance and accuracy. The pedestrian detection system will initially classify the input image. Image Classification is the next process of computer vision that draws the boxes around each object and label them with the class. This step will basically categorize the objects present in the image into different groups. Object Localization is a process that identifies the location of the most visible object present in the image. Object classification and localization processes play a major role in any object detection system.

Many algorithms are available in the literature to detect pedestrians as shown in Fig. 1, they are Haar Cascade [2], HOG [1], YOLO [6], SSD [7], and R-CNN [8]. All of these algorithms have their own categories, such as Haar Cascade and HOG, which are the traditional models in which object characteristics must be manually entered into the computer. However, the advanced models are self-training models. In the real world, the traditional paradigm remain inefficient. Recently developed models such as YOLO and SSD are one step models in which the detection process is completed in a single step. Likewise, R-CNN is a recent model, but it detects objects in two phases. In the first stage, it retrieves the existing feature from the image, and in the second stage, the particular objects are classified. This paper has proposed a two stage model to utilize the advantage of two stage models and attempted to remove the limitation with its heavy weight model by modifying the RPN of Mask R-CNN and speedup the MobileNetV2, which is used in the place of VGG 16 or ResNet-101.

The proposed model is a combination of two efficient deep learning models. First is MobileNetV2, which is best-suited for the mobile application and embedded vision application. It was developed by Google researchers in 2017 to power the next generation mobile. MobileNetV2 is a convolutional neural network that is 53 layers deep. It is a pre-trained network that can easily classify images into more than 1000 object categories. The biggest advantage of using MobileNetV2 is it reduces their parameters, which means less number of calculations and this will speed up the proposed model.

MobileNetV2 establishes a residual connection to prevent the deep convolutional layer from vanishing gradient problem. It has a depthwise convolutional layer and pointwise convolutional layer to assist the model in working pixel wise information to provide the accurate result. In the incoming section we will show how these layers are working together to give the best result. The second component of the proposed model is Lighter version of mask r-CNN. Lighter version of mask R-CNN means



Fig. 1 An illustration of the proposed model for the pedestrian detection system

doing some modification in RPN so that it becomes lighter than the standard one without compromising the performance and accuracy. This may be accomplished by substituting the spatial convolution operation with 512 convolution kernel and 3×3 RPN filter with depthwise convolution operation with 256 convolution kernel and 3×3 filter. ReLU6 is the activation function used in the proposed model.

2 Related Works

Pedestrian detection system detects and locates the pedestrians from the image or live video and alerts the driver or takes the system appropriate action. In recent years, researchers have attempted to develop algorithms that can be implemented in real time. Nowadays many algorithms are designed and developed to leverage efficient results. These algorithms are firstly divided into two categories, first is traditional algorithms like HOG features [1], Haar features [2], LBP features [3], LUV features [4]. These algorithms are the manual design models, which represent the attribute of pedestrians based on the low-level features. After that, a classifier like support vector machine [5] or decision tree is employed to detect the pedestrian in the image. These models take lots of time to train the model manual for each type of object and this is why these models are not used nowadays. Traditional methods need to incorporate artificially complex features to solve the emerging challenges and combat the challenges faced in the process of pedestrian detection. Deep learning techniques are also employed in the real-time applications like human activity recognition [26], parking slot allotment using predictive control strategy [27], and traffic light cycle control using deep reinforcement techniques [29]. Different challenges will emerge in vision-based intelligent vehicle systems [23].

All the deep learning based algorithms are included in this category. Modern methods are more efficient than the traditional methods and it also overcomes the existing limitations. Traditional methods are considered as the manual design models, which consumes more time to train the model for each object. Deep learning methods are indeed a self-learning paradigm, which justifies its enormous success in computer vision applications. Further, deep learning-based approaches are separated into two parts: Single stage detectors such as YOLO [6], SSD [7], etc. and two stage detectors such as Fast R-CNN [9], Faster R-CNN [10], Mask R-CNN [11], R-FCN [12], etc. A single stage detector model is the one that requires a single run through the neural network to identify all objects and create bounding boxes all at once. YOLO model interprets the image in real-time at a speed of 45 frames per second. After some modifications, YOLO increased it to 155 frames per second [6]. The major drawback of YOLO is the generation of localization error but it predicts less false positives in the background. YOLO uses two fully connected layers but SSD uses multiple convolutional layers of different sizes. In SSD, the Lightweight layer present in a neural network may not generate enough high level features to perform prediction for small objects but YOLO overcomes this problem. YOLO is also suitable for performing small object detection.

Two stage detector also called as Region proposal based algorithms is more efficient than all other algorithms employed in object detection. The different categories of algorithms are: R-CNN, Fast R-CNN, Faster R-CNN, Mask R-CNN, R-FCN, etc. These algorithms leverage high accuracy, low error rate and high speed. Despite the hype, the main limitations observed in these algorithms include high computation cost and high complexity. Proposed in 2014, R-CNN [10] is the first algorithm with high capability of object detection. In this paper, authors have explained the working of R-CNN in three parts. In the first part the model will generate the region proposal, which means the system will draw bounding boxes around each object present in the image. In the second part, feature extraction will be performed for each object region by using a fully convolution layer. In the third part, a final output will be generated by the fully connected layer. As an output, all the object gets detected and further indicated by including a bounding box around the object. The major limitations of R-CNN are its slow process and expensive training [10]. To overcome these drawbacks, researchers came up with a new framework called fast R-CNN. The fast R-CNN uses the concept of RoI pooling to enhance the accuracy and speedup the training process. Next framework included in this series is Faster R-CNN, which added the region proposal network concept to make the process faster. The Faster R-CNN is just an extension of Fast R-CNN. Furthermore, the Mask R-CNN was proposed by the researchers after achieving a huge success in the object detection task [12]. It has provided more accuracy than all other algorithms with high precision. Mask R-CNN also uses segmentation concepts to increase the visibility of detected objects. This algorithm has also come with some challenges such as high complexity and processing time. To overcome this limitation, this research study has proposed a lightweight Mask R-CNN and MobileNet V2. To make sure about the working of autonomous vehicle in real time, it should do many work automatically and to do so the researchers have proposed many deep learning based techniques like VNet (Vehicle Detection Network) [22], Road Detection[24], Lane Detection[25], vehicle motion prediction[28], Vehicle Detection[30], Vision Based Lane Region Detection Network (VLDNet) [31], Suspicious Human Activity Detection system[32], Image Generation Using GAN and its Classification using SVM and CNN[33], fusion technique for finger knuckle print recognition[34], finger knuckle print images using Gabor feature[35]. The literature survey is shown in the below Table 1.

Table 1 Literature survey

Ref No.	Author year	Proposed method	Inference
[11]	M. A. Malbog 2019	Mask R-CNN	High Accuracy achieved but the complexity is very high
[14]	Z. Zhao et al., 2021	Faster RCNN with MobileNetV2	Improve the ability to detect the pedestrian in a crowded background using SE-RPN
[15]	N. N. F Giron et al., 2020	Faster R-CNN with Inception	Feature extractor gives better accuracy but size and speed are not better
[16]	X. Li et al., 2020	SSD with Inception	This framework achieved good accuracy as well as good speed using sparse connection and multi-fusion but not efficiently in crowded streets
[17]	Y. XU et al., 2021	SSD with ResNet	Achieved 85% mAP which is 12% more than the base SSD model
[18]	N. Zhang et al., 2021	Lightweight-YOLOV3	YOLOv3 has high computational complexity. To overcome this problem, L1 Regularization is used on the batch normalization layer
[19]	M. A. Malbog et al., 2019	Mask R-CNN with ResNet-101	Vanish Gradient and exploding gradient problems are sorted by this model but the 100 deep layers make this model slower
[20]	L. Chen et al., 2021	Faster R-CNN with ResNet-50	The detection accuracy is average accuracy but less than the ResNet-101 backbone
[15]	N. N. F Giron et al., 2020	SSD with MobileNet V2	Reduce the Size of the model and achieve good speed but the accuracy is a bit low
[20]	L. Chen et al., 2021	SSD with MobileNet V1	This model is the one of the lightest model but slow than SSD with MobileNetV2

3 Methodology

The proposed model is composed of a combination of two most powerful algorithms such as Mask R-CNN Lite and MobileNetV2. The MobileNetV2 helps to speed up the tasks involved in the model whereas RPNLite makes a lighter version of Mask R-CNN to avoid the high complexity and computational cost. As depicted in Fig. 2,

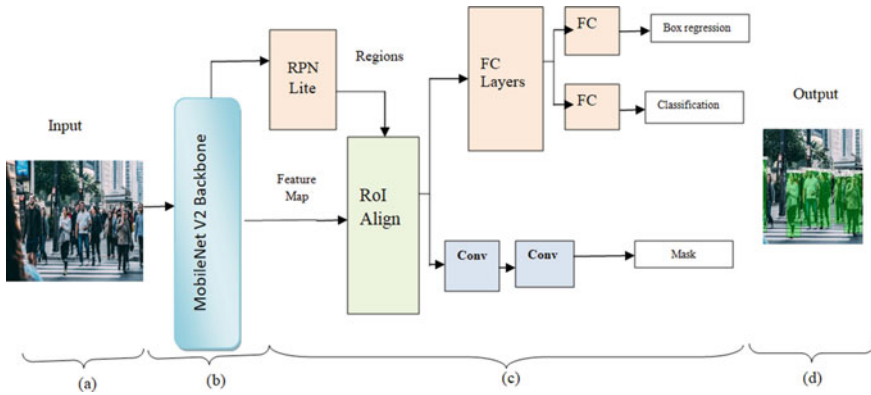


Fig. 2 Overall architecture of proposed model

the proposed architecture is divided into four parts. In the first part, the obtained video will be fed as an input. Then, the input data will be sent to a camera, which records the live video and extract images out of it will be sent to the system. The image will be further sent to the MobileNetV2 architecture to extract the feature map. This is considered as the second part of proposed architecture. The feature map will be then used by the Mask R-CNN to detect the pedestrian accurately in the third part. At the last part of architecture, the segmented pedestrian output will be obtained. In the further section we will elaborate each module in a detailed manner with the real time experimental results.

In general, our model has two major parts as we can see in the Fig. 2. And each part has different algorithms.

- A. MobileNet V2
- B. Mask R-CNN Lite

3.1 MobileNetV2

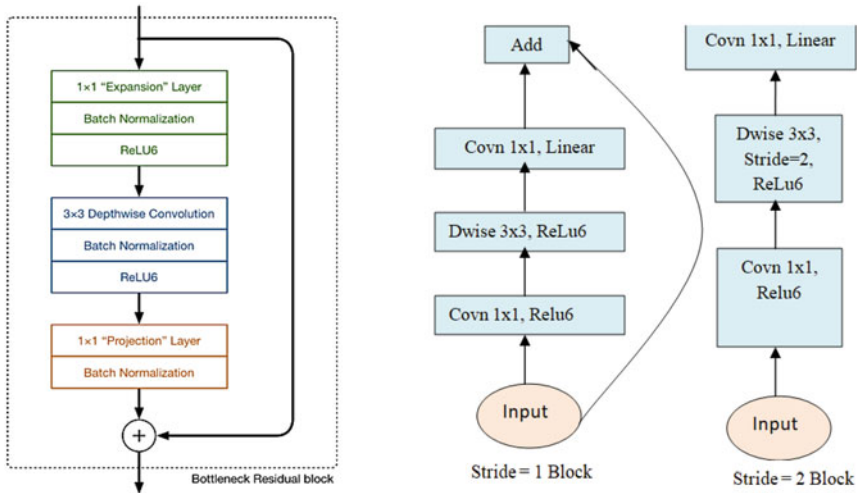
MobileNet V2 is a lightweight Convolutional Neural Network [CNN] with 53 deep convolutional layers. It works well on mobile applications and embedded systems due to its lightweight nature. It is a pre-trained network based on the image dataset, which can classify more than 1000 of categories easily. It was developed by Google in 2017 to get better accuracy in object detection tasks. MobileNet V2 is based on inverted residual structure established between the bottleneck layers. MobileNetV2 introduces two most important concepts, first is depthwise separable convolution and second is linear bottlenecks. MobileNetV2 replaces the full convolution operator with a separate version, which divides the convolution into two different layers. The first layer is depthwise convolution, and the main task of this layer is to perform lightweight filtering by using a single convolutional filter on each input channel. The

second layer is pointwise convolution also known as 1×1 convolution, and the task of this layer is creating new features by computing linear combinations of the input channels. MobileNetV2 also makes use of Batch Normalization technique used between the intermediate convolution layers to help the input flow over the network. This technique prevents the layer from internal covariate shift which means when the changes happened in the input distribution due to this training process gets slow down. To avoid this kind of problem we used batch normalization technique after each layer. ReLU6 is used as an activation function here to prevent the information even at low-dimension. It has 6 as a maximum value of activation that helps in fixed point inference.

Figure 3(a), (b), (c) represents the internal functionality of MobileNetV2. As we can see in the Fig. 3(a) it is the bottleneck Residual block that is the key concept of mobileNetV2. The first layer of this block is 1×1 Expansion layer as its name suggests it does the same work it expands the number of channels in input data for the next layer which is Depthwise Convolution. That means this model can process the image easily, even if the quality of image is not good. Hence, the Expansion layer produces more output channels than the input channels. The Expansion factor decides how much data channel should be expanded and by default it is 6. This is a kind of hyperparameter for experimenting with different models. After expansion, batch normalization technique is used to make the output of the expansion layer into the standard format for the depthwise convolution layer. ReLU6 is used as an activation function to increase robustness. The next layer is a 3×3 Depthwise convolution layer that is used to filter the input. This layer also helps in reducing the computational cost at only a small cutting in accuracy. After this layer Batch Normalization is used then the ReLU6 activation function. At the layer layer we have a 1×1 pointwise convolution layer that is also called a projection layer. The task of this layer is to reduce the number of output channels and make it same as the number of input channels. It converts the high number of dimensions into a low number of dimensions. This is why we called it the bottleneck layer because it reduces the flow of data all over the network. After this we do use the batch normalization but we don't use the ReLU6 activation function. The reason behind not using activation function at last is to prevent the information loss. Experts say if we use ReLU on a low dimension channel it results in information loss. This is why we call it Linear Bottleneck Inverted Residual Structure. Inverted Residual connection is used to prevent the network from vanishing gradient problem. It helps in the flow of gradients all over the network when the number of input channels is the same as the number of output channels.

Let's take an example if 24 input channel is going into the Bottleneck block then expansion layer convert this tensor into $24 * 6 = 144$ channels. These channels will pass through the 3×3 depthwise convolution for the filter. The output of the previous layer will pass to the 1×1 pointwise convolution layer that project the outgoing channel in low number of dimension which is 24 channels again as we can see in the Fig. 4.

Inverted Residual refers to the connection between bottleneck layers. As we can see in the Fig. 4. The very first and last layers are the bottleneck means either these blocks expand or compress the channels. It just controls the flow of channels in the



(a): Bottleneck Residual Block.

(b): Types of blocks in MobileNetV2

Input	Operator	t	c	n	s
$224^2 \times 3$	conv2d	-	32	1	2
$112^2 \times 32$	bottleneck	1	16	1	1
$112^2 \times 16$	bottleneck	6	24	2	2
$56^2 \times 24$	bottleneck	6	32	3	2
$28^2 \times 32$	bottleneck	6	64	4	2
$14^2 \times 64$	bottleneck	6	96	3	1
$14^2 \times 96$	bottleneck	6	160	3	2
$7^2 \times 160$	bottleneck	6	320	1	1
$7^2 \times 320$	conv2d 1x1	-	1280	1	1

(c): Architecture of MobileNetV2.

Fig. 3 a Bottleneck residual block b Types of blocks in MobileNetV2 c Architecture of MobileNetV2

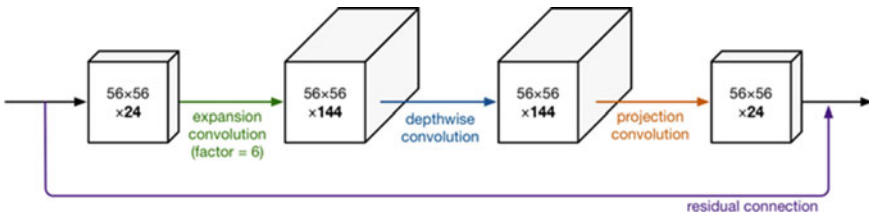


Fig. 4 Work flow diagram of bottleneck residual block

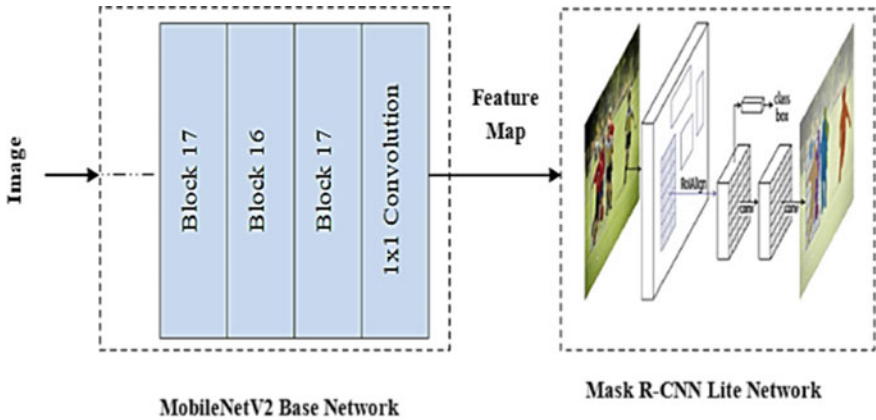
whole network. Standard Residual connection means connection between the layers with a high number of channels and inverted residual connection is just opposite of it.

MobileNetV2 has two kinds of block, first with stride = 1 and other is stride = 2 as Fig. 3(b) shows. Stride is the part of mobilenetv2 it helps in compression of image. ‘s’ indicates the Stride parameter of the neural network’s filter as we can see in the Fig. 3(c). In mobilenetv2 network we have four parameters like t indicates expansion factor, c indicates number of output channels, n represents repeating number of blocks and s is stride parameter. In Fig. 3(c) we have represented all the details of the network in tabular format. The 3×3 kernel is used for depthwise convolution and 1×1 pointwise convolution. The first layer of this model is 2D convolution layer and the task of this layer is to convert the input image into the tensor form for the next layer.

MobileNetV2 is the backbone of our proposed model that extracts the feature map from input and passes it to the next module of the model which is Lite Mask R-CNN. To extract the feature map MobileNetV2 has 52 deep convolution layers that help it. Last layer of mobilenetv2 has been removed because we only want a feature map and the last layer is used to classify the object which is not necessary here. Figure 5(a) represents the working flow of MobileNetV2. The first layer of mobilenetv2 algorithms is a 2D convolution layer that converts the input image into a specified size of image and passes it to the Bottleneck Residual block. We have already discussed this block in the upper section. We can see the architecture of this block in Fig. 3(a). Next to Bottleneck Residual block 16 the same block is present as we can see in the Fig. 5(a). After this we have a 1×1 convolution layer that converts all the output of the previous layer into the tensor for the next module of the proposed model. Figure 5(b) represents the input size at each convolution layer and what operation has been performed on it and what is the output size getting.

3.2 Proposed Mask R-CNN

Mask R-CNN is the most efficient and powerful model for object detection among all other deep learning models. This algorithm gives high accuracy with less false detection which means very less error rate and that makes this algorithm best among all. Mask R-CNN has the special feature which is segmentation that masks or color different the positive RoI with different color. This property increases the visibility for the user that can see clearly detected objects. That is the best thing about this algorithm. Mask R-CNN is a heavy weight model and it requires high memory space to execute. Mask R-CNN has high computation complexity because of the high number of parameters on the network. If the network contains a high number of parameters that means more calculation on each neuron this leads to slower processing. To overcome this problem in this paper we have proposed lightweight Mask R-CNN. To make the base Mask R-CNN light in weight we modified the RPN of the model. RPN is the Region Proposal Network of Mask R-CNN that helps the model to detect



(a): Work Flow of proposed Model.

Input	Operator	Output
$h \times w \times k$	1x1 conv2d, ReLU6	$h \times w \times (tk)$
$h \times w \times tk$	3x3 dwise s=s, ReLU6	$\frac{h}{s} \times \frac{w}{s} \times (tk)$
$\frac{h}{s} \times \frac{w}{s} \times tk$	linear 1x1 conv2d	$\frac{h}{s} \times \frac{w}{s} \times k'$

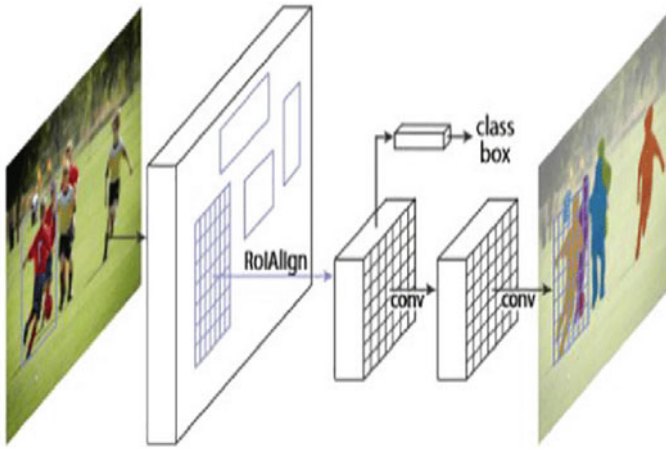
(b): Input Output of model.

Fig. 5 a Work flow of proposed model b Input output of model

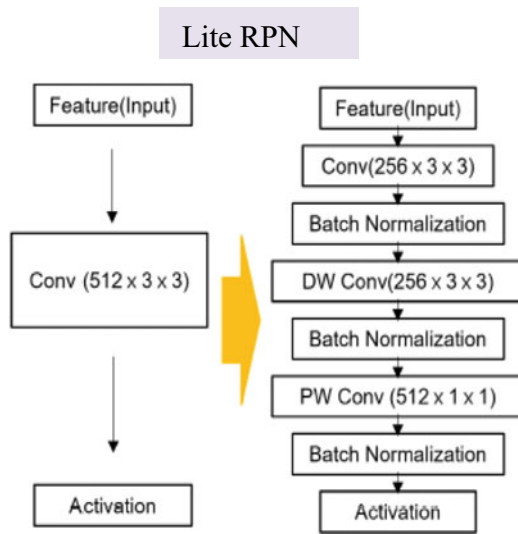
all possible areas where objects can be found. Figure 6(a) represents the architecture of Mask R-CNN and we have different sub-layers inside of it. The First layer is RPN and we have done some modification on this layer to make the all over model lightweight. In the next section we will see the detailed architecture of RPN.

Region Proposal Network (RPN)

Region Proposal Network takes the Input as a feature map from the backbone of the model and draws the bounding boxes around all the objects which are present in the image. In Fig. 6(b) we can see the modified RPN architecture of RPN. In the Base RPN we have convolution operation and we have replaced it with Depthwise and pointwise convolution operation to reduce the parameter. Batch normalization and activation function ReLU6 is used after each convolution layer as we can see in Fig. 6(b) Lite RPN.



(a): Architecture of Mask-RCNN.



(b): Architecture of Lightweight RPN.

Fig. 6 a Architecture of Mask-RCNN b Architecture of Lightweight RPN

Lite RPN takes the input from backbone network and passed to the spatial convolution with 256 kernels and with 3×3 filter size. After processing the output gets formatted into standard format for the next layer which is Depthwise convolution with 256 and with 3×3 filter size. After DW Conv batch normalization is applied on the output and send it to the next layer which is Pointwise Convolution layer. Pointwise convolution with 512 kernels and the filter size is 3×3 . At last batch normalization is used then ReLu6 activation function on the output. The final output will be passed to the RoI pooling for the next process. This is the whole process of Lite RPN that makes it light and accurate. Lite RPN makes the Mask-RCNN light in weight that leads this model to the less computation cost and less complexity with little difference in accuracy that can be neglected.

ROI Pooling

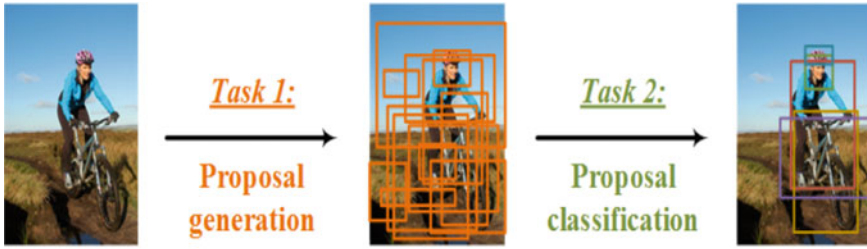
Region of Interest are the areas on the image where the necessary object can be found. RoI pooling is a technique in which we convert all the region proposals or feature map in fixed size for the further fully connected convolution layer. In our proposed model we have used MaxPooling algorithms with 7×7 window size to reduce the feature map. This technique select the maximum weight element from the particular region of feature map covered by the 7×7 filter. The output of this layer is the feature map that holds only high features of the last one feature map.

Classification and Segmentation

Classification and Segmentation is the last process of this model. After passing through the two fully connected convolution layer it will go for the classification and segmentation. Classification technique gives the category to the object from it belong. In this model we have used Softmax classifier that calculate the probability for each class type. After calculation we select those bounding boxes who has higher positive score for the next stage. Regression technique is also used here to refine or predict localization of object in the image. Segmentation is the next step which we will apply on the selected region this process also called masking. As we can see in the Fig. 7(a) and (b) that shows the output of Region proposal network and the final output.

4 Result and Experimentation

To validate the performance of the proposed model, the INRIA dataset is utilized. For the experiment on the INRIA dataset, 80% of input data is used for training and 20% of the input data is used for model testing and validation. Totally, 2000 images of pedestrians are collected. The custom dataset is divided into two parts 1800 pictures for training and 200 for validation. For the performance measurement, two types of accuracy are calculated: First is location accuracy and second is classification accuracy. Location prediction accuracy can be calculated based on the Intersection of Union (IOU), which means the ratio of predicted area and actual area of interest. The classification accuracy is calculated based on the precision, recall, mean and



(a): Region Proposal Network (RPN)



(b): Final Output

Fig. 7 a Region Proposal Network (RPN) b Final output

accuracy. Precision is also calculated by using a confusion matrix. Using Eq. (1) and (2), we have calculated the Precision, Recall and mAP.

$$Precision = \frac{TP}{TP + FP}, Recall = \frac{TP}{TP + FN} \tag{1}$$

$$mAP = \frac{\sum_{i=1}^N \int_0^1 P(r)dr}{N} \tag{2}$$

In the Eq. (1) TP indicated the results that comes when the model predict the positive category. TN indicates the results, where the model correctly forecast the negative category. FP indicates the result when the model wrongly forecast the positive category. FN indicates the result when model wrongly forecast the negative category.

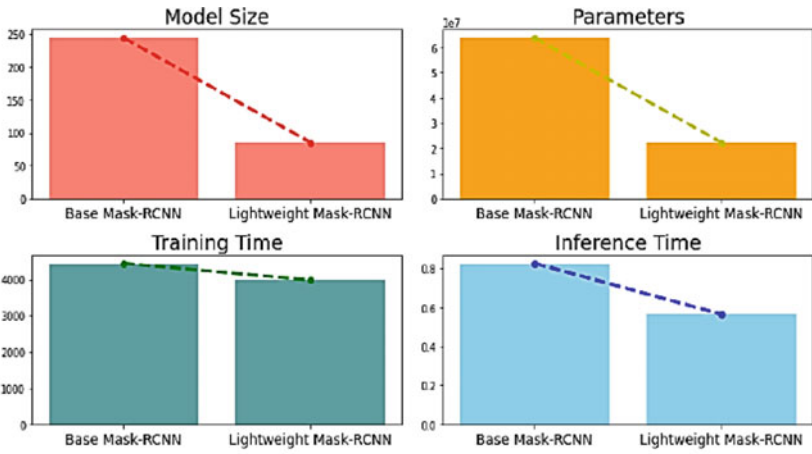
Table 2 Comparison of Standard Mask R-CNN and Lite Mask R-CNN

Types	Standard Mask R-CNN	Lite Mask R-CNN
Backbone Network	ResNet-101	MobileNetV2
No. of Backbone Parameters	44,706,276	3,538,980
No. of RPN Parameters	1,189,400	7,36,780

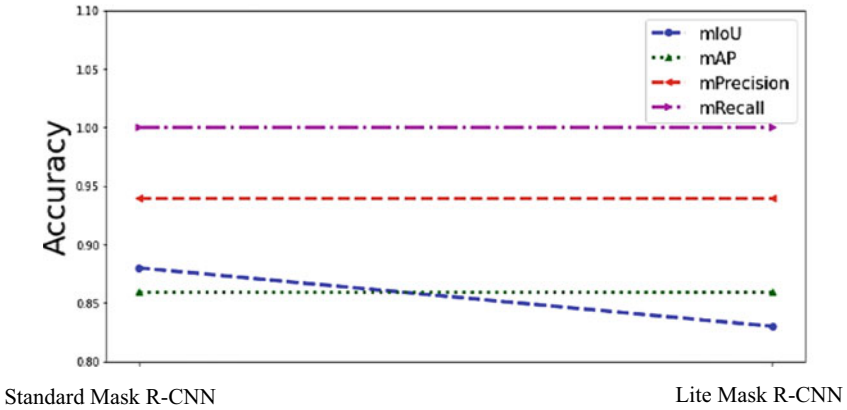
Table 3 Comparison of accuracy results

Types	Standard Mask R-CNN	Lite Mask R-CNN
mIoU	0.89	0.84
Precision	0.95	0.95
Recall	1.10	1.10
mAP	0.87	0.87

Table 2 has compared the proposed model with Mask R-CNN and ResNet-101 in terms of number of parameters in the network and number of parameters present at RPN. MobileNetV2 has fewer parameters compared to the ResNet-101. Due to modification in RPN, less number of parameters are obtained at RPN when compared to the base Mask R-CNN. Table 3 has compared the accuracy results between the Base Mask R-CNN and Lite R-CNN and revealed that the Lite Mask R-CNN has performed better than the base Mask R-CNN. Figure 8(a) represents the comparison between base Mask R-CNN and Lite Mask R-CNN in terms of model size, parameters, training time and inference time. Hence, Lite Mask R-CNN outperforms all other existing models. Further, the based Mask R-CNN and Lite Mask R-CNN are compared in terms of mAP, Precision, Recall and IoU and represented the result in the graph form in Fig. 8(b). Figure 9(a), (b) is the real time experiment screenshots.



(a): Comparison Graph.

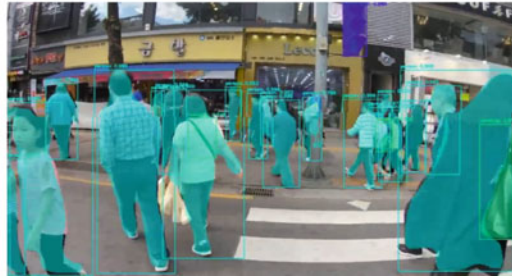


(b): Accuracy Comparison Graph.

Fig. 8 a Comparison graph b Accuracy comparison graph



(a): Real-Time Experiment output 1.



(b): Real-Time Experiment output 2.

Fig. 9 a Real-time experiment output 1 b Real-time experiment output 2

5 Conclusion

This research study has successfully analyzed the most effective and light weight model used for integrating pedestrian detection in autonomous vehicles. The proposed model has been successfully developed with the combination of two models, where each leverages best quality and combined both to get both the qualities in one frame to overcome the existing limitations. From the results, it is observed that the MobileNetV2 is a light weight model that works well on mobile devices but delivers low accuracy rate. By combined the MobileNetV2 with Mask R-CNN, the proposed hybrid model has attained high accuracy. Mask R-CNN is a very heavy weight model, hence the proposed study has modified the architecture by incorporating Region Proposal Network (RPN). After all these modifications made, the proposed model has become highly accurate and fast. Based on the experimentation results, the proposed model has outperformed the existing models and achieved 98.9% detection rate with 3.40% error rate. The proposed model is also compared the standard mask R-CNN with ResNet-101 Backbone and obtained less parameters at Backbone Network as well as at RPN and that makes the proposed model much faster and accurate. In the future, the proposed model can be modified to combat the future challenges and achieve a better accuracy.

References

1. Dalal N, Triggs B (2005) Histograms of oriented gradients for human detection. In: IEEE conference on computer vision and pattern recognition, pp 886–893
2. Viola P, Jones M (2003) Rapid object detection using a boosted cascade of simple features. In: IEEE conference on computer vision and pattern recognition, pp 511–518
3. Wang X, Han X, Yan S (2010) An HOG-LBP human detector with partial occlusion handling. In: IEEE conference on computer vision, pp 32–39
4. Dollár P, Tu Z, Perona P et al (2009) Integral channel features. In: British machine vision conference, pp 25–31
5. Felzen F, Girshick B, Mcallester D et al (2010) Object detection with discriminatively trained part-based models. *IEEE Trans Pattern Anal Mach Intell* 32(9):1627–1633
6. Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: unified, real-time object detection. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), pp 779–788. <https://doi.org/10.1109/CVPR.2016.91>
7. Liu S, Lv S, Zhang H, Gong J (2019) Pedestrian detection algorithm based on the improved SSD. In: 2019 Chinese control and decision conference (CCDC), pp 3559–3563. <https://doi.org/10.1109/CCDC.2019.8832518>
8. Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: 2014 IEEE conference on computer vision and pattern recognition, pp 580–587. <https://doi.org/10.1109/CVPR.2014.81>
9. Girshick R (2015) Fast R-CNN. In: 2015 IEEE international conference on computer vision (ICCV), pp 1440–1448. <https://doi.org/10.1109/ICCV.2015.169>
10. Ren S, He K, Girshick R, Sun J (2017) Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans Pattern Anal Mach Intell* 39(6):1137–1149. <https://doi.org/10.1109/TPAMI.2016.2577031>
11. He K, Gkioxari G, Dollár P, Girshick R (2020) Mask R-CNN. *IEEE Trans Pattern Anal Mach Intell* 42(2):386–397. <https://doi.org/10.1109/TPAMI.2018.2844175>
12. Kolarow A, Schenk K, Eisenbach M et al (2013) APFEL: the intelligent video analysis and surveillance system for assisting human operators. In: IEEE conference on advanced video and signal based surveillance, pp 195–201
13. Kadam K, Ahirrao S, Kotecha K, Sahu S (2021) Detection and localization of multiple image splicing using MobileNet V1. *IEEE Access* 9:162499–162519. <https://doi.org/10.1109/ACCESS.2021.3130342>
14. Zhao Z, Ma J, Ma C, Wang Y (2021) An improved faster R-CNN algorithm for pedestrian detection. In: 2021 11th international conference on information technology in medicine and education (ITME), pp 76–80. <https://doi.org/10.1109/ITME53901.2021.00026>
15. Giron NNF, Billones RKC, Fillone AM, Del Rosario JR, Bandala AA, Dadios EP (2020) Classification between pedestrians and motorcycles using Faster RCNN inception and SSD MobileNetv2. In: 2020 IEEE 12th international conference on humanoid, nanotechnology, information technology, communication and control, environment, and management (HNICEM), pp 1–6. <https://doi.org/10.1109/HNICEM51456.2020.9400113>
16. Li X, Luo X, Hao H, Chen F, Li M (2020) Pedestrian detection method based on multi-scale fusion inception-SSD model. In: 2020 IEEE 9th joint international information technology and artificial intelligence conference (ITAIC), pp 1549–1553. <https://doi.org/10.1109/ITAIC49862.2020.9338909>
17. Xu Y, Cao Y, Liu Y (2021) Research on pedestrian detection based on improved SSD algorithm. In: 2021 international conference on information science, parallel and distributed systems (ISPDS), pp 192–196. <https://doi.org/10.1109/ISPDS54097.2021.00044>
18. Zhang N, Fan J (2021) A lightweight object detection algorithm based on YOLOv3 for vehicle and pedestrian detection. In: 2021 IEEE Asia-Pacific conference on image processing, electronics and computers (IPEC), pp 742–745. <https://doi.org/10.1109/IPECS1340.2021.9421214>

19. Malbog MA (2019) MASK R-CNN for pedestrian crosswalk detection and instance segmentation. In: 2019 IEEE 6th international conference on engineering technologies and applied sciences (ICETAS), pp 1–5. <https://doi.org/10.1109/ICETAS48360.2019.9117217>
20. Chen L et al (2021) Deep neural network based vehicle and pedestrian detection for autonomous driving: a survey. *IEEE Trans Intell Transp Syst* 22(6):3234–3246. <https://doi.org/10.1109/TITS.2020.2993926>
21. Zhang S, Wang X (2013) Human detection and object tracking based on histograms of oriented gradients. In: 2013 ninth international conference on natural computation (ICNC), pp 1349–1353. <https://doi.org/10.1109/ICNC.2013.6818189>
22. Ojha A, Sahu SP, Dewangan DK (2022) VNet: vehicle detection network using computer vision and deep learning mechanism for intelligent vehicle system. In: Noor A, Sen A, Trivedi G (eds) *Proceedings of Emerging Trends and Technologies on Intelligent Systems*. ETTIS 2021. AISC, vol 1371. Springer, Singapore. https://doi.org/10.1007/978-981-16-3097-2_9
23. Dewangan DK, Sahu SP (2022) Towards the design of vision-based intelligent vehicle system: methodologies and challenges. *Evol Intel*
24. Dewangan DK, Sahu SP (2022) Optimized convolutional neural network for road detection with structured contour and spatial information for intelligent vehicle system. *Int J Pattern Recogn Artif Intell* 36(06):2252002
25. Dewangan, D.K., Sahu, S.P. (2021). Lane detection for intelligent vehicle system using image processing techniques. In: Verma GK, Soni B, Bourennane S, Ramos ACB (eds) *Data Science*. TCSN. Springer, Singapore. https://doi.org/10.1007/978-981-16-1681-5_21
26. Banjarey K, Sahu SP, Dewangan DK (2022) Human activity recognition using 1D convolutional neural network. In: Shakya S, Balas VE, Kamolphiwong S, Du KL (eds) *Sentimental Analysis and Deep Learning*. AISC, vol 1408. Springer, Singapore. https://doi.org/10.1007/978-981-16-5157-1_54
27. Dewangan DK, Sahu SP (2021) Predictive control strategy for driving of intelligent vehicle system against the parking slots. In: 2021 5th international conference on intelligent computing and control systems (ICICCS), pp 10–13. <https://doi.org/10.1109/ICICCS51141.2021.9432362>
28. Pardhi P, Yadav K, Shrivastav S, Sahu SP, Kumar Dewangan D (2021) Vehicle motion prediction for autonomous navigation system using 3 dimensional convolutional neural network. In: 2021 5th international conference on computing methodologies and communication (ICCMC), pp 1322–1329. <https://doi.org/10.1109/ICCMC51019.2021.9418449>
29. Sahu SP, Dewangan DK, Agrawal A, Sai Priyanka T (2021) Traffic light cycle control using deep reinforcement technique. In: 2021 international conference on artificial intelligence and smart systems (ICAIS), pp 697–702. <https://doi.org/10.1109/ICAIS50930.2021.9395880>
30. Ojha A, Sahu SP, Dewangan DK (2021) Vehicle detection through instance segmentation using Mask R-CNN for intelligent vehicle system. In: 2021 5th international conference on intelligent computing and control systems (ICICCS), pp 954–959. <https://doi.org/10.1109/ICICCS51141.2021.9432374>
31. Dewangan DK, Sahu SP, Sairam B et al (2021) VLDNet: vision-based lane region detection network for intelligent vehicle system using semantic segmentation. *Computing* 103:2867–2892
32. Bhambri P, Bagga S, Priya D, Singh H, Dhiman HK (2020) Suspicious human activity detection system. *J IoT Soc Mob Anal Cloud* 2(4):216–221
33. Singh A, Bansal A, Chauhan N, Sahu SP, Dewangan DK (2022) Image generation using GAN and its classification using SVM and CNN. In: Noor A, Sen A, Trivedi G (eds) *Proceedings of Emerging Trends and Technologies on Intelligent Systems*. ETTIS 2021. AISC, vol 1371. Springer, Singapore. https://doi.org/10.1007/978-981-16-3097-2_8

34. Bhattacharya N, Dewangan DK (2015) Notice of removal: fusion technique for finger knuckle print recognition. In: 2015 International conference on electrical, electronics, signals, communication and optimization (EESCO), pp 1–4. <https://doi.org/10.1109/EESCO.2015.7253990>
35. Bhattacharya N, Dewangan DK, Dewangan KK (2018) An efficacious matching of finger knuckle print images using gabor feature. In: Saini A, Nayak A, Vyas R. (eds) ICT Based Innovations. AISComputing, vol 653. Springer, Singapore. https://doi.org/10.1007/978-981-10-6602-3_15

Localization of Calcifications in Mammograms Using CNN with GAP Layer



Praneeth Vykuntam, Venkata Rohith Vykuntam, Pragun Srivastav, Sri Sai Bharat Uppalapati, and Poornima Mohan

Abstract Breast cancer is the most often diagnosed cancer in women around the world, accounting for one out of every four cancer cases. When breast cancer is identified early and treatment is available, the survival rate is extremely high. It is vital to enhance access to early detection to combat the rising breast cancer burden. With technological innovations in recent decades, the healthcare sector has undergone a huge transformation. The detection of abnormalities is aided with the image pre-processing techniques, namely Median filter and CLAHE (contrast limited adaptive histogram equalization). This paper describes an automated model for localization of calcifications in mammograms using CNN with Global Average Pooling layers. Our method will help physicians in determining the lesions of the breast. The model is not just restricted to x-ray mammograms but can be used also for other medical images.

Keywords Localization · Global average pooling layer (GAP layer) · Median filter · CLAHE (Contrast Limited Adaptive Histogram Equalization)

P. Vykuntam (✉) · V. R. Vykuntam · P. Srivastav · S. S. B. Uppalapati · P. Mohan
Department of Electronics and Communication Engineering, Amrita Vishwa Vidyapeetham,
Kollam, Amritapuri, India
e-mail: vykuntampraneeth@am.students.amrita.edu

V. R. Vykuntam
e-mail: venkatarohith@am.students.amrita.edu

P. Srivastav
e-mail: pragunsrivastava@am.students.amrita.edu

S. S. B. Uppalapati
e-mail: usrisaibharath@am.students.amrita.edu

P. Mohan
e-mail: poornimamohan@am.amrita.edu

1 Introduction

Breast cancer ranks fifth in mortality worldwide after skin and lung cancer. Although it is more commonly found in women, men are also susceptible to this disease. This affects around one in eight women, which is more than fifty percent compared to the cases occurring in men every year. In 2020, the mortality to incidence ratio (MIR) was 0.30 around the world [1].

In the early 1980s, an invention called the mammogram machine allowed doctors to detect breast cancer in its early stages, which significantly increased the survival rate by 15–20% [2]. This technology has been extremely useful in identifying breast cancer before it has progressed too far. A mammography test helps the physician see through the tissues of the breast, which is induced with a low dose x-ray, which has an accuracy rate of 80 to 90% in detecting cancer [3].

Mammograms are used to diagnose breast cancer in which symptoms may not be observed. The diagnosis can only be made after the screening of the breast and the confirmation from the radiologist. Classification of mammograms is sometimes difficult for the human eyes, due to noise effects in the image and low contrast. To counter it, many researchers have implemented techniques related to noise removal and contrast enhancement—usually called image pre-processing techniques. Image preprocessing aims to enhance the quality of the images, so that one can analyse images efficiently [4]. There are various noise removal and contrast enhancement techniques applied to images. Median filtering and CLAHE are applied on the mammogram for noise removal and contrast enhancement [5].

Image segmentation is a better way to identify abnormalities in mammograms, which will lead to a better diagnosis. It is the process of dividing the components of the image into parts or objects, and there are numerous ways for that. Many researchers have used different techniques to segment images. Localization is a process used to identify cancerous masses in mammograms [6].

We propose a new method for the localization of cancerous lesions in mammograms using convolutional neural networks with global average pooling layers. Our method works for any type of mammogram, which may vary in the density and grade of the tissues in the breast. Sections 2 and 3 describes previous related works in literature and description of our proposed model, respectively. The results are included in Sect. 4.

2 Related Work

With the increase in the mortality rate associated with breast cancer, breast cancer detection has brought the attention of many researchers to develop systems to diagnose it.

Authors of [7] have used a dataset which is a collection of images from the picture archiving and communication system (PACS) of Hallym University. They pre-processed the data by resizing the images into 1000×1200 pixels and contrast enhancement using CLAHE. After this process, they trained a convolution neural network for identifying abnormalities in the form of calcification and mass.

Abdollah Jafari Chashmi and Mehdi Chehelamirani have proposed a model for breast cancer detection using the Mini—MIAS dataset. For denoising, they have used median filtering [8]. Median filtering was used because it retains sharpness while denoising. Their precision in denoising the image was measured by RMSE and PSNR, which gave good results.

With the vast availability of medical images in all segments, it is impossible to obtain an image with no noise or tampered with, So to remove or improve the quality of the image it is necessary to use a filter to enhance the image. And this issue is addressed in the recent study” Performance of image pre-processing filters for noise removal in transformer oil images at different temperatures” Authors of [9] have explored various denoising spatial filters. They have used a sample image to analyze different filtering techniques which are the Mean filter, Median filter, and Wiener filter. The one filter that stands out is the Median filter for its edge-preserving technique [9]. They have concluded their work by proving the differences and evaluating the filters based on PSNR and MSE values obtained.

Nowadays the resolution and to identify the different regions of an image has become a major issue in the medical sector, with various masses and the evolution or deterioration of human being cells has made it impossible to understand and distinguish the masses of lumps from general cells. to counteract the authors this [10] has used the CLAHE technique for the recognition of faces. In their study, they collected a few images and converted them into grayscale images. Following that, the image was pre-processed which includes morphological operations, contrast enhancement, and then CLAHE. After these processes, they used a local binary pattern histogram approach for face recognition and finally were successful in recognizing two faces per frame. They were successful in recognizing faces that are present in bright and dark images. They extended their work to a system that is authenticated using biometric data [10].

Aiman Ai-Sabaawi, Hassan Muayad Ibrahim, and Zinah Mohsin in their research have discussed CNNs with global average pooling layers in object localization. The GAP layer is a good replacement for fully connected layers in CNN which are used for image classification and these layers deal with the issue of over-fitting which is a limitation of deep learning [11].

Authors of [12] have proposed a system that uses a three-tumor brain magnetic resonance image data set which consists of 3064 images. They have developed a neural network model with the help of GAP layers and ResNet-50 for the classification of Multi tumor brain images. They have explained the classifications of brain tumors and used a global average pooling layer to localize the region of interest as they solve the issue of vanishing gradient and the problem of overfitting, and showcased the accuracy of the GAP and the loss of GAP. With and without data augmentation, the model could achieve an accuracy's around 97 percent. The model works better than others.

In [13], the authors have done a survey on-various diagnosis methods for breast cancer and various related imaging modalities. In [14], the authors have made a classification attempt on mammogram images from the CBIS-DDSM database using a feature vector framed from a co-occurrence matrix and local binary pattern. They have compared the performance of two machine learning algorithms-KNN and SVM-in classification.

In the recent studies of dimensionality reduction, the challenging part of it is to reduce the dimension while preserving the image with its original content. This issue is addressed by the authors of "An Efficient Dimension Reduction based Fusion of CNN and SVM Model for Detection of Abnormal Incidents in Video Surveillance"[15]. In this study, the authors have proposed a method that takes the input images and processes them through the Background subtraction method which uses Gaussian mixture procedure to frame the pixel of every mounting and followed by the classification part where it utilizes a few static pooling layers which are convolutional layers and pooling layers and then the features are extracted using CNN (convolutional neural network) and the model is trained and tested using SVM classifier to produce accurate results to recognize and track every activity of any individual.

In [16], the authors have proposed a model which uses the YOLO algorithm for detecting litter thrown out by moving vehicles and then detecting the license plate of that particular vehicle.

3 Methodology

Figure 1 shows the workflow diagram of the proposed model. As mentioned before, we use median filtering and CLAHE techniques to improve image quality. Once the image has been improvised, it is segmented to showcase the abnormalities in the mammogram.

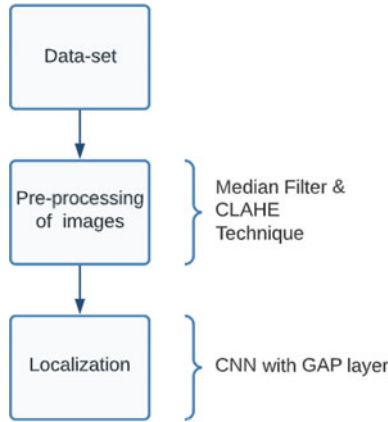


Fig. 1 Block diagram of proposed method

3.1 Dataset

The data set we used for the work was collected from the IN breast data set. It comprises 7,632 images. These images were collected from medical databases. The data set consists of 2,520 images of benign class and 5,112 images of malignant class. All the images were resized into images of size 227 * 227.

3.2 Pre-processing of Data

Image preprocessing helps the model in reducing the training time and also improves the quality of images. On images taken from the data set, the preprocessing techniques are applied, which are median filtering and contrast limited adaptive histogram equalization.

3.2.1 Median Filtering

It is a nonlinear method that is used to denoise the image. It is effective in removing the noise while maintaining the sharpness and the edges of the image. The basic working of this technique is to process the image from pixel to pixel while replacing each pixel value with the median of the neighbouring pixels. Every pixel in the noise affected image will be gone through and the intensity value is replaced by the median value of the pixel intensities in it's neighbourhood. An example of result of applying median filter for a noise affected pixel is shown in Fig. 2. The size of neighbourhood considered is 3 × 3 here.

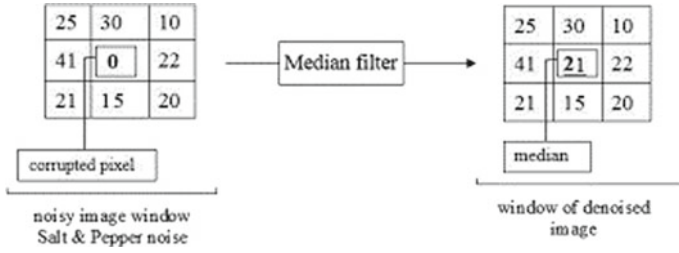


Fig. 2 An example of median filtering in 3×3 neighbourhood

3.2.2 Contrast Limited Adaptive Histogram Equalization

CLAHE is a method that enhances the local contrast of images. The basic working is that it divides the images into multiple sub-images and performs histogram equalization in those sub-images with contrast limiting. After getting filtered from the noise (salt and pepper noise) using a median filter, the image undergoes CLAHE, in which the contrast of an image is enhanced. The mathematical formula used for CLAHE is mentioned below.

$$\text{ClipLimit} = \left[\frac{\alpha}{256} \right] \pm \left[\beta(\delta) \left(\alpha \mp \left[\frac{\alpha}{256} \right] \right) \right]$$

It is limited to 8-bit gray images in which the pixel intensities should not exceed 255, as it ranges from 0–255. α , β and δ are the parameters defining block size, clipping parameters, and truncating value. It also depends on the number of bins, which should always be smaller than the number of pixels in a block. After getting filtered from the noise (salt and pepper noise), the image runs through the algorithm of the CLAHE technique, which enhances the contrast of some areas of the image.

3.3 Localization

In general, localization is the process of locating the presence of objects in an image by marking their location with a bounding box. These bounding boxes are defined by a point, Width, and height. We have used CNNs with global average pooling layers for localization.

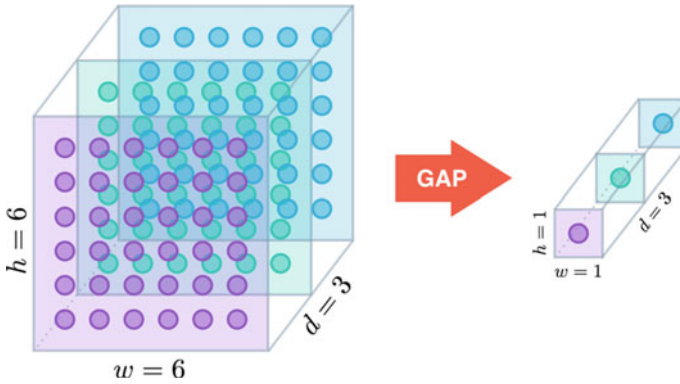


Fig. 3 Dimension reduction

3.3.1 Global Average Pooling Layer

It is used as an alternative to traditional fully connected layers in a convolution neural network. The global pooling layer reduces the entire feature map to a single value. It holds out among the remaining methods, due to its ability to generalize feature maps and categorize them. Since there are no parameters, overfitting is avoided. After passing through the algorithm, the resulting image will be ready to be segmented to showcase the masses with bounding boxes, that have parameters of width, height, and distance. To understand the working of this model, a basic example of dimension reduction is presented below in Fig. 3. After applying global average pooling, it can be observed that the dimension is reduced.

3.4 Model

After preprocessing the images from the dataset using the mentioned techniques, the images are stored in a directory. This directory is considered as input to the model, which predicts whether a particular mammogram consist of lesions that are harmful to the corresponding individual. We have split the dataset into training and testing with the train size of 0.7 and the test size of 0.3. Without scaling; the training may not converge. So after splitting the dataset, we have normalized the values resulting in all values falling between 0 and 1. In the model, we have used convolution and global average pooling layers. Also, the first few convolutional blocks are non-trainable, and only the last block is trained. This is only to expedite the training process. Lastly, we have fit the model with 14 epochs. The model has resulted in 95.32% accuracy. The localized images are saved in a new directory, and then the output images were produced from the last convolutional layer by resampling and resizing the image to the original image size.

4 Results and Analysis

Figure 4 represents the image before processing whereas Figs. 5, 6, and 7 are the results obtained after applying median filtering, CLAHE, and gap filter respectively. Localization for calcification was performed on preprocessed images. From Fig. 7, it is clear that our model is successful in the localization of calcifications. With 95.23% accuracy, we could do localization. The proposed model is implemented using google collab and to verify our proposed model, the performance and accuracy of this model are compared with various other models, such as ANN (artificial neural network), which resulted in 92.45% of accuracy. While processing the dataset with the YOLO model, the accuracy was 94.5%, which was noticeably good, but the issue with these models occurs when the dataset is large, resulting in over-fitting and reducing the accuracy. Thus, making the model vulnerable and avoiding it from detecting the lesions. Therefore, stating that the proposed model stands out among other models with high accuracy rates.



Fig. 4 Image before preprocessing

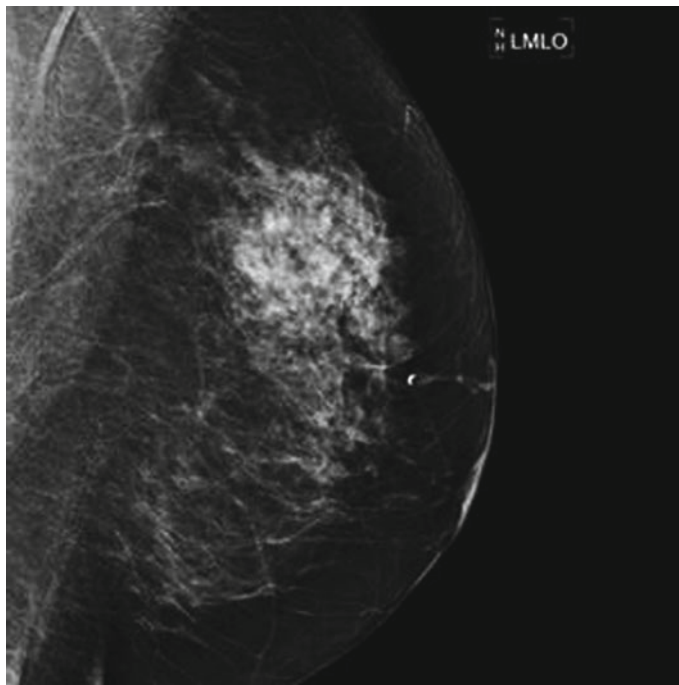


Fig. 5 Image after applying median filter

5 Conclusion

To detect breast cancer, radiologists need to diagnose cancer as soon as possible. For that, we need to ensure that effective techniques are used to detect the lesions. We aim to provide the reader with the working of the model and the working of each technique used. Using median filter and CLAHE, we have preprocessed the mammogram images in the dataset, and the preprocessed mammogram images were fed to the convolutional neural network with global average pooling layers for localization of calcifications, which could help a medical practitioner in better diagnosis. This model can be characterized for breast cancer detection and localizing the abnormalities in the mammography data.

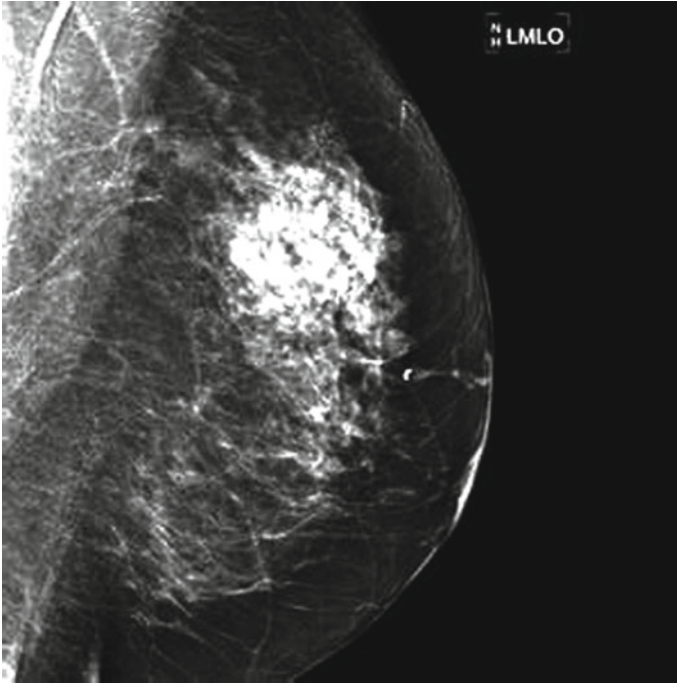


Fig. 6 Image after applying CLAHE

6 Future Work

The proposed study will be extended for categorising the images to its respective classes-benign and malignant.

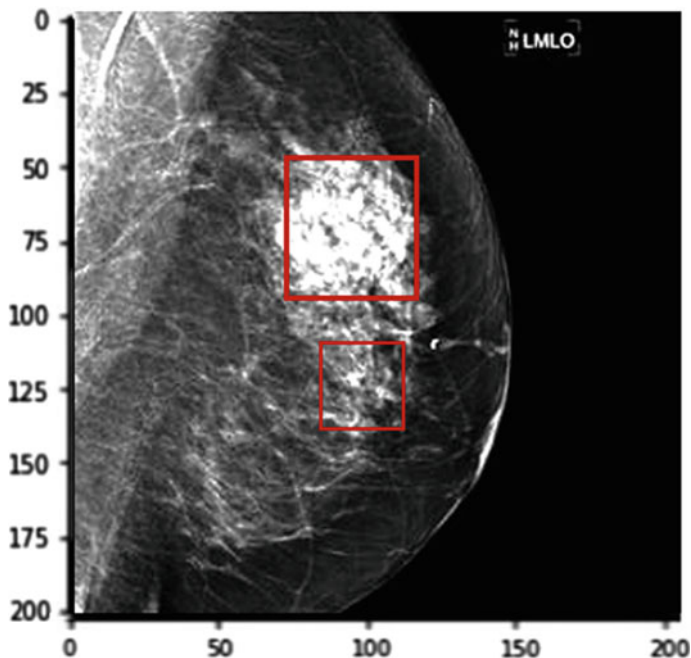


Fig. 7 Localized image using gap filter

Acknowledgements We express our gratitude to our beloved chancellor, Mata Amritanandamayi Devi, for her encouragement and inspiration in all of our endeavors. We also express our gratitude to our mentor, Mrs. Poornima Mohan, who helped us in the successful completion of the project.

References

1. Łukasiewicz S et al (2021) Breast cancer-epidemiology, risk factors, classification, prognostic markers, and current treatment strategies-an updated review. *Cancers* 13(17):4287. <https://doi.org/10.3390/cancers13174287>
2. Abdillahi B (2017) Image processing based detection of lung cancer on CT scan images. *IOP Conf Ser J Phy Conf Ser* 893:012063. <https://doi.org/10.1088/1742-6596/893/1/012063>
3. Sood R et al (2019) Ultrasound for breast cancer detection globally: a systematic review and meta-analysis. *J Glob Oncol* 5:1–17. <https://doi.org/10.1200/JGO.19.00127>. PMID: 31454282, PMCID: PMC6733207
4. Wason JV, Nagarajan A (2019) Image processing techniques for analyzing CT scan images towards the early detection of lung cancer. *Bio Inf* 15(8):596–599. <https://doi.org/10.6026/97320630015596>
5. Sarki R, Ahmed K, Wang H (2021) Image preprocessing in classification and identification of diabetic eye diseases. *Data Sci Eng* 455–471 (2021). <https://doi.org/10.1007/s41019-021-00167-z>
6. Rao SS, Shreyas R, Maske GL, Choudhury AR (2020) Survey of Iris image segmentation and localization. *J Innov Image Process (JIIP)* 02(03). <https://doi.org/10.36548/jiip.2020.3.005>

7. Suh YJ, Jung J, Cho BJ (2020) Automated breast cancer detection in digital mammograms of various densities via deep learning. *J Personalized Med* 10(4):211. <https://doi.org/10.3390/jpm10040211>. PMID: 33172076, PMCID: PMC7711783
8. Chashmi AJ, Chehelamirani M (2019) Merit Res J Eng Pure Appl Sci 5(1):014–018 (2019). <https://doi.org/10.5281/zenodo.3374916>
9. Mahesh CM, Prasanna Kumar H (2020) Performance of image pre-processing filters for noise removal in transformer oil images at different temperatures. *Appl Sci App* 2:67. <https://doi.org/10.1007/s42452-019-1800-x>
10. Musa P, Al Rafi F, Lamsani M (2018) A review: contrast-limited adaptive histogram equalization (CLAHE) methods to help the application of face recognition, October 2018. <https://doi.org/10.1109/IAC.2018.8780492>
11. Al-Sabaawi A, Ibrahim HM, Arkah ZM, Al-Amidie M, Alzubaidi L (2021). Amended convolutional neural network with global average pooling for image classification. In: Abraham A, Piuri V, Gandhi N, Siarry P, Kaklauskas A, Madureira A (eds) *Intelligent Systems Design and Applications. ISDA 2020. AISC*, vol 1351. Springer, Cham. https://doi.org/10.1007/978-3-030-71187-0_16
12. Kumar RL, Kakarla J, Isunuri BV et al (2021) Multi-class brain tumor classification using residual network and global average pooling. *Multimed Tools Appl* 80:13429–13438. <https://doi.org/10.1007/s11042-020-10335-4>
13. John KA, Mohan P (2020) A comparison between KNN and SVM for breast cancer diagnosis using GLCM shape and LBP features. In: 2020 third international conference on smart systems and inventive technology (ICSSIT), pp 1058–1062. <https://doi.org/10.1109/ICSSIT48917.2020.9214235>
14. Swathi TV, Krishna S, Ramesh MV (2019) A survey on breast cancer diagnosis methods and modalities. In: 2019 international conference on wireless communications signal processing and networking (WiSPNET), pp 287–292. <https://doi.org/10.1109/WiSPNET45539.2019.9032799>
15. Sharma R, Sungeetha A (2021) An efficient dimension reduction based fusion of CNN and SVM model for detection of abnormal incident in video surveillance. *J Soft Comput Paradigm (JSCP)* 3(02):55–69
16. Amrutha JM, Nandini S, Anjali T, (2022) Real-time litter detection system for moving vehicles using YOLO. In: 2022 4th international conference on smart systems and inventive technology (ICSSIT), pp 1311–1315. <https://doi.org/10.1109/ICSSIT53264.2022.9716512>.

Comparative Analysis of Machine Learning and Deep Learning Algorithms for Real-Time Posture Detection to Prevent Sciatica, Kyphosis, Lordosis



Palavalasa Venkata Satish and Meena Belwal

Abstract People must incorporate a “work from home” strategy because of the COVID-19 outbreak. In today’s pandemic situation, due to working from home, employees are working for long hours, and spending long hours is a pretty challenging task. Nowadays, irrespective of age concerns, sciatica, Kyphosis, and lordosis are becoming a significant problem even for youngsters. The longest nerve in our body is sciatica, which causes severe pain due to stress applied while sitting in the wrong posture. It gets compressed with our lower back discs, which may lead to severe radiating pain from our lower back disc to the entire right leg, and a person can’t even perform his daily activities comfortably. To prevent these problems, sitting posture while working should be maintained correctly. This work mainly focuses on preventing employers, and students from sciatica, Kyphosis, and Lordosis health issues. We used all kinds of sitting postures that interact while working with a laptop, classified which posture was good, and predicted which stance led to health issues. We used Convolutional Neural Network and K-Nearest Neighbor machine learning algorithms to predict the correct sitting postures. In KNN, we followed two techniques to improve the performance: using Edge detection. The other method we used was detecting facial landmark detection and plotting their respective rotational angles. So by using this technique, we improved the accuracy and precision rate compared to the classical Edge detection. We also trained the model with CNN, which gives good results. We performed a comparative analysis to pick the best model to integrate with OpenCV to make it real-time.

Keywords Sciatica · Kyphosis · Lordosis · Convolution Neural Network · K-Nearest Neighbor

P. V. Satish (✉) · M. Belwal
Department of Computer Science & Engineering, Amrita School of Engineering, Bengaluru,
Amrita Vishwa Vidyapeetham, Bengaluru, India
e-mail: venkatasatish615@gmail.com

M. Belwal
e-mail: b_meena@blr.amrita.edu

1 Introduction

In recent days due to the Covid Outbreak employers are working from home and not maintaining their postures.

Working from home can cause a significant shift in your ergonomic setup, making it challenging to maintain a healthy posture. The main motto of this paper is by applying AI power to sitting postures, one can overcome bad sitting postures and maintain a good stance for extended periods. Poor postures can cause problems to a person's physical and mental health, resulting in near-sightedness, abdominal cervical infection, Kyphosis, Lordosis, Sciatica, and other ailments. In this project, we mainly focus on three major issues Sciatica, Kyphosis, and Lordosis which are caused by sitting postures. In this work [1] the author mainly focuses on how sitting postures in our daily life will cause issues like sciatica and explains how our spine is affected by continuous long periods of sitting and sitting with bending forward. Good posture aids in preventing this serious threat to people's health. Late tests have also revealed a significant relationship between sitting tendencies and efficiency; for example, a consistent sitting position addresses constant focus. In this work [2] the author mainly shows how the four major regions like cervical, thoracic kyphosis, lumbar lordosis, and pelvic forward regions are getting affected while sitting in the wrong postures. The work [3] shows that the solution for Kyphosis and Lordosis is due to bad sitting postures and this can be avoided by changing the movements of the postures. Our work mainly focuses on image processing using deep learning techniques and machine learning techniques to distinguish whether the sitting stance is Good or Bad, causing severe medical problems.

In any event, We can see that even for younger generations, these health issues are common. They are suffering from a lot of agonies, which is a bit uncomfortable, and they are losing their confidence in this aggressive world.

Therefore, prevention by adhering to good postures is the ideal way to overcome this, taking some time off from routine to make our bodies easier to adapt and help us relax more. It is difficult for any human to work for a long time without good assets, so they are attempting to change their position as expressed by their comfort. However, they may have chosen the wrong positions, and they operate continuously in the postures which are very bad [4]. They will lose efficiency while working. So they cannot play their regular businesses at their level of comfortable posture. The significant issues are caused due to spine issues since it prompts total anxiety on it.

The term sciatica depicts torment along the way of the sciatic nerve from the lower back, through the hips and backside, and down the rear of the legs to the feet. Sciatica ordinarily influences just a single leg yet can be felt on the two sides. Sciatica can change from gentle distress to severe torment and incorporate deadness, shivering, consumption, and shortcoming. Specific individuals experience expanded agony after standing, while others experience expanded suffering when sitting.

Uneven postural characteristics cause many diseases such as spinal stenosis, degenerative plate illness, spondylolisthesis, bone spikes, piriformis disease, muscle fits, and herniated and burst lumbar circles.

For the most part, these are actual reactions to grating, stress, and strain because the body has lost ideal postural arrangement, and the compromised act is making pressure and shear that harms the body's tissues. Kyphosis and Lordosis are different back and forth movements found in the body of the spine. Excessive kyphosis in the thoracic spine—similarly returned to change or hunchback.

0-h Superfluous lordosis, also known as swayback, is categorized by an absurd turnaround bowing of the spine where the midriff projects out. The two conditions can cause torture and comfort and have an impact on a person's certainty. These are related to spinal pain issues so maintaining sitting positions while sitting in front of a Laptop mainly focuses on the upper portion of the body and by changing the movements of the head, and neck there is a change in the spine position as well.

The remaining portion of the paper is organized as follows. Section 2 goes over the Related work. Section 3 of this paper discusses the Methodology. Section 4 describes the experiments. Section 5 discusses the results. Section 6 discusses the conclusions.

2 Related Work

Various types of research on sitting postures have previously been conducted for multiple purposes. A comparison of the impact of sitting postures on the back and referred pain was published by the Auckland Institute of Technology in Auckland, New Zealand, that focused on the causes of the kyphotic or lordotic posture on the low-back and referred pain [5]. It is frequently seen in many working professionals who don't maintain the correct body postures.

Several methods for measuring and categorizing sitting posture have been developed. One of the approaches for recognizing and capturing sitting positions is to utilize a camera.

Sensors attached to the user's back can also be used to detect sitting postures based on the body's movement are discussed in this paper [6–8]. There is a disadvantage of being capable of causing unusual movements due to the attached sensors is mainly discussed in this paper [6]. To address these issues, Pressure sensors are installed on chair components that are used to categorize sitting postures. To address the issues, previous research was conducted to classify user sitting postures by inserting pressure sensors into chair structures at the seat and backrest. Because this sensor-based detecting sitting posture has a disadvantage while capturing posture movement, the alternative is to opt for camera capture. Most studies categorize sitting upright stances, and those with the waist straight, while capturing photographs with the camera. Poses in which the feet are flat on the ground, postures in which the upper section is tilted forward, backward, left, or right, and poses in which the left or right leg is crossed. Therefore, in the work [8] the authors mainly focus on the postures like in front of the system or table of the upper part of the body.

Another work [9] shows how to create the images based on the specific positions when a the person sits in front of the camera. Article [9] shows how the feature extraction is done by one of the most prominent edge detections called Sobel Edge

detection. The same work [9] has another feature extraction approach by using Topological Feature Extraction. The Topological Feature extraction is done by the Foreground extraction using OpenCV to locate the body joints of the posture. By using these two approaches, one can able to load labels to the Picture. Once the labeling part is completed, it uses KNN architecture to classify the Postures based on the labels.

The work in [10] describes how the classification of images is done by various deep learning algorithms and their advantages over other algorithms. One among them is classification using a Convolution Neural Network. It also states how many layers a convolution layer should take in the architecture design and how it works for the category of images.

The work in [11] shows how sitting postures in aircraft provide comfort for the passengers. For this posture detection, the authors used Machine learning algorithms to classify the postures by Support vector machine. This classification helps the passengers who travel for long hours so that designers of aircraft seats can provide good comfort for the passengers.

Another work shows how AI power is used to build providing good postures [12]. It uses sensing technology and machine learning to create the Life chair Smart Cushion and provides the vibrator when it detects wrong postures.

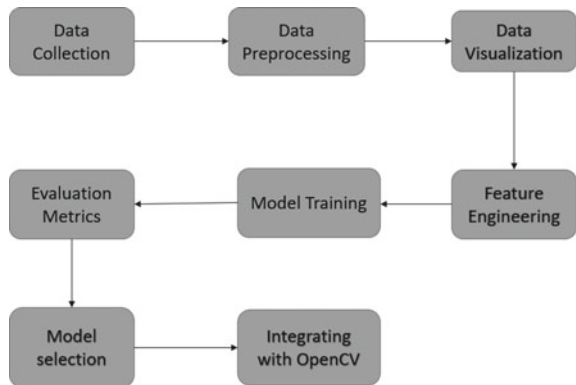
Other work mainly shows how habituating a good posture in children does not cause Lumbar disease [13].

The authors used the Convolution Neural Network Algorithm to detect the sitting postures for the children to avoid the habitual postures.

3 Proposed Work

Figure 1 depicts the proposed work for the paper. There are five steps to be followed explained in this section.

Fig. 1 Design model



3.1 Data Collection

The information we seek is a person's sitting body image which shows the postures from the front and includes the nine different types of images mentioned in this paper [9]. We considered upper portion images while sitting on the chair usually how a person will interact when working with a laptop and these pictures will help to classify the postures for real-time purposes, Participants are encouraged to participate in the collection by demonstrating the nine postures: upright, downright, right partial, left partial, head down, left-hand cheeks, right-hand cheeks, body left oblique, body right oblique, and head up. These types of postures will help to understand that facial movement, neck, and head movements completely depend on our spine which causes the issues. This Dataset is collected for training purposes and is of the size is 2030 images, both Good and Bad Postures.

3.2 Data Preprocessing

Data preprocessing is a technique for increasing the amount of a training dataset by generating many versions of the same image. For training, the models using Deep Neural Network data augmentation help to prepare the best fits for the models. Data Augmentation is a technique that allows experts to expand the amount of data available for model preparation. We use simple augmentation techniques like blurring, contrast, scaling, and illumination used to reduce the overfitting data for modeling. This method is utilized to improve the outcomes.

3.3 Data Visualization

Data visualization is used to know how the image properties are varied after applying image processing techniques to the data set. This Data visualization helps to know if there are any problems in the data set like a balanced dataset or an Unbalanced dataset, the size of the data set, and the shape of the images, how edges are detected.

3.4 Model Training

The K-Nearest Neighbor and Convolution Neural Network Algorithms construct the model. We use Feature Extraction Techniques for K-Nearest Neighbor before passing the images to the model so that the model can classify the images. We may not use any techniques for Convolution Neural networks because the architecture finds ways to classify itself.

3.5 *Evaluating the Model*

For evaluating the models, we use metrics like *Accuracy*, *Precision*, *F1 score*, and *Recall* are calculated to evaluate the models.

3.6 *Model Selection*

In this experiment, we built three models, two models using KNN with two different approaches and the other using CNN Model, and we calculated the evaluation metrics. Based on these evaluation metrics, we picked up the best model.

3.7 *Integrating with OpenCV*

Pickling is done to serialize and de-serialize python objects. Models that are built for training are converted to byte streams. Un-pickling is done to convert byte streams to python objects. These pickle files are used in real-time posture detection while reading the images from the webcam.

The inference is to grab the input from the webcam and classify the posture as good or bad. The pose will be recorded using the webcam, and the predicted posture will be displayed in the window. To complete this task, we must build Machine Learning and Deep Learning algorithms such as CNN and KNN. Before fitting the model, the Machine Learning Algorithm requires feature engineering techniques, also known as image processing techniques should be used before training the model. In Convolution neural network, convolution layers in the architecture will handle it automatically.

These algorithms are divided into two phases: training and testing. We trained the model using the 80% dataset and tested it using the remaining 20%.

Training the Model using Machine Learning

We used the data set that was created as explained in Sect. 3 under data collection, to train the model. For the Machine Learning model, we trained using Image processing to classify the postures.

Steps to follow for image processing: Read the images from the dataset labeled as Good Posture and Bad Posture. After reading the photos in the dataset, resize all the images to the required size.

We followed two techniques for feature extraction.

- **Local Edge Feature Extraction**
- **Topological Feature Extraction**

Local Edge Feature Extraction

In Local Edge Feature Extraction, we converted the image to a grayscale image to make the picture suitable for edge detection and good pre-processing time. After converting to the grayscale image, applied the Gaussian blur to remove the noise or outliers in the photos to get better results while training. Next, we used the Sobel Edge Detection to detect the edges of the images. The Sobel Operator employs kernels and the convolution operation to see edges in a snap. The algorithm uses two kernels.

- A kernel to find the intensity changes in the horizontal direction.
- A kernel to find the intensity changes in the vertical direction.

Sobel Edge detection is worked by convolving the grayscale image pixels to the kernel in the X-direction and Y-direction. The gradient approximations in the X-direction and in Y-direction is denoted as G_x and G_y respectively. The formulas to compute G_x and G_y are:

$G_x = \text{horizontal direction kernel} * (3 \times 3 \text{ portion of image } A \text{ with } (x, y) \text{ as the center cell}).$

$G_y = \text{vertical direction kernel} * (3 \times 3 \text{ portion of image } A \text{ with } (x, y) \text{ as the center cell}).$

Wherein (x, y) represents the pixel around which we want to apply the gradient approximation. Finally, the magnitude (G) is calculated using G_x and G_y as:

$$\text{magnitude}(G) = \sqrt{G_x^2 + G_y^2}$$

The direction of the gradient θ at pixel (x, y) is calculated as:

$$\theta = \text{atan} \frac{G_y}{G_x}$$

In this way, the edge is detected in the images and given the data to train the model for classification.

Topological Feature Extraction

In Topological Feature Extraction, first, we should perform the foreground extraction. This foreground extraction is achieved by using the python OpenCV library. Foreground extraction is used to remove the background of the images so that the image training is not dependent on the background image.

After performing the foreground extraction, we used pre-trained models like Dlib 68 points Face landmark detection to detect the facial parts and calculated the rotation angles of the facial features like eyes, mouth, and Nose angles and plotted the facial landmarks for every image this technique will help to increase the accuracy and used these data to train the model for classifying the images.

Considering the above two techniques, we created two models using KNN to get better results. We used both the predicted values and if both models predict the same output, it will classify the posture.

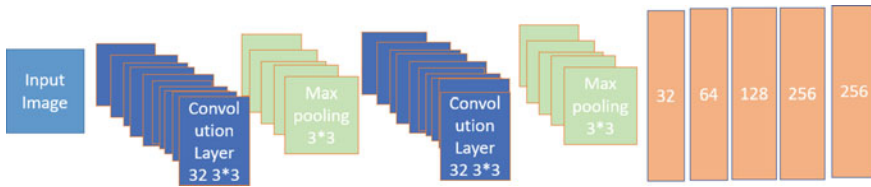


Fig. 2 Architecture for CNN

Training the Model using Deep Learning

To train the model using deep learning method, we use Convolution Neural Network Algorithm. For Convolution Neural Networks, there is no need for Feature Extraction techniques. This is taken care of by the convolution layers, which get convolved by the kernels present inside the layers. These convolution layers perform edge detection. Figure 2 shows how the CNN architecture is built.

To build the convolution layer, we took the input as an image size of 64×64 pixels and added a first convolution layer with 3×3 kernels of 32 with a max-pooling layer of 3×3 . Similarly, we build the second convolution layer. After that, we used fully connected layers with 32, 64, 128, and 256. On the output side, SoftMax acts as an activation function. For this, we used Adam Optimizer to build the Convolution Neural Network. This model is used for training the dataset. For Training, we used 80% of the data from the dataset.

After performing the Training, we saved both the models into a pickle file so that there is no need to train the model repeatedly. The remaining 20% of the dataset is used for testing purposes.

This testing method is used to estimate whether a model's performance is adequate for deployment. Testing the model should follow similar pre-processing steps as Training for the CNN and KNN models. After performing the pre-processing steps, it will infer the results. Compare the results with the targeted outputs and draw the evaluation metrics like Accuracy, Precision, Recall, and F1Score for both Models.

We created a user interface with Tinker library and OpenCV library, which detects whether the posture is good or bad and shows the labeled image. GUI works by Converting the video into frames. From that frame, it undergoes all the pre-processing steps mentioned for both deep learning and machine learning models and shows the results in the window.

4 Results Analysis

Python 3 is used for detection, extraction, and classification. Figs. 3, 4, and 5 depict the obtained testing confusion matrices for all three models.

Below mentioned are key terms related to confusion matrices:

True-Positive (TP): A model that predicted Good Posture accurately as Good Posture.

False-Positive (FP): A model that incorrectly predicted Bad Posture as Good Posture.

False Negative (FN): A model that incorrectly predicted Good Posture as Bad Posture.

True Negative (TN): A model that correctly predicted Bad Posture as Good Posture.

For the proposed system’s evaluation, the following performance matrices are computed:

Fig. 3 Confusion matrices for CNN

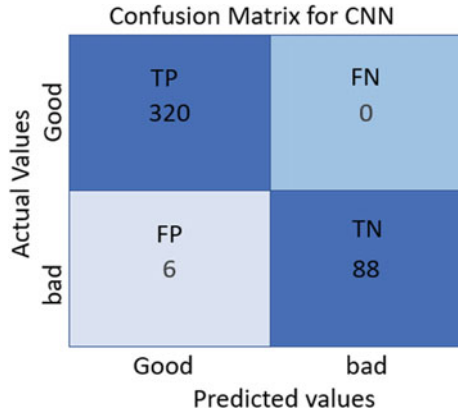


Fig. 4 Confusion matrices using KNN (Sobel edge)

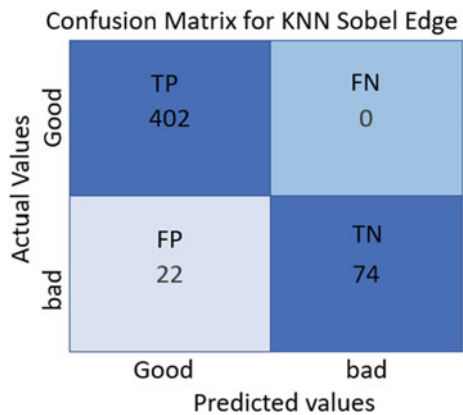


Fig. 5 Confusion matrices using KNN (Topological feature extraction)

Confusion Matrix for Topological features extraction

Actual Values	Good	TP 404	FN 0
	bad	FP 6	TN 98
		Good	bad
		Predicted values	

Accuracy: This is the proportion of exactly labeled postures to the total count of poses.

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

Precision: Precision is defined as the percentage of exactly labeled good postures compared to all good ones. A good classifier should have a precision of a high value.

$$Precision = \frac{TP}{TP + FP}$$

Sensitivity or Recall: The recall is the ratio of exactly labeled good postures divided by the total count of wrong postures.

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

F1 Score: The harmonic mean of recall and precision is used to calculate the F1 score. It only becomes 1 when both precision and recall are 1.

$$F1Score = \frac{2 * precision * Recall}{Precision + Recall}$$

Specificity: It shows the percentage of exactly labeled bad postures to the total count of Good Postures.

$$Specificity = \frac{TN}{TN + FP}$$

Table 1 Performance results

	Accuracy	Precision	Recall	F1 Score
CNN	0.98	0.981	0.99	0.99
KNN edge detection	0.94	0.991	0.737	0.848
KNN Face landmarks detection	0.988	0.99	0.942	0.947

Confusion Matrix

We had built the confusion matrix for all the three models shown in Figs. 3, 4, and 5 from those confusion matrices we extracted the evaluation metrics for the three classification models that we had shown in Table 1. We can compare which algorithm is the best among others through this table. The accuracy of KNN edge detection is 94% and recall of this classification is 0.737 these two metrics are improved by following KNN face landmarks detection here we spotted the facial landmarks and calculated the translation motion and rotational motion. Translation motion means it represents uniform motion, Rotational motion represents circular motion. The angle for this motion is calculated by Rodrigues's formula. From these angles, we extracted landmark movements and provided resultant data to canny edge detection before training the model using KNN so that we had improved both accuracy and Recall is increased. We also trained the model by using CNN architecture we get an accuracy of about 0.98 and all other parameters are also good by using the Deep learning technique. We plotted the ROC curves for all the three models shown in Fig. 6, it gives the relation between sensitivity and specificity. AUC value is more in facial landmark detection approach compared to edge detection technique. CNN has the good AUC curve. We Integrated these models as a Real time approach by integrating with the webcam for this we used OpenCV and Tinkter as a GUI. There is a Interface created which allows to open the webcam from this video frames are continuously readied and get the Inference by following the similar pre-processing steps as training. Output of the Real-Time is shown in Fig. 7.

Figure 7 shows how the results are displayed with labeling in real-time using OpenCV.

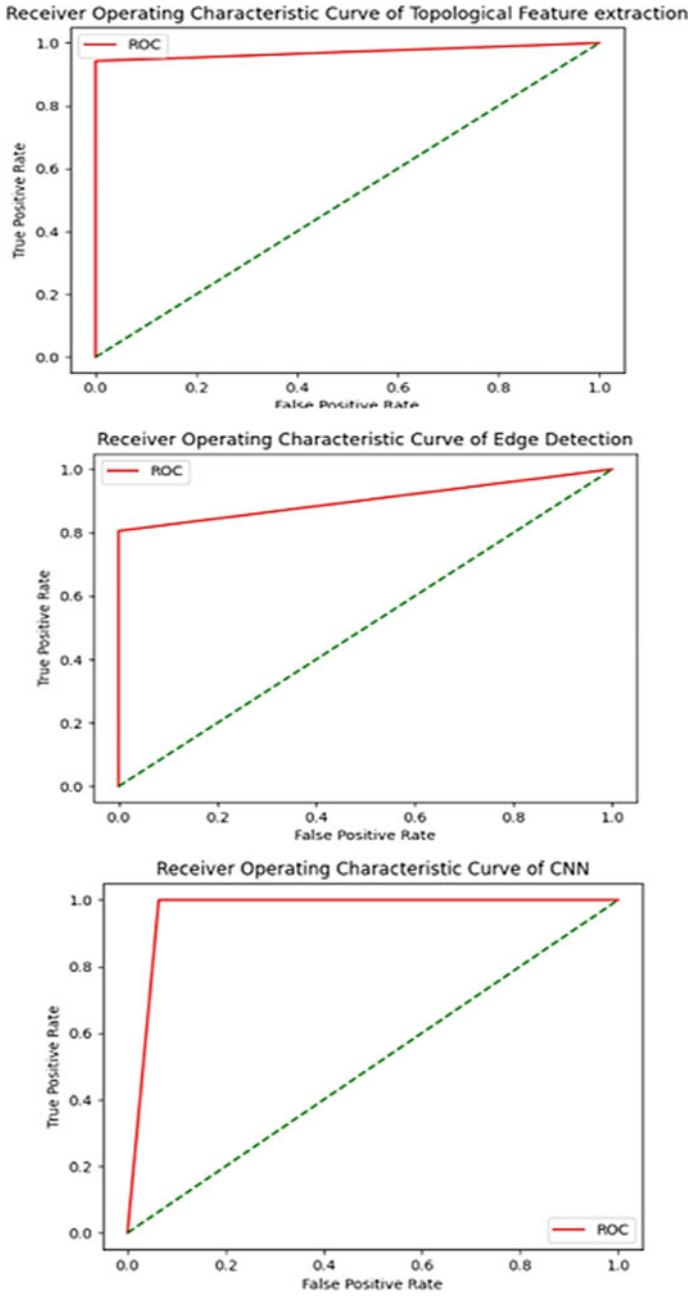


Fig. 6 ROC curves for three models

Fig. 7 Real-time results after classification



5 Conclusions

To train the model using machine learning requires feature engineering techniques, whereas a deep learning process does not need feature engineering techniques but does require pre-processing data. We took two different approaches for KNN, a machine learning technique — Sobel Edge Detection and Topological Feature Extraction (Face landmarks detection). While using Sobel Edge detection, we achieved an accuracy of 93%. With the facial landmark approach, we increased accuracy to 98%. From Fig. 6, we can state the ROC curve is Good. We summed up the classification results and predicted the outputs to get better results. We used the Deep Learning technique Convolution Neural Network to achieve 98% accuracy. Comparatively, For CNN, other metrics like *Precision*, *Recall*, and *F1 Score* are higher than KNN. Using different CNN algorithms like VGG16, VGG19, and ImageNet, the models were undergoing overfitting.

The Future Scope for this work is it should be placed in the working environment so that it will be helpful for employers to overcome health issues like sciatica, neck pain, and back pain and help maintain good postures and maintain long sustainability in their respective fields.

References

1. Liebenson C (2012) What can I do for sciatica? *J Bodyw Mov Ther* 16(3):369–371
2. Zhai M et al (2022) Effects of a postural cueing for head and neck posture on lumbar lordosis angles in healthy young and older adults: a preliminary study. *J Orthop Surg Res* 17(1):1–12
3. Pynt J, Higgs J, Mackey M (2001) Seeking the optimal posture of the seated lumbar spine. *Physiother Theory Pract* 17(1):5–21
4. Williams MM, Hawley JA, Mckenzie RA, Van Wijmen PM (1991) A comparison of the effects of two sitting postures on back and referred pain. *Spine* 16(10):1185–1191
5. Grunseit Anne C (2017) Patterns of sitting and mortality in the Nord-Trøndelag health study (HUNT). *Int J Behav Nutr Phys Act* 14:1–7
6. Dunne Lucy E (2008) Wearable monitoring of seated spinal posture. *IEEE Trans Biomed Circ Syst* 2:97–105
7. Knight JF (2007) Uses of accelerometer data collected from a wearable system. *Pers Ubiquitous Comput* 11:117–132
8. Foerster F, Smeja M, Fahrenberg J (1999) Detection of posture and motion by accelerometry: a validation study in ambulatory monitoring. *Comput Hum Behav* 15:571–583
9. Bei S (2017) Sitting posture detection using adaptively fused 3D features. In: *IEEE 2nd Information Technology Networking Electronic Automation Control Conference (ITNEC)*
10. Manjula PM, Adarsh S, Ramachandran KI (2020) Driver inattention monitoring system based on the orientation of the face using convolutional neural network. In: *2020 11th international conference on computing, communication and networking technologies (ICC- CNT)*
11. Cun W (2021) Sitting posture detection and recognition of aircraft passengers using machine learning. *AI EDAM* 35:284–294
12. Bourahmoune K, Amagasa T (2019) AI-powered posture training: application of machine learning in sitting posture recognition using the LifeChair smart cushion. In: *IJCAI*
13. Kim YM et al (2018) Classification of children’s sitting postures using machine learning algorithms. *Appl Sci* 8(8):1280
14. Vijayakumar T (2020) Posed inverse problem rectification using novel deep convolutional neural network. *J Innov Image Proc (JIIP)* 2(03):121–127
15. Chen Joy Iong-Zong, Chang Jen-Ting (2020) Applying a 6-axis mechanical arm combine with computer vision to the research of object recognition in plane Inspection. *J Artif Intell* 2(02):77–99
16. Somepalli, Meghana Rao, et al (2021) Implementation of single camera markerless facial motion capture using blendshapes. In: *2021 IEEE International Conference on Computation System and Information Technology for Sustainable Solutions (CSITSS)*. IEEE
17. Madireddy R et al (2021) Driver drowsiness detection system using conventional machine learning. In: Smys S, Palanisamy R, Rocha Á, Beligiannis GN (eds) *Computer networks and inventive communication technologies*, vol 58. *Lecture Notes on Data Engineering and Communications Technologies*. Springer, Singapore, pp 407–415. https://doi.org/10.1007/978-981-15-9647-6_31
18. Babu A, Nair S, Sreekumar K (2022) Driver’s drowsiness detection system using Dlib HOG. In: Karuppusamy P, Perikos I, García Márquez FP (eds) *Ubiquitous intelligent systems*, vol 243. *Smart Innovation, Systems and Technologies*. Springer, Singapore, pp 219–229. https://doi.org/10.1007/978-981-16-3675-2_16
19. Akama S, Matsufuji A (2018) Successive human tracking and posture estimation with multiple omnidirectional cameras. In: *2018 Conference on Technologies and Applications of Artificial Intelligence (TAAI)*
20. Xing Y et al (2017) Identification and analysis of driver postures for in-vehicle driving activities and secondary tasks recognition. *IEEE Trans Comput Soc Syst* 5(1):95–108
21. Esmaeili B, Alireza A, Alireza B (2020) An ensemble model for human posture recognition. In: *2020 International conference on machine vision and image processing (MVIP)*. IEEE

22. Neha R, Nithin S (2018) Comparative analysis of image processing algorithms for face recognition. In: 2018 International conference on inventive research in computing applications (ICIRCA). IEEE
23. Gupta R, Devesh S, Shubham M (2020) Posture detection using deep learning for time series data. In: 2020 Third international conference on smart systems and inventive technology (ICSSIT). IEEE
24. Suja P (2019) A robust pose & illumination invariant emotion recognition from facial images using deep learning for human-machine interface. In: 2019 4th International conference on computational systems and information technology for sustainable solution (CSITSS), Vol. 4. IEEE
25. Vamsi M, Soman KP (2020) In-vehicle occupancy detection and classification using machine learning. In: 2020 11th International conference on computing, communication and networking technologies

Lightweight Block Cipher for Resource Constrained IoT Environment—An Survey, Performance, Cryptanalysis and Research Challenges



M. Abinaya and S. Prabakeran

Abstract Nowadays IoT (Internet of Things) is becoming a more popular environment and has a variety of applications like Smart Home, Smart Healthcare, Vehicles, and many Industries. There is plenty of information shared among the devices with the use of the internet. Due to this importance of data sharing, there is a possibility of security attacks, and threats. IoT environments have many security challenges including Providing Confidentiality, Integrity, and Availability in addition Privacy, Authentication. These challenges can be fulfilled by many cryptographic algorithms. Since IoT has limited memory, resources, power, those cryptographic primitives may not be suitable. The best solution for this problem is lightweight cryptographic algorithms. This paper presents the importance of lightweight cryptography algorithms. We analysed the performance of current algorithms in terms of throughput, latency, ROM/RAM, software efficiency, and energy. We are comparing the cryptanalysis of some popular algorithms. Also, we are discussing the research challenges and research gaps in the area of lightweight cryptography for providing better performance, cost and software implementation without affecting high security.

Keywords Internet of Things · Security issues · Lightweight block Cipher · Performance metrics. Cryptanalysis

1 Introduction

This is a fast-moving field known as the Internet of Things. By 2024, there are expected to be 50 billion gadgets on the market, and it is imperative that we understand what it will take to get there.

M. Abinaya (✉) · S. Prabakeran
Department of Networking and Communication, SRMIST, Chennai, India
e-mail: am8580@srmist.edu.in

S. Prabakeran
e-mail: prabakes@srmist.edu.in

The Internet of Things (IoT) is an environment that encompasses connecting devices to the internet and using that connection to facilitate remote monitoring or control of those objects [15]. No substantial security concerns were raised when the Internet of Things (IoT) technologies were first built by linking small devices equipped with sensors, as was the case when they were first developed.

The Internet of Things (IoT) is becoming increasingly significant in terms of security as more and more devices are connected to exchange private and sensitive information. Each stage of the Internet of Things design lifecycle has a distinct set of security and research issues.

IoT devices are classified into two types. One has many resources, while the other does not. Servers, personal computers, tablets, and smartphones are examples of high-resource devices. On the other hand, insufficient resources are sensor nodes, RFID tags, actuators. These resource-constrained devices have some security flaws. Every smart device must be protected and maintained as new vulnerabilities are discovered.

Confidentiality, Integrity, Availability, Privacy, Authenticity, and Lightweight Solutions are all well-known security objectives. During any transfer, confidentiality is a way of securing all information from unauthorised nodes [52]. It may be performed by the sender and recipient exchanging safe keys. In addition, encrypt the data before delivering it to the recipient and then decode the data after receiving it using this key to obtain the original information. It is imperative that data stored in the cloud remain private.

The integrity of the transmission guarantees that it does not alter throughout transmission. When it comes to data transport, a symmetric cryptographic approach is frequently employed to generate signatures. Another function is the Message Integrity Check, which checks to see if the data received is correct before displaying it. The system must be capable of displaying the route if a change is detected and an activity log must be created in order to demonstrate the change. They may be held locally or centrally, for a short period of time or for an extended period of time, and for any reason at all.

Availability ensured that authorized users could always access IoT services and applications. When the connected things are needed, they should be available and functional. It is essential that the system has the ability to protect itself and heal itself in the case of a failure or an attack. Hierarchical organisation of Internet of Things nodes can help to improve scalability.

Privacy is the ability of an individual or a group to separate information about themselves or themselves and thereby selectively express themselves. To keep the nodes flexible and consider a wide range of IoT applications, RFID tags provide robust privacy. The personal information of other users should not be used to create profiles by unidentified individuals. An IoT device's current or previous location can't be revealed.

Lightweight solutions is a novel characteristic for IoT devices since IoT devices are generally computationally lightweight and have limited memory.

For Internet of Things devices, authenticity is essential since it enables them to verify and authenticate the active users of the connection. User authentication, context

authentication, and device authentication all need to be confirmed for a user to be authenticated in the context of a given system. Secondly, there's Trust Management, which focuses on IoT security and network performance in general. It further states that IoT devices and the central unit must validate a user's identity on their own, without the assistance of a third party.

Hardware is more important to execute the algorithm in faster. In IoT environment all the hardware devices are very small having low energy and its working based upon the battery power. Devices used in IoT environment is execute only RISC and CISC oriented architecture. So, the size of the algorithm should be minimized with the consideration of high security.

Let us consider a resource-constrained device like IoT. The best solution for security achievement is light cryptography algorithms. LWC (Lightweight Cryptography) is encryption with a tiny footprint or low computing complexity. When developing a security solution for devices with limited resources, lightweight cryptography aims to use less memory, computing resources, and power than classical encryption in order to provide a more reliable security solution. When compared to traditional encryption, lightweight cryptography is believed to be less complicated and faster to implement than the latter.

In this article, we will discuss in Sect. 2 of related work from various research papers and our contribution towards our research, in Sect. 3 on lightweight block ciphers, in Sect. 4 on performance metrics for lightweight block ciphers, in Sect. 5 Cryptanalysis of Various Algorithms, in Sect. 6, the discussion of research gaps and challenges in Sect. 7 will be the conclusion.

2 Related Work

In [1], authors described light cryptography algorithms to secure the IoT environment. Discussed about the security level, chip area, throughput, latency time, hardware and software efficiency, and figure of merit are all the important factors for validating the encryption algorithm. Based upon these metrics, they concluded that AES is the most competitive algorithm which provides high-level security. They also indicated that ECC is still a viable solution for providing authentication and non-repudiation.

The creators of [2] offered a new Speck version, dubbed Speck-R, to the world. Here dynamic substitution layer had been introduced to improve security level of encryption algorithm Speck-R. The ARX (Addition, Rotation, and XOR) method of encryption is used to secure this Speck-R. The most significant contribution of this study is the number of rounds of original Speck algorithm is reduced from 26 to 7 and also high level of safety is satisfied.

In [3], the authors discuss the common and well-known attacks and threats that are affecting the different IoT design, as well as the problems identified with them. Eavesdropping on the sender's messages, identity theft, unauthorised access, trojans, and malicious software insertion into the code are all examples of risks. As a result of their research, they developed SDN-based Internet of Things designs.

The authors concentrate on the end-to-end security model, which allows the end nodes to communicate securely over an unprotected channel, as described in [4]. An IoT security middleware that is adaptable may safeguard intermittent network devices when they are connected as well as convert security protocols between cloud and edge networks in this configuration. No matter whether one of the devices is an active communication node or not, it ensures a secure connection between the two.

According to the authors of [5] the era of Internet of Things (IoT), its supporting technologies, and a complicated security strategy in conjunction with the conventional internet were offered. Many security attacks, threats, and reactions have been examined, as well as their consequences. Finally, they came to the conclusion that Internet of Things (IoT) Availability is essential. The discussion has come to a conclusion with regard to current approaches, implementation challenges, and future research objectives.

The authors of [6] presented a lightweight security system (LSS) for IoT in their paper. LSS protects the Internet of Things while lowering energy consumption. LSS is divided into three stages: The system was made immune to CPA attacks by generating secret compressed samples during the key generation, key exchange, and compression with encryption stages. The success of their technique is that it extends the network lifetime compared to existing encryption algorithms.

The authors of [7] concentrate on the security of resource-constrained systems, indicating the need for lightweight cryptographic methods. Lightweight cryptography, which is a realistic way for securing communication by modifying data, may be beneficial for Internet of Things devices with little resources. The well-defined LWC characteristics are compared and contrasted with one another. This paper highlights the research gaps and outstanding research problems that have been identified. They concluded that the block ciphers PRESENT and CLEFIA are acceptable. SIMON and SPECK are the most suitable encryption algorithms for hardware and software implementations respectively.

The performance of ten lightweight block cyphers is investigated and evaluated in the work of [8] researchers using the Raspberry Pi 3 and the Arduino Mega 2560 devices, respectively. The performance of encryption and decryption operations on payloads is measured in terms of memory usage, execution time, throughput, and energy consumption, with memory utilisation being the most important factor to consider. This research is really beneficial in establishing the most appropriate setting and encryption approach for us.

In [9], the authors proposed an algorithm that is based upon a symmetric key block cipher with a 64-bit key. Every symmetric key strategy has a number of encryption rounds, and the process of encrypting data is one of them. It is necessary to use custom substitution-permutation networks and the Feistel architecture in this case. Two fundamental concepts are applied through the usage of the Genetic Algorithm. When it comes to measurement, FELICES, a Linux-based benchmark application, is employed, whereas MATLAB is used when it comes to encryption quality testing.

Key scheduling can be used to construct the encryption keys needed for IoT devices in medical care to increase the security of data transferred in healthcare environments, according to the author of [10]. First, a unique input is transformed

into a 128-bit input key separated into four 4-bit segments. The Fibonacci scrambling algorithm is used to generate the encryption key sequences in the second stage.

A review of the many lightweight solutions and the security dangers they pose to the authentication and data integrity of the Internet of Things application can be found in [11]. The main application area of the Internet of Things has been discussed. In their examination, researchers discovered that the main security part of these protocols is to execute with the least amount of computing in order to avoid attacks such as “man in the middle,” “replay attacks,” “denial of service attacks,” “forgery,” and “chosen-ciphertext attacks,” among others. The article demonstrates how to use Microsoft’s threat modelling tool for the safe development life cycle of IoT-based applications.

In [12], a brief summary of the evolution of the Internet of Things with an emphasis on security vulnerabilities and countermeasures is proposed. Several innovative approaches to enhancing IoT security are discussed in this study, which includes cloud-fog, lightweight algorithms, block chain, machine learning, SDN/NFV, PUF, and neural networks. A discussion of cybersecurity issues such as privacy concerns, limited resources, vulnerabilities, trust management, access control, and several lightweight cryptographic techniques is provided in this work.

It is discussed in [13] how DDoS assaults inflict substantial harm to an existing system and how available solutions are utilised to fight these attacks. It also looks at resource limits in the context of resource-constrained devices and how to overcome them.

[14], proposes the unique taxonomy for IoT vulnerabilities, attacks and threats, security impacts on IoT and research impact related to security on IoT. The research contributions were discussed, covering several security issues of the IoT paradigm. This paper elaborates on the IoT vulnerabilities, Taxonomy Overview, Layers of IoT.

[15], elaborates the effects of security and privacy for some IoT features and available research challenges to be solved. This article provides up-to-date information on a variety of industries and highlights the most recent Internet of Things security research as well as how IoT aspects influence existing security research.

A simple and successful model for lightweight cipher performance measurements was devised in [16]. The devices can encrypt communications in low-energy mode using this paradigm. The algorithm balanced the encryption throughput, energy, and execution time. Their next task will be to keep an eye on unusual behaviour in the gadgets.

In [17], the authors introduced a new Fuzzy with Black Widow for cluster the query solution and Spider Monkey Optimization Algorithms query optimization. This proposed model solve privacy preserving in crowdsourcing for minimizing the cost and latency effectively. This model expresses optimal communication and computation time efficiency.

The authors in [18], presented a HCPDS (Hybrid Chaotic Particle Dragonfly Swarm Algorithm) based system for detection of DDoS attacks in VANETs. In the HCPDS approach, the dragonfly algorithm is added for enhancing the PSO updating algorithm and also, the performance metrics like processing delay, network accuracy, false alarm detection ratio and communication overhead are evaluated.

The research article [19] authors presented a hybrid crypto model for satisfying privacy and security for the cloud data. They use Elliptic Curve Cryptography (ECC) with Homomorphic for encryption. The process includes first implementing ECC at level 1 then implementing Homomorphic Algorithm. To provide more security the encryption process is done at 2 levels. After that the cipher text has been stored in the cloud. However, the implementation of this model seeks high cost.

Apart from these literature survey, our contribution of this article has summarized below,

- Our research addresses the important of the lightweight block cipher for IoT Security for providing better security without affecting the resource constraint.
- A comparison of the performance of different lightweight block cipher algorithms based on latency, throughput, chip area, security, power, and energy efficiency.
- Based on several assaults, cryptanalysis of some lightweight block cipher algorithms

3 Lightweight Block Cipher Algorithms

The majority of IoT devices have limited storage size, are small in size, and have limited resources. The following are the significant obstacles to implementing traditional cryptography algorithms:

- Limited memory
- Reduced battery power
- Real time response

RFID tags, sensors, contactless smart cards, and healthcare equipment need a lightweight cryptography method or protocol for deployment in limited contexts [20]. Hundreds of billions of heterogeneous lightweight gadgets will be connected in the future.

Lightweight cryptography is a specialty of cryptography that focuses on the optimization of encryption algorithms based on the fundamental cryptographic primitives such that they can run on small devices that have limited resources [21]. There are several types of cryptography, including:

Lightweight Block Cipher: Lightweight block cipher focuses on implementing a lightweight version of existing block ciphers and inventing new and secure cipher specifically for memory constrained devices [22]. There are two types of designs for block ciphers: Substitution-Permutation Networks and Feistel Networks.

Lightweight Stream Cipher: Lightweight stream cipher generates a key for a input data with a secret key and initialization vector. A stream cipher with low battery power low computational complexity and high level of security is called as lightweight stream cipher. Chacha and FSR (Feedback shift register)-based designs are two famous lightweight stream ciphers [21].

In order to accomplish encryption, Block Cipher makes use of Electronic Code Block (ECB) and the Cipher Block Chaining (CBC), whereas Stream Cipher makes

use of Output Feedback (OFB) and Cipher Feed-back (CFB). On the other hand, decryption of the block cipher is more difficult than decryption of the stream cipher. The implementation of the block cipher is carried out using the Feistel cipher, while the stream cipher is carried out using the Vernam cipher. The structure of a block cipher is straightforward, whereas the structure of a stream cipher is more involved.

The National Institute of Standards and Technology (NIST) has announced that FIPS 197, The Advanced Encryption Standard (AES), has been approved. The United States Government has full confidence in AES, which results in a very high level of security. In addition, it employs 192-bit and 256-bit keys for its heavy-duty encryption function [26].

When developing cryptographic algorithms for extremely low-resource devices, it is important to consider design criteria that are distinct from those used for more common devices. Despite the fact that no specific criteria for lightweight cryptography algorithms have been established, the features typically include any one or more of the following:

- the lowest feasible implementation cost
- the highest possible level of security
- the smallest size of the memory necessary for hardware implementation
- the low computing capability of microprocessors or microcontrollers

The length of the key is related to the cost and security of cryptographic algorithms, while the number of rounds in encryption provides security, performance, and hardware architecture for cryptographic algorithms that use these algorithms as well as for other algorithms that don't use cryptographic algorithms. Key length is also related to the cost of implementing cryptographic algorithms [50].

Cryptography includes two basic characteristics: To make the cipher more intriguing, Claude Shannon introduced confusion and diffusion. The link between cipher text and key is as complicated as employing a substitution box because of the ambiguity. Diffusion, on the other hand, indicates that plaintext merely generates cipher text. If a single letter in the plaintext is changed, the cipher text is completely transformed. Stream ciphers rely primarily on the property of confusion, whereas block ciphers incorporate both confusion and diffusion principles [23].

For the reasons stated above, a block cipher is favoured over a stream cipher. The focus of this research paper is on lightweight block cipher methods. Symmetric block cipher designed by the structures categorized by Feistel Network, Substitution-Permutation Network, Add-Rotate-XOR, General Feistel Network, Non-Linear Feedback Shift Register, Hybrid. Table 1 shows the structure-based categorization of several algorithms [4].

Table 1 Structure wise category of algorithms

Structure type	Description	Algorithms
SPN	The data is tweaked with the use of a set of substitution boxes and a permutation table, and the procedure is repeated defined no. of rounds	AES, Present, SKINNY, mCrypton, Iceberg, SAFER, SHARK, Square, Prince, Klein
FN	Divides the input into equal halves and applies diffusion to one half in each round	DESL/DESXL, TEA, Simon, SEA, Lblock, ITUbee, RC2, Skipjack
GFN	Splits the input data into a number of sub-blocks, with each pair of sub-blocks being applied to the Feistel function as a result of this division	CLEFIA, Piccolo, Twis, Twine, HISEC
ARX	For generating ciphertext, it combines the operations of addition, rotation, and XOR	Speck, IDEA, HIGHT, BEST-1, LEA
NLFSR	Current state is obtained from earlier state in both stream and block ciphers	KeeLoq, KATAN/KTANTAN, Halka
Hybrid	Any three types of ciphers or combination of block and stream cipher combined	Hummingbird, Present-GRP

4 Performance Metrics and Cryptanalysis

In this part of the article, we will evaluate and contrast a large number of lightweight cryptographic techniques based on a predetermined set of performance standards. In this part, we will evaluate a large number of lightweight cryptographic algorithms by contrasting them against a predetermined set of performance standards. In this part, we will evaluate a large number of lightweight cryptographic algorithms by contrasting them against a predetermined set of performance goals. The performance metrics details as follows,

- **Security performance:** It is measured in bits and can be assessed against several forms of assaults. The key size that measured in bits is the deciding factor for the security level
- **Throughput:** It is evaluated in bits and can be weighed against a variety of potential dangers. The amount of security is proportional to the key length, which is measured in bits. If it is at the maximum, then it is satisfactory. It is possible to calculate it using the formula $T = (B F)/N$, where T is throughput, B is the amount of data in bits that is encrypted or decrypted, F denoted as frequency, and N is the number of cycles take place in each block [3].

If any security attack occurs, the receiver can compute throughput from the security constraints and the channel states during the reception of the frame. The link adaptive scheme can be presented for the optimization between security and throughput.

- **Latency:** measured in terms of the number of clock cycles needed to process a single block of plaintext during encryption and cypher text during decryption. It is the equivalent of seconds. It is denoted by the equation $L = k \text{ tcycle}$, in which k represents the number of clock cycles required to compute one block of cypher text and tcycle represents the number of clock cycles required to compute one block of cypher text.

Latency can be measured in two ways: i) One Way Latency is the time taken for data to travel in one direction and it is used to diagnose the network problem. ii) Two-Way Latency is the time taken for the round-trip time for the data packet and it used to calculate Mean Opinion Score. It also called as round-trip latency.

- **Power and energy consumption:** The power and energy consumption of 8-bit and 16-bit microprocessors that operate at 4 MHz frequency with 0.9 V voltage are measured by taking the average power of the processors into consideration [3].

$$\text{Energy [J]} = (\text{Latency} [\text{number of cycles per block}] * \text{Power [W]}) / \text{block size [bits]} \quad (1)$$

The quantity of clock cycles needed to encrypt a block, the amount of power used by the hardware or software implementation, and the number of bits contained in a block of data are all described in terms of latency and power respectively.

- **Efficiency:** It indicates a balance achieved between performance and implementation size.

$$\text{Efficiency} = \text{Throughput [Kbps]} / \text{Code size [KB]} \quad (2)$$

4.1 Comparative Analysis

Over the past few years, many different types of work have been done to compare various analyses in order to determine which one is the best suitable for providing security to resource-constrained Internet of Things devices.

To optimize the encryption algorithm, we need to compare the algorithms based upon their speed, efficiency, performance and how it is to be secure the protected data against attacks. There are so many efficient new edition encryption algorithms available to decrease the security threats. The various optimization algorithms like Binary Particle Swarm Optimization, Swarm Intelligence Based Approach, Ant Colony Optimization are used for encryption algorithm to improve the performance and security.

These investigations are based on a number of trials carried out on several platforms, including NXP, AVR, and ARM microcontrollers [23]. We consider some popular lightweight block cipher algorithms and figure out their latency, throughput, security, power, and energy efficiency. The software implementation on an 8/16/32 bit microcontroller is summarised in Table 2.

Table 2 Performance metrics analysis of various algorithm

lwc algorithms	Key size	Block size	No. of rounds	Rom	Ram	Latency	Energy	Through put	SW efficiency
AES [26] [27]	128, 192, 256	128	10,12,14	918	0	4192	16.7	122	132.9
KLEIN [37]	64, 80, 96	64	12,16,20	2980	50	7901	10.6	32.4	10.87
LED [41]	64,128	64	32,48	2164	368	35,161	-	7.28	3.36
NOEKEON [47]	128	128	16	364	32	23,517	95.9	21.7	59.62
PRIDE [28]	128	64	20	266	0	1514	6	169	635.34
mCrypton [42]	64,96,128	64	12	1076	28	16,457	68	15.5	14.41
PRINCE [35]	128	64	12	1108	0	3614	14.4	70.8	63.9
ITUbee [30]	80	80	5	716	0	2607	10.4	122.7	171.37
SPECK [24]	64 72/94 96/128 96/144 128/192 156	32 48 64 96 128	22 22/23 26/27 28/29 32/33/34	134	0	408	1.6	750.5	3511.19
SIMON [24]	64 72/96 96/128 96/144 128/192/256	32 48 64 96 128	32 36 42/44 52/54 68/69/72	170	0	594	2.3	323	1900
HIGHT [36]	128	64	32	5718	47	6377	25.5	40.14	7.02
GOST [52]	256	64	32	4748	190	10,240	13.8	25	5.27
PICCOLO [52]	128,80	64	25,31	966	70	21,448	28.9	11.93	12.35
LEA [31]	128/192/256	128	24/28/32	590	32	5231	-	97.8	165.76

(continued)

Table 2 (continued)

lwc algorithms	Key size	Block size	No. of. rounds	Rom	Ram	Latency	Energy	Through put	S/W efficiency
CLEFIA [33]	128/192/256	128	18,28,26	1920	78	3646	4.9	140.42	73.14
LBLOCK [45]	80	64	32	976	58	18,988	25.6	13.48	13.81
SEA [38]	96	96	93	426	24	41,604	173.7	9.2	21.6
TEA [39]	128	64	30	648	24	7408	30.3	34.5	53.24
BORON	80/128	64	25	140	0	500	2.3	350.25	2344.20

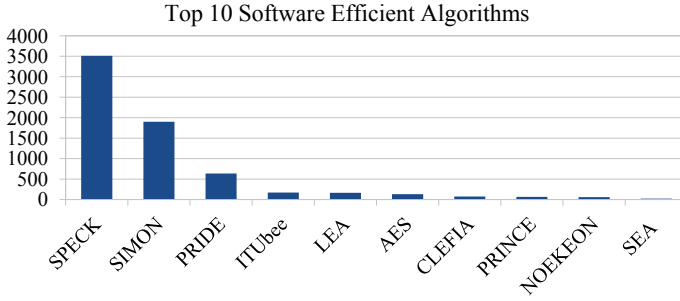


Fig. 1 Software efficient algorithms

Based upon the Fig. 1, software efficiency of various algorithms, we conclude that Speck is the best solution for IoT security. Memory power for various LWC algorithms has been shown in the Fig. 2. Again, Speck has won the competition. Other essential metrics like latency, throughput is shown in the Fig. 3, Fig. 4. Again, Speck has the lowest latency and high throughput.

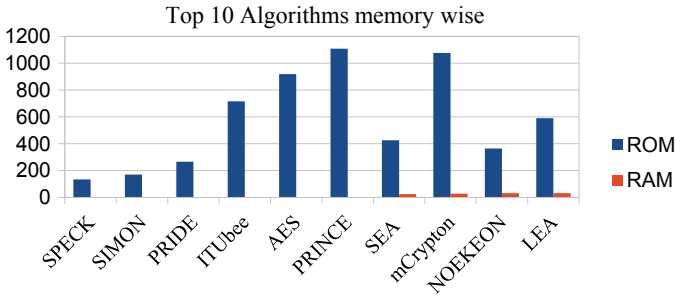


Fig. 2 Memory wise algorithms

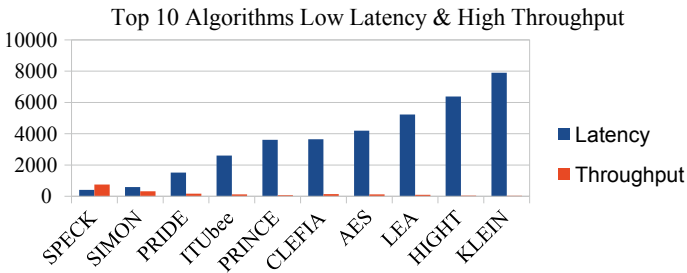


Fig. 3 Low latency and high throughput algorithms

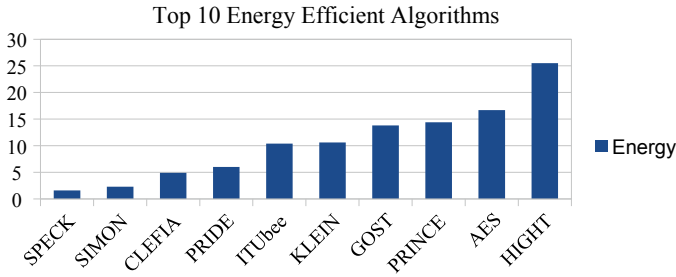


Fig. 4 Energy efficient algorithms

Table 3 Types of cryptanalysis

Cryptanalysis	Description
Differential cryptanalysis	Analysis of output against various types of inputs
Linear cryptanalysis	Experimenting with plaintext, ciphertext, and the secret key in a linear way
Integral cryptanalysis	It’s especially useful for block ciphers that use the Substitution and Permutation Network
Algebraic cryptanalysis	Based upon solving the mathematical equation

4.2 Cryptanalysis of LWC Algorithms

Security is one of the most important factors, along with performance and cost. Every LWC has a certain amount of assault resistance. However, in order to acquire our information, the attacker devises a new type of assault. As a result, examining the security element of algorithms is critical. We can get the information of security efficiency through cryptanalysis. The Table 3 depicts the many types of cryptanalysis.

These cryptanalysis employ on Cipher text Only, Known Plaintext, Chosen Plaintext, and Chosen Cipher text with Man-In-The-Middle, Brute Force, and Side Channel Attacks. Related Key Attack, Boomerang Attack, Biclique Attack, and Algebraic Attack are some of the other attacks [22]. In Table 4 the various popular algorithms are analysed based upon varous attacks and also covered merits of those algorithms discussed.

5 Research Challenges and Research Gap

The main challenge in the IoT environment is security. It has a high demand in Confidentiality, Integrity, Availability, and Authentication. Many researchers depict that cryptographic algorithms will be effective to provide security. The problem of cryptographic algorithms requires large resource allocation, large memory usage,

Table 4 Cryptanalysis of popular algorithms

Ref No	LWBC Algorithm	Merits	Attacks/Analysis
[25]	SIMON	Provides support for many key sizes and executes effectively on hardware	Attacks based on differential faults and assaults on reduced version
[25]	SPECK	Performs better in software, high security	Key recovery, Boomerang attack
[27]	AES	Excellent Security, Flexible	Related key attack, Boomerang, Biclique
[28]	PRIDE	Low Latency and low energy consumption	Differential Related Key attack on 17 th round,
[29]	PRESENT	Ultra-Lightweight cipher, Energy Efficient	Attacks such as truncated differential cryptanalysis, integral bottleneck assaults, and statistical saturation attacks
[30]	ITUbee	It provides 80-bit security against related key models as well as single level attack types. algorithm aimed toward software	Self-similarity cryptanalysis on 8-round
[32]	LEA	Fast encryption on common processors, small code size and low power	Boomerang attack on 15,16,17 th round, truncated differential attack
[34]	CLEFIA	Efficient encryption on a variety of processors, a compact code footprint, and low power consumption	In the 10th round, an attack on key recovery and saturation cryptanalysis
[35]	PRINCE	Low Latency in hardware, low energy consumption	Key recovery attack and truncated differential attack on 7 th round
[36]	HIGHT	Ultra lightweight, offers exceptional security, and is suitable for RFID tagging	Impossible differential attack on 27 th round, Integral attack is on 9th round
[37]	Klein	High software-based performance,	Biclique attack of the full round, key recovery attack up to 8 th round
[38]	SEA	Low-cost encryption, low memory, small code size,	Slide attack, Related key attack, algebraic attack
[42]	mCrypton	Cost and Energy Efficient	Related Key Attack
[48–50]	BORON	Ultra-Lightweight Cipher, Secure against DPA attack	Key Recovery attack on 8 th round, Related key attack on 10 th round

high battery power since IoT is a resource constrained environment. NIST standards represent AES as the most competitive block cipher algorithm among other algorithms. But it has a larger block size, larger rounds, and S-Box. Considering memory and computational power the traditional block cipher algorithm is not suitable for the IoT environment.

Based upon these challenges, we came to know some problem,

- Confusion is one of the two fundamental features of cryptographic algorithms, and it can be created by the use of the S-Box method. Nonetheless, S-Box takes a significant investment of time and resources.
- Larger block sizes like 128 bits, 256 bits are slowing down the computational power.
- Cryptographic algorithms are only as secure as their keys, and the key plays the most critical role in this. The goal is to produce random subkeys from the starting key for all rounds while utilising the same initial key as the first round.
- Some cryptographic algorithms have many rounds for ensuring security. But increasing the no. of rounds will affect the performance and cost. So, the problem is how to reduce the no of rounds without affecting security.

6 Future Work

To provide better security, we should think about those comparative statements and performance analysis. From the study, performance comparison and cryptanalysis we conclude that SPECK and AES have better security and limited Resource consuming, higher software efficiency. According to cryptanalysis this algorithm has some attacks. To overcome this problem, we plan to implement a reduced version of SPECK named SPECK-R. To enhance the security of this algorithm, we plan to introduce a new key scheduling algorithm. Additionally, research is going on for authentication purposes.

7 Conclusion

Nowadays IoT has become an important one in our day-to-day life. There is plenty of sensitive information that has been shared among the devices. There are many challenges for securing the IoT environment. The main security goals are Confidentiality, Integrity, and Availability. Lightweight cryptography algorithms are much better compared with traditional cryptographic algorithms since IoT devices are resource-constrained devices. Based on performance and cryptanalysis measures, we've analysed the various methods in this research study. Research difficulties and gaps in research have been addressed in this work. With a new ultra-lightweight block cipher approach to be launched in the not-too-distant future, it will be possible

to substantially increase IoT device security while still consuming little power and memory.

References

1. Dhanda SS, Singh B, Jindal P (2020) LightWeight cryptography - a solution to secure IoT. *Wirel Pers Commun* 112:1947–1980
2. Sleem L, Couturie R (2020) Speck-R: an ultra light-weight cryptographic scheme for internet of things. *Multimedia Tools Appl* 80:17067–17102
3. Iqbal W, Abbas H, Daneshmand M, Rauf B, Abbas Y (2020) An in-depth analysis of IoT security requirements, challenges and their countermeasures via software defined security. *IEEE int Things j* 7:10250–10276
4. Thakor V, Razzaque MA, Khandaker M (2020) Lightweight cryptography for IoT: a state-of-the-art
5. Khanam S, Ahmedy IB, Idris MYI, Jawarad MH, Sabri AQB (2020) A survey of security challenges, attacks taxonomy and advanced countermeasures in the internet of things. *IEEE Access* 8:219709–219743
6. Aziz A, Singh K (2019) Lightweight security scheme for internet of thing. Springer Science
7. Thakori VA, Razzaque MA, Khandaker MRA (2021) Lightweight cryptography algorithms for resource-constrained IoT devices: a review, comparison and research opportunitie. *IEEE Access* 9:28177–28193
8. Panahi P, Bayılmış Ç, Çavuşoğlu U, Kaçar S (2021) Performance evaluation of lightweight encryption algorithms for IoT-based applications. Springer, *Arabian Journal For Science And Engineering*
9. RanaS Hossain S, Shoun HI, Kashem MA (2018) An effective lightweight cryptographic algorithm to secure resource-constrained devices. *(IJACSA) Int J Adv ComputSci Appl* 9(11):267–275
10. PoojariA Nagesh H, Kiran KVG, Sangharama RC (2020) A novel key scheduling algorithmfor lightweight cryptographic applications. *Int J Adv Trends Comput Sci Eng* 9:682–684
11. Rao V, Prema KV (2020) A review on lightweight cryptography for internet of things based applications. *J Am Intell Human Comput* 12:8835–8857. <https://doi.org/10.1007/s12652-020-02672-x>
12. Ali M A Abuagoub (2019) IoT security evolution: challenges and countermeasures review. *Int J Commun Netw Inf Secur (IJCNIS)*
13. Al-Hadhrami Y, Hussain FK (2021) DDoS attacks in IoT networks: a comprehensive systematic literature review. *World Wide Web* 24:971–1001. <https://doi.org/10.1007/s11280-020-00855-2>
14. Neshenko N, Bou-Harb E, Crichigno J, Kaddoum G, Ghani N (2019) Demystifying IoT security: an exhaustive survey on IoT vulnerabilities and a first empirical look on internet scale IoT exploitations. *IEEE Commun Surv Tutor*
15. Zhou W, Jia Y, Peng A, Zhang Y, Lia P (2018) The effect of IoT new features on security and privacy: new threats, existing solutions and challenges yet to be solved. *IEEE Int Things J*
16. Bassam JM, Hayajneh T (2018) Light weight block ciphers for IoT: energy optimization and survivability techniques. *IEEE Access, Special Section on Survivability Strategies for Emerging Wireless Networks*
17. Prabakeran S (2021) Fuzzy with black widow and spider monkey optimization for privacy-preserving-based crowdsourcing system. *Soft Computing*
18. Prabakeran S, Sethukarasi T (2020) Optimal solution for malicious node detection and prevention using hybrid chaotic particle dragonfly swarm algorithm in VANETs. *Wireless Networks*
19. Saravanan P, Kumar RH, Arvind T, Narayanan B (2019) Hybrid crypto system using homomorphic encryption and elliptic curve cryptography. *i-Manager's J Comput Sci*

20. Bansod G, Raval N, Pisharoty N (2015) Implementation of a new lightweight encryption design for embedded security. *IEEE Trans. Inf. Forensics Secur* 10(1):142–151
21. Philip MA, Vaithyanathan (2018) A survey on lightweight block Ciphers for IoT Devices. In: *Proceedings IEEE region conference*, October 2018. pp 1784–1789
22. Hatzivasilis G, Fysarakis K, Papaefstathiou I, Manifavas C (2018) A review of lightweight block ciphers. *J Cryptograph Eng* 8(2):141184
23. Dinu D, Biryukov A, Groÿschädl J, Khovratovich D, Corre YL, Perrin L (2015) Felics fair evaluation of lightweight cryptographic systems. In: *Proceedings NIST Workshop Light Cryptograph*, p 128
24. Beaulieu R, Shors D, Smith J, Treatman-Clark S, Weeks B, Wingers L (2015) The simon and speck: block ciphers for internet of things. *DAC 2015*, 07–11 June 2015. San Francisco
25. Abed F, List E, Lucks S, Wenzel J (2015) Differential cryptanalysis of round-reduced Simon and Speck. *International Association of Cryptologic Research*
26. Pub N (2001) 197: Advanced encryption standard (AES). *Federal Inf Process Standards* 197(441):0311
27. Verma A, Kaur S, Chhabra B (2016) Improvement in the performance and security of advanced encryption standard using AES algorithm and comparison with blowfish. *Int Res J Eng Technol (IRJET)* 03(10):10–14
28. Albrecht MR, Driessen B, Kavun EB, Leander G, Paar C, Yağcı T (2014) Block Ciphers – Focus on the Linear Layer (feat. PRIDE). In: Garay JA, Gennaro R (eds) *Advances in Cryptology – CRYPTO 2014*, vol 8616. *Lecture Notes in Computer Science*. Springer, Heidelberg, pp 57–76. https://doi.org/10.1007/978-3-662-44371-2_4
29. Z'aba MR, Jamil N, Rusli ME, Jamaludin MZ, Yasir AAM (2014) I-PRESENT: An involutive lightweight block cipher. *J Inf Secur* 2014:25
30. Karakoç F, Demirci H, Harmanc AE (2013) ITUbee: A software oriented lightweight block cipher. In: Avoine G, Kara O (eds) *Lightweight Cryptography for Security and Privacy*, vol 8162. *Lecture Notes in Computer Science*. Springer, Heidelberg, pp 16–27. https://doi.org/10.1007/978-3-642-40392-7_2
31. Hong D, Lee J-K, Kim D-C, Kwon D, Ryu KH, Lee D-G (2014) LEA: A 128-bit block cipher for fast encryption on common processors. In: Kim Y, Lee H, Perrig A (eds) *Information Security Applications*, vol 8267. *Lecture Notes in Computer Science*. Springer, Cham, pp 3–27. https://doi.org/10.1007/978-3-319-05149-9_1
32. Kim Y, Yoon H (2014) First experimental result of power analysis attacks on a FPGA implementation of LEA. *Proc IACR 2014*:999
33. Pyrgas L, Kitsos P (2019) A very compact architecture of CLEFIA block cipher for secure IoT systems. *Euromicro Conference on Digital System Design (DSD)*
34. Tezcan C (2010) The improbable differential attack: Cryptanalysis of reduced round CLEFIA. In: Gong G, Gupta KC (eds) *Progress in Cryptology - INDOCRYPT 2010*, vol 6498. *Lecture Notes in Computer Science*. Springer, Heidelberg, pp 197–209. https://doi.org/10.1007/978-3-642-17401-8_15
35. Borghoff J, Canteaut A, Güneysu T, Kavun EB, Knezevic M, Knudsen LR, Leander G, Nikov V, Paar C, Rechberger C, Rombouts P, Thomsen SS, Yağcı T (2012) PRINCE – A Low-Latency Block Cipher for Pervasive Computing Applications. In: Wang X, Sako K (eds) *Advances in Cryptology – ASIACRYPT 2012*, vol 7658. *Lecture Notes in Computer Science*. Springer, Heidelberg, pp 208–225. https://doi.org/10.1007/978-3-642-34961-4_14
36. Hong D, Sung J, Hong S, Lim J, Lee S, Koo B-S, Lee C, Chang D, Lee J, Jeong K, Kim H, Kim J, Chee S (2006) Hight: A new block cipher suitable for low-resource device. In: Goubin L, Matsui M (eds) *Cryptographic Hardware and Embedded Systems - CHES 2006*, vol 4249. *Lecture Notes in Computer Science*. Springer, Heidelberg, pp 46–59. https://doi.org/10.1007/11894063_4
37. Gong Z, Nikova S, Law YW (2012) KLEIN: A new family of lightweight block ciphers. In: Juels A, Paar C (eds) *RFID. Security and Privacy*, vol 7055. *Lecture Notes in Computer Science*. Springer, Heidelberg, pp 1–18. https://doi.org/10.1007/978-3-642-25286-0_1

38. Standaert F-X, Piret G, Gershenfeld N, Quisquater J-J (2006) SEA: A scalable encryption algorithm for small embedded applications. In: Domingo-Ferrer J, Posegga J, Schreckling D (eds) *Smart Card Research and Advanced Applications*, vol 3928. *Lecture Notes in Computer Science*. Springer, Heidelberg, pp 222–236. https://doi.org/10.1007/11733447_16
39. Andrews B, Chapman S, Dearstyne S (2020) Tiny encryption algorithm (TEA) cryptography 4005.705. 01 graduate team ACD_nal report, Rochester Inst. Technol., Rochester, NY, USA, Tech. Rep. 33695183. https://www.coursehero.com/_le/33695183/TEApdf/
40. Sekar G, Mouha N, Velichkov V, Preneel B (2011) “Meet-in-the-Middle Attacks on Reduced-Round XTEA”, *The Cryptographers’ Track at the RSA Conference 2011*. CA, USA, San Francisco
41. Guo J, Peyrin T, Poschmann A, Robshaw M (2011) The LED block cipher. In: Preneel B, Takagi T (eds) *Cryptographic Hardware and Embedded Systems – CHES 2011*, vol 6917. *Lecture Notes in Computer Science*. Springer, Heidelberg, pp 326–341. https://doi.org/10.1007/978-3-642-23951-9_22
42. Lim CH, Korkishko T (2006) mCrypton – A Lightweight Block Cipher for Security of Low-Cost RFID Tags and Sensors. In: Song J-S, Kwon T, Yung M (eds) *Information Security Applications*, vol 3786. *Lecture Notes in Computer Science*. Springer, Heidelberg, pp 243–258. https://doi.org/10.1007/11604938_19
43. Standaert F-X, Piret G, Quisquater J-J (2003) Cryptanalysis of block ciphers: a survey. UCL Crypto Group Laboratoire de Microelectronique Universite Catholique de Louvain
44. Mala H, Dakhilalian M, Shakiba M (2011) Cryptanalysis of mCrypton—A lightweight block cipher for security of RFID tags and sensors. *Int J Commun Syst*
45. Wu W, Zhang L (2011) LBlock: A lightweight block cipher. In: Lopez J, Tsudik G (eds) *Applied Cryptography and Network Security*, vol 6715. *Lecture Notes in Computer Science*. Springer, Heidelberg, pp 327–344. https://doi.org/10.1007/978-3-642-21554-4_19
46. Knudsen L, Leander G, Poschmann A, Robshaw MJB (2010) PRINTCipher: A Block Cipher for IC-Printing. In: Mangard S, Standaert F-X (eds) *Cryptographic Hardware and Embedded Systems, CHES 2010*, vol 6225. *Lecture Notes in Computer Science*. Springer, Heidelberg, pp 16–32. https://doi.org/10.1007/978-3-642-15031-9_2
47. Daemen J, Peeters M, Assche G, Rijmen V (2000) The Noekeon block cipher. In *Proceedings 1st Open NESSIE Workshop* pp 1–5
48. Bansod G, Pisharoty N, Patil A (2013) BORON: an ultra-lightweight and low power encryption design for pervasive computing. *Front Inf Technol Electron Eng* 18:317–331
49. Liang H, Wang M (2019) Cryptanalysis of the lightweight block cipher BORON. *Secur Commun Netw* volume 2019, article ID 7862738
50. Teha JS, Thama LJ, Jamil N, Yapd W-S (2021) New differential cryptanalysis results for the lightweight block cipher BORON. *J Inf Secur Appl*
51. Alizadeh M, Salleh M, Zamani M, Shayan J, Karamizadeh S (2015) Security and performance evaluation of lightweight cryptographic algorithms in RFID. *Recent Researches in Communications and Computers*
52. Aldabbagh SSM, Fakhri I, Shaikhli T, Sulaiman AG (2016) Lightweight Block Cipher Algorithms: Review Paper. *Int: International Journal of Enhanced Research in Science, Technology & Engineering*, vol 5 issue 5, ISSN: 2319–7463
53. Hassija V, Chamola V, Saxena V, Jain D, Goyal P, Sikdar B (2019) A survey on IoT security: application areas, security threats, and solution architectures. *IEEE Access* 7:5–10
54. Prabakaran S (2021) Pulmonary disease diagnosis using African vulture optimized weighted support vector machine approach. *International Journal of Imaging Systems and Technology (IMA)*
55. Indumathi V, Prabakeran S (2021) A Comparative Analysis on Sensor-Based Human Activity Recognition Using Various Deep Learning Techniques. In: Pandian AP, Fernando X, Islam SMS (eds) *Computer Networks, Big Data and IoT*, vol 66. *Lecture Notes on Data Engineering and Communications Technologies*. Springer, Singapore, pp 919–938. https://doi.org/10.1007/978-981-16-0965-7_70

56. Prabakeran S (2021) Women's mental health chatbot using seq2seq with attention. *Turkish J Comput Math Educ* 12(10):919–938
57. Saravanan P, Sethukarasi T, Indumathi V (2018) An efficient software defined network based cooperative scheme for mitigation of distributed denial of service (DDoS) attacks. *J Comput Theor Nanosci* 15:2221–2226
58. Saravanan P, Sethukarasi T, Indumathi V (2018) Authentic novel trust propagation model with deceptive recommendation penalty scheme for distributed denial of service attacks. *J Comput Theor Nanosci* 15:2383–2389
59. Saravanan P, Sekar S, Kulam S, Selvaraj S (2018) An automatic helmet detection and penalty system using image descriptors and classifiers. *J Comput Theor Nanosci* 15:2245–2250
60. Murugeswari B, Sudharson K, Panimalar SP, Shanmugapriya M, Abinaya M (2020) SAFE – secure authentication in federated environment using CEG key code (A Novel Method to Enhance Cloud Security). The Mattingley Publishing Co, Inc
61. Chatterjee R, Chakraborty R, Mondal JK (2019) Design of lightweight cryptographic model for end-to-end encryption in IoT domain. *IRO J Sustain Wireless Syst* 1(4):215–224
62. Dhaya R (2021) Light weight CNN based robust image watermarking scheme for security. *J Inf Technol Digital World* 3(2):118–132

Image Classification Using Quantum Machine Learning



Amrit Raj and Jayakumar Vaithiyashankar

Abstract When quantum algorithms are used in machine learning systems, it is referred to as “quantum machine learning.” An approach known as “quantum-enhanced machine learning” utilizes a quantum computer to evaluate classical data to boost machine learning. Data may be processed and stored more quickly and efficiently with the help of quantum machine learning. Using neural networks as analogies for physical systems is an important part of quantum machine learning. This paper summarizes the CIFAR-10 dataset. For the dataset, “five training batches and one testing batch” are used to divide the ten thousand photographs. One thousand images from each class are randomly selected for inclusion in the test batch. Even though each batch comprises all of the remaining photographs, some batches have a greater number of images from a particular category. It is estimated that each training batch contains around 5000 photographs. This section includes an evaluation of the classifier’s overall performance. Quantum neural networks describe “a parameterized quantum computational model best” implemented “on a quantum computer” (QNN). Third-party libraries such as PyTorch, Qiskit, and matplotlib are frequently loaded into the program. PyTorch is a popular option for GPU and CPU-based Deep Learning applications because it is built on Torch rather than merely Python. All of your quantum computing needs may be met by Qiskit, a Python library. Importing it will be necessary after the system is installed. Creating static, animated, and interactive graphics is easy with Matplotlib, a Python toolkit. To begin, we need to identify the quantum layers that will make up the circuit’s structure.

Keywords Quantum machine learning · Matplotlib · GPU & CPU based deep learning · PyTorch

A. Raj (✉) · J. Vaithiyashankar
Department of CSE Galgotias University, Greater Noida, Uttar Pradesh 201301, India
e-mail: Amritraj825@gmail.com

J. Vaithiyashankar
e-mail: jayakumar.v@presidencyuniversity.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
P. P. Joby et al. (eds.), *IoT Based Control Networks and Intelligent Systems*,
Lecture Notes in Networks and Systems 528,
https://doi.org/10.1007/978-981-19-5845-8_26

367

1 Introduction

We derive our classical understanding from daily living, but this is not the fundamental mechanism at work in nature. Quantum mechanics, the underlying and more fundamental mechanics, is just beginning to manifest in our surroundings. Our everyday intuition does not fit quantum phenomena. In truth, as far back as we can go in human history, we have been baffled by these fundamental physics. Nature has only been revealed to us in the last century. We developed theories and mathematical tools as the research advanced, thanks to our renowned scientists. Due to its status as a probabilistic theory, quantum theory has sparked numerous philosophical arguments. Expanding permutations are the key to quantum computing's ability to store twice as much data per qubit as conventional computers. For a classical bit system with "N bits of binary numbers", we require N bits of binary numbers [14].

1.1 *Classical Machine Learning*

An industrial control system (ICS) of the modern era is a cyber-physical one. It has a similar network infrastructure to that of a corporate system. Aside from that, it's equipped with tools for monitoring and controlling industrial processes. A cyber attack on an industrial control system (ICS) is extremely risky because of the damage it can wreak. In addition, these attacks are frequently difficult to detect. To address these issues, a new, intelligent method of identifying assaults based on anomaly detection is being developed. Cyber assaults on industrial control systems (ICS) can express themselves in a variety of ways, including abnormal behavior in both software and hardware [3].

A.) **"MACHINE LEARNING FROM DATA AND LEARNING FROM INTERACTION"**

Co-creation of value in collaborative networks is becoming increasingly dependent on the sharing of data between organizations. Most corporate and scientific research applications necessitate anonymization when sharing data across organizational boundaries. This data can be used for a second purpose, however, pseudonymization or anonymization is required depending on the intended use, according to the European General Data Protection Regulation (GDPR). Anonymous data and machine learning algorithms' ability to process highly anonymized, unstructured textual input is not well understood. There are numerous benefits to collaborating with other firms through sharing data as shown in Fig. 1 [8].

- **Supervised Learning**

The machine learning technique called "supervised learning" is the reduction of an input-output relationship to a small collection of training samples. There are a variety

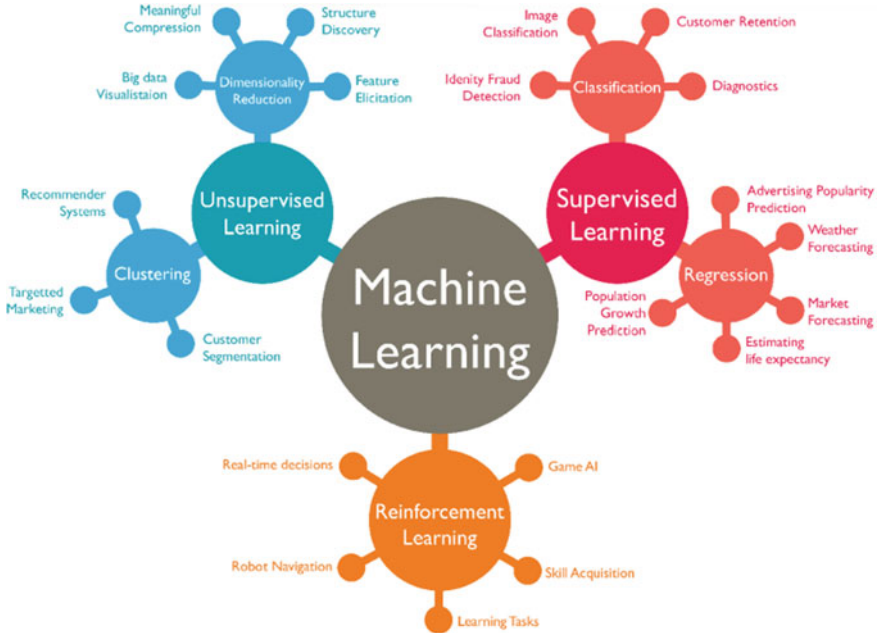


Fig. 1 Explanation of machine learning

of names for this type of data, including labeled training datasets and supervised training datasets. Inductive machine learning is also known as learning from labeled data or learning by induction [17]. These are of various types are as follows:

- Model Selection
- Learning
- Interference

• **Unsupervised Learning**

Using no labels and as little human interaction as possible, an unsupervised machine learning approach seeks patterns in a dataset that have never been discovered previously. Probability densities can be modelled across inputs rather than utilizing human-labeled data. This is known as unsupervised learning. Reinforcement and supervised learning are the other two main types of machine learning. Unsupervised and supervised procedures are also used in semi-supervised learning [20].

These are of various types are as follows:

- Density Estimation
- Clustering
- Dimensionally Reduction

- **Reinforcement Learning**

As a term for detecting structure concealed in large swaths of data that haven't been labeled, machine learning researchers call "unsupervised learning," reinforcement learning differs from that. Machine learning paradigms do not fit well into the categories of supervised and unsupervised. It is tempting to conceive of reinforcement learning as an unsupervised method because it does not rely on examples of the right behavior, but reinforcement learning aims to maximize a reward signal rather than seek out a hidden structure [18].

Machine Learning

Computer programs that can learn from data are at the heart of Machine Learning (ML). It is possible to contrast machine learning's inductive inference, i.e., the generalizations from a set of observed instances, against the deductive inference of early Artificial Intelligence (AI) systems. There are numerous other scientific areas where machine learning (ML) interacts such as statistics, cognitive science, and information theory, even though it is a subfield of AI. Data mining is a branch of machine learning that focuses on discovering interesting patterns in massive datasets. A wide range of real-world issues has been addressed using machine learning approaches over the years, in addition to speech recognition and fraud detection, customer relationship management, gene function prediction, and, so on. Finally, we sum up our findings and address the future of machine learning on the Semantic Web [1].

A.) Artificial neural networks-

generative artificial neural networks. Many diverse methods and approaches from both statistics and computer science fall under this umbrella term. It would take a long time to go over every possible use of this powerful tool, so our goal is to get a handle on the basics and see how it may be put to use. We begin by comparing neural networks to the human brain [5].

These are of various types are as follows:

- **Feed Forward Neural Networks**

The biological brain and nervous system are the inspiration for artificial neural networks, which go by the moniker of "artificial neural networks."

- **Convolutional Neural Networks**

Many of the tools we use daily are now powered by deep learning. Students and professionals alike are clamoring to learn and implement this technology because of its enduring success and wide range of possible applications.

When it comes to understanding CNNs, this is the most significant obstacle intricacy of the interplay between low-level math procedures and higher-level concepts and high-level network integration of these tasks [7].

Support vector machines of supervised learning:

Unlike neural networks and SVMs, they were developed in the opposite order (NNs). Examples include SVMs, which went from sound theory to implementation

and experiments, while NNs went from applications and extensive experimentation to theory. Although SVMs have a theoretical underpinning, they haven't gained much popularity. Even while SVMs are theoretically fascinating, the statistics and/or machine learning community often considers them unsuitable or irrelevant. The only time they were taken seriously was when they performed exceptionally well on practical learning standards (in numeral recognition, computer vision, and text categorization).

- **Computational Learning Theory**

It is the goal of computer science theory to develop learning computer systems and to understand the computational limitations of machine learning. A branch of theoretical computer science is computational learning theory. The ability of learning algorithms to solve sample problems has traditionally been used by artificial intelligence researchers. Such evaluations are difficult to make meaningful comparisons between competing algorithms, even if they provide a wealth of valuable information and insight [19].

2 Quantum Machine Learning

It has been more than half a century since machine learning began, and with the advancement in computing power, it has become a vital aspect of computer science. Unsupervised learning supervised learning, and reinforcement learning is the most common types of machine learning. In this examination, we will not focus on the principles of these three forms of learning, but rather on the algorithms that implement them in machine learning. Even though processing power has increased significantly over the past few decades and new algorithms have been developed regularly, the increase in data has outpaced this expansion. Machine learning, which is heavily reliant on huge data, will eventually suffer from a lack of computational capacity as a result. The principles of quantum physics, such as superposition and entanglement, underlie quantum computing. Parallelism is a crucial element for high-speed computing; therefore, it can be tailored to suit specific issues. It is difficult to answer classical problems as quickly as in the quantum world. An exponential speedup can be achieved by employing Shor's algorithm1 to solve the problem of huge integer factorization, which cannot be achieved by any classical method. Quantum algorithms to solve specific issues are then proposed. A quadratic speedup in the search of an unstructured database has been demonstrated for Grover's technique 2. It's possible that quantum computing and machine learning as in Fig. 2, could work well together because machine learning is being hampered by a shortage of computer capacity, therefore this is being explored. Some progress has been made recently in the development of a quantum computing device [22].

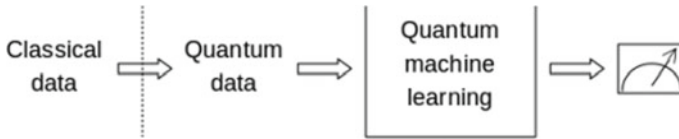


Fig. 2 The fundamentals of quantum machine learning. Quantum data is used in the learning process

2.1 Review of Literature

[6] One of the most quickly emerging scientific fields today is classical machine learning theory and quantum computation theory. In recent years, academics have looked into whether quantum computing can aid in the improvement of classical machine learning. Hybrid approaches combining classical and quantum algorithms are part of quantum machine learning. Quantum techniques can be utilized instead of classical data to analyze quantum states. Quantum algorithms, on the other hand, have the potential to revolutionize data science. Quantum machine learning concepts are introduced in this section. Machine learning algorithms and quantum computing approaches can be combined in a variety of ways. On an IBM quantum processor, we present the multiclass tree tensor network technique.

[9] The intersection of quantum computing and classical machine learning is the focus of quantum machine learning. An investigation of quantum computing and machine learning is conducted, to discover how the two fields may work together to solve issues. Computer models that have been in use for decades are rapidly reaching their theoretical limits. As a result, quantum computing can provide an advantage in certain kinds of machine-learning tasks.

[2] Interact when the attribute values of both qualities may be observed in conjunction. An example of a two-way interaction is the one seen above. There are k -way interactions if we can't rebuild the relationship between a collection of qualities X and l -way interactions, which is the case in most cases. In a nutshell, these two definitions capture the concept of interaction. The concept of context is another important one to keep in mind.

[12] Machine Learning from User Interaction for Visualization and Analytics was organized at IEEE VIS 2018. We wanted to bring together researchers from the visualization community to talk about how machine learning can help visualizers, with a focus on how users' interactions may help enhance visualization systems. A summary and categorization of participants' comments and suggestions were completed at the end of the workshop. This work's research agenda is the culmination of this compilation.

[10] IoT (Internet of Things), cybersecurity, mobile, business, social media, and health-related data are just a few of the many types of data available in the digital age. Data analysis and application development require the use of artificial intelligence (AI) and in particular machine learning (ML). Methods such as reinforcement learning and unsupervised learning can be found in the industry.

[21] The class of machine learning known as “deep learning” does substantially better on unstructured data than other classes. Currently, deep learning approaches are outperforming current machine learning methods. Computer models can learn features incrementally from data at different levels. Deep learning’s popularity grew as the amount of data available and the development of powerful computers made it more accessible.

This article discusses a wide range of deep learning approaches, architectures, methodologies, and applications.

[13] There are several applications of deep learning in signal and information processing that are covered in this monograph. For this project, we focused on three types of applications: those that have already been transformed by deep learning technology, like speech recognition and computer vision, as well as those that have the potential to be significantly impacted by deep learning and are seeing increased research activity, like natural language processing and text analysis.

[11] Reinforcement learning (RL) and deep learning (DL) are combined in deep reinforcement learning (DRL). Complex decision-making problems previously out of the grasp of a machine have been solved through this branch of research. As a result, deep reinforcement learning opens up a slew of new possibilities in fields as diverse as healthcare, robotics, smart grids, and finance. Deep reinforcement learning models, methods, and approaches are introduced in this book. The elements of generalization and practical application of deep RL are of particular interest. Machine learning ideas are assumed to be familiar to the reader.

[4] Analysis of learning models and their pattern categorization ratings in a higher education environment is discussed here. While the error back-propagation method supplied by the supervised learning model is particularly efficient for many non-linear real-time problems, the current study’s KSOM of the unsupervised learning model provides an efficient solution and categorization.

[15] Big data and the discipline of data science as a whole are increasing at a similar rate to machine learning. Research publications published between 2015 and 2018 addressing or using supervised or unsupervised machine learning approaches in various problem-solving paradigms were the focus of this systematic review. PRISMA features were used to identify 84 scientific publications published in various journals during the evaluation process.

[16] Quantum Machine Learning was developed as a means of bridging the divide between quantum computing and machine learning. Machine learning methods are explained in terms of quantum mechanics, which simplifies and unites the various disciplines. Computer experts, and even researchers in the field, have a difficult time keeping up with the latest quantum computing theoretical breakthroughs. The lack of a step-by-step guide to this emerging interdisciplinary body of study hinders its wider understanding.

3 Methodology

“Quantum Convolutional Neural Network”

To use the PyTorch interface in Penny Lane, firstly install PyTorch.

Pre-requisite

Data modeling, programming language, and some knowledge of statistics and probability are required. To succeed in the profitable field of machine learning you need a lot of effort and expertise, but it is well worth it. This is not a task that can be completed in a single day.

All the most commonly used PyTorch libraries, as well as the Qiskit quantum software framework and matplotlib, are imported.

Pytorch

It is designed for GPU and CPU-based applications and is built on Python and Torch as a Deep Learning Tensor Library. A key advantage of PyTorch over Deep Learning frameworks like Keras and TensorFlow is the fact that it uses dynamic computation graphs.

PyTorch has two primary features:

GPU (Graphical Processing Unit) acceleration for Tensor Computation (similar to NumPy).

QISKIT

Qiskit is a Python library for all of your quantum computing needs. Installing it is a must if it isn't already done. You'll have to import it when it's been installed.

“MATPLOTLIB”

Mathematical plotting libraries such as Matplotlib allow Python programmers to create dynamic, animated, and interactive graphs. Matplotlib makes simple tasks simple and difficult tasks attainable.

- Create work worthy of publishing.
- Make figures that can be zoomed, panned, and updated in real-time.
- Style and layout can be altered to suit your needs.
- Many different file types can be exported.
- Embedded in Jupyter Lab and GUIs.
- Matplotlib is the foundation for several third-party tools.

3.1 Model Building and Training

We first define some quantum layers that will compose the quantum circuit. We follow the transfer learning approach: First load the classical pre-trained network ResNet18

from the torch vision. model's zoo. Freeze all the weights since they should not be trained. Replace the last fully connected layer with our trainable dressed quantum circuit (Quantum net). Alternatively, if quantum == False, an entirely classical analogue is used.

3.2 *Quantum Variational Circuits*

Variational circuits are also called parameterized quantum circuits. They play a role in Quantum Computing akin to the role played by Neural Networks in Classical Computing.

“Pretrained Models”

Using a pre-trained model means that the model has already been trained on a specific dataset, and it has the weights and biases that represent the features of the dataset in which it was tested. Learned characteristics can frequently be applied to new types of data.

Resnet-18

Deep residual learning framework for picture classification. This allows for a good balance between productivity and quality by supporting a variety of architectural options. ResNet-18 is a convolutional neural network with 18 layers of processing. Use a network that has already been trained on millions of photos from the ImageNet to get started.

The setting of the main hyper-parameters of the model:

1. **Number of qubits used = 4**

One qubit can encode the entirety of a single bit. With superdense coding, a qubit can hold up to two bits of information instead of the standard one bit.

2. **Learning rate = 0.001**

The learning rate is a hyperparameter with a modest positive value, often between 0.0 and 1.0, that is used to train neural networks. There are two ways to control the learning rate of an algorithm.

3. **Batch size = 32**

The “batch size,” or simply “batch,” is a hyperparameter for the learning algorithm that determines how many training samples are utilized to estimate the error gradient. 32 samples are used to estimate the error gradient before updating the model weights in a batch size of 32.

4. **Number of training -epochs = 5**

The technique of preparing athletes according to scientific concepts to enhance and sustain their performance.

5. **Depth of the quantum circuit = 5**

The longest path from input to output or measurement gate, traveling forward in time along qubit wires, is known as circuit depth.



Fig. 3 A batch of the test data

Dataset Loading

The PyTorch packages torchvision and torch.utils.data are used for loading the dataset.

```
trainset_full = torchvision.datasets.CIFAR10(root='./data', train=True, download=True, transform=data_transforms['train'])
testset_full = torchvision.datasets.CIFAR10(root='./data', train=False,
                                           download=True, transform=data_transforms['val'])
image_datasets_full={'train': trainset_full, 'val': testset_full}
```

We can initialize the CIFAR training set using train set_full = torchvision.datasets.CIFAR10 with the parameters root = './data', train = True, download = True, and transform = data_transforms['train'].

Also performing standard preliminary image operations:

- **RESIZE:**

Resizing an image alters its pixel information. When an image is reduced in size, for example, the photo editor discards any unnecessary pixel information (Photoshop).

- **CENTRE-CROP:**

When a picture is cropped from the center, an equal amount of padding is added to both the vertical and horizontal sides.

- **NORMALIZE**

Changing the range of pixel intensity values, which is a common image processing technique, is called “image normalization.”

- **HORIZONTAL FLIP**

Layers and selections are flipped when you pick Flip. Layers and selections are mirrored horizontally when you select Mirror (left to right).

Showing a batch of the test data as in Fig. 3, just to have an idea of the classification problem.

3.3 Variation Quantum Circuit: Variational Quantum Circuit

The structure is that of a typical variational quantum circuit:

We first define some quantum layers that will compose the quantum circuit.

- **EMBEDDING LAYER:** All qubits are initially initialized in a balanced superposition of up and down states and then rotated based on the input parameters (local embedding).
- **VARIATIONAL LAYERS:** A series of trainable rotation layers and constant entangling layers are applied sequentially.
- **MEASUREMENT LAYER:** Finally, the local expectation value of the Z operator is calculated for each qubit.

Now define a custom torch. No. Module representing a dressed quantum circuit.

This is a concatenation of:

- A classical pre-processing layer (nn. Linear).
- A classical activation function (torch.tanh).
- A constant $\pi/2.0$ scaling.
- The previously defined quantum circuit (quantum_net).
- A classical post-processing layer (nn. Linear).

The module takes a batch of vectors with 512 real parameters (features) as input and produces a batch of vectors with n real outputs as outputs (one for each class of images).

3.4 Finally Ready to Build Our Full Hybrid Classical-Quantum Network

We follow the transfer learning approach:

- First load the classical pre-trained network ResNet18 from the torchvision.models zoo
- Freeze all the weights since they should not be trained.
- Replace the last fully connected layer with our trainable dressed quantum circuit (Dressed Quantum Net).

The ResNet18 model is automatically downloaded by PyTorch and it may take several minutes (only the first time).

4 Training and Results

Before training the network, we need to specify the loss function.

The function takes parameters as optimizer, criterion, model, num epochs, and scheduler.

Table 1 Performance evaluation results

“Training accuracy”	“Training loss”	“Testing accuracy”	“Testing loss”
0.7144	0.5546	0.7860	0.4831

- We use, as usual in the classification problem, the cross-entropy which is directly available within the torch. no.
- Also, initialize the Adam optimizer which is called at each training step to update the weights of the model.
- Schedule to reduce the learning rate by a factor of gamma_lr_scheduler every 10 epochs.

Visualizing the Model Predictions

First, define a visualization function for a batch of test data.

To load the pre-trained weights, it is necessary to first define the model. So, before this cell, one should have run all the cells above the Training and results section.

Results: As a result, comprehending physical systems and designing quantum algorithms are among the goals of this research as in Table 1.

5 Conclusion and Future Work

In this work, we summarized the impact of quantum computers on machine learning so far and in the future. Only a few years ago, most of the research in these fields was purely theoretical. Now, we have provable quantum machine learning algorithms. New algorithms are faster and more efficient than the previous ones, as expected. This technology has the potential to revolutionize the field of machine learning. When larger quantum computers with more qubits come to fruition, we’ll be able to test more quantum algorithms and discover how quantum computers affect machine learning. Quantum machine learning appears to be a methodology since it will lead to a better future in today’s society, where large volumes of data are generated every day and processed minute-by-minute, new and creative research methods can have major effects on life and the economy. Our efforts in this direction are heavily focused on the computational aspects of machine learning (ML).

References

1. Ławrynowicz A (2014) Introducing machine learning. Research Gate
2. Jakulin A (2013) Machine learning based on attribute interactions. Sežana
3. Sokolov A (2018) Research of classical machine learning methods and deep learning models effectiveness in detecting anomalies of industrial control system. Research Gate.

4. Abraham A (2013) Comparison of supervised and unsupervised learning algorithms for pattern classification. Research Gate.
5. Gallo C (2015) Artificial Neural Networks: a tutorial. Research Gate
6. Fastovets DV (n.d.) Machine learning methods in quantum computing theory. Valiev Institute of Physics and Technology of Russian Academy of Sciences, Russia
7. Hohman F (2020) CNN explainer: learning convolutional neural networks with interactive visualization
8. Ketamo H (2019) Interactive machine learning: managing information richness in highly anonymized conversation data. Research Gate.
9. Rakesh H (2019) Quantum machine learning: a review and current status. Research Gate
10. Iqbal HS (2021) Machine learning: algorithms, real-world applications, and research directions
11. Pineau J (2018) An introduction to deep reinforcement learning
12. Wenskovich J (n.d.). Machine learning from user interaction for visualization and analytics: a workshop-generated research agenda
13. Deng L (2014) Deep learning: methods and applications. 07
14. Mishra N et al (2021) Quantum machine learning: a review and current status. Adv Intell Syst Comput 1175:101–145. https://doi.org/10.1007/978-981-15-5619-7_8
15. Alloghani MA (2020) A systematic review on supervised and unsupervised machine learning algorithms for data science. Research Gate.
16. Wittek P (2014) Quantum machine learning: what quantum computing means to data mining. Research Gate
17. Liu Q (2012) Supervised learning. Research Gate
18. Richard S (2015) Reinforcement learning: an introduction
19. Goldman SA (n.d.) Computational learning theory
20. Siadati S (2018) What is unsupervised learning? Research Gate
21. Sivakumari S (2021) Deep Learning Techniques: An Overview. Research Gate
22. Zhang Y (n.d.) Recent Advances in Quantum Machine Learning. School of Computing and Communications, Lancaster University, United Kingdom.

An Efficient Algorithm for Multi Class Classification in Deep Neural Network



Pranamita Nanda and N. Duraipandian

Abstract In comparison to other machine learning techniques, deep neural networks are effective in classifying non-linearly separable data. Because of its simplicity, contemporary gradient-based algorithms such as momentum Stochastic Gradient Descent (SGD) are commonly employed in Deep Neural Networks (DNN). However, the process of convergence is slowed by the choice of an appropriate learning rate and the local minima problem. To address these issues, this research proposes a unique approach for training DNNs called Simulated Annealing Based Gradient Descent (SAGD), which involves optimizing weights and biases. The SAGD technique optimizes the function by combining gradient information with the simulated annealing notion. The learning rate does not need to be manually adjusted with this method. Instead, using the simulated annealing approach, the learning rate is modified automatically for each epoch. The approach is tested utilizing VGG16, ResNet 18 and InceptionV3 architectures on typical multi-class classification data sets such as Iris, MNIST, and CIFAR10. The performance of SAGD and other state-of-the-art gradient descent optimization methods is compared, and it is demonstrated that SAGD performs comparably to existing gradient descent methods.

Keywords DNN · Classification · Simulated annealing · Optimization · Gradient descent

1 Introduction

The problem in which the instances are classified as one out of three or more classes is called as multi class classification. Multi class classification has several application

P. Nanda (✉)

Department of Computer Science and Engineering, Velammal Institute of Technology, Chennai 601204, India

e-mail: pranamita.nanda@gmail.com

N. Duraipandian

Department of Computer Science and Engineering, Saveetha Engineering College, Chennai 602117, India

areas like face recognition, biometric identification etc. The most common neural networks are deep neural networks, which are trained using a back-propagation learning method [1, 2]. The network is made up of neurons arranged in layers. The input layer is the first layer, the output layer is the last layer, and the intermediate levels are called hidden layers [2, 3]. The complexity of the network depends on the number of layers and the number of neurons in each layer. The DNN with few parameters will not be able to classify a non-linearly separable data and also a network trained with too much of parameters may produce over-fitting. Several machine learning algorithms have been applied on classification problems [4]. But for non-linearly separable data Deep Neural Network performs better than the other machine learning algorithms. Hence here DNN is used with the various optimization algorithms, activation functions along with the hyper parameter tuning like regularization, dataset augmentation and early stopping. The activation function transforms a neuron's activation level into an output signal [5].

This paper develops a method to optimize the weights and biases of the deep neural network. It is based on the existing Gradient Descent (GD). A strength of GD is that they are simple to implement and also fast for problems that have many training examples [6]. However, the drawback of Gradient Descent is that it suffers from local minima as well as hyper parameter tuning. The Simulated Annealing algorithm is a probabilistic heuristic that approximates the global minimum of an optimization problem by simulating the annealing process in metallurgy [7].

In this article the drawback of gradient descent is overcome by applying the condition of simulated annealing for updating of weights and biases as well as auto-tuning of learning rate.

2 Related Work

The performance of the DNN can be improved by improving the learning algorithms, activation functions, initialization methods and regularization methods.

2.1 Initialization

Xavier [23] and He [24] Weight initialization with various nonlinear activation functions like sigmoid, ReLU etc. is done [8].

2.2 Optimization

The commonly used Gradient Descent update rule for weights is

$$w_{i+1} = w_i - \eta \nabla w_i \tag{1}$$

where η is the learning rate.

For computing the gradients, the strategies used are batch, mini-batch and stochastic. Because they process all of the training examples in a big batch, optimization algorithms that employ the complete training set are known as batch or deterministic gradient techniques [9]. Stochastic optimization methods are those that employ just a single example at a time.

The gradients are used in different ways to develop better learning algorithms: [10, 11, 25, 26]. Few of the algorithms are given below:

- *Momentum based gradient descent*

Momentum based gradient descent is able to take large steps even in the regions having gentle slopes. It oscillates in and out of the minima valley.

$$p_i = \gamma * p_{i-1} + \eta \nabla w_i \tag{2}$$

$$w_{i+1} = w_i - p_i \tag{3}$$

- *Adagrad*

In Adagrad, parameters corresponding to sparse features get better updates, but the learning rate decays belligerently.

$$p_i = p_{i-1} + (\nabla w_i)^2 \tag{4}$$

$$w_{i+1} = w_i - \frac{\eta}{\sqrt{(p_i) + \epsilon}} \nabla w_i \tag{5}$$

- *RMSProp*

In comparison to Adagrad it decays the denominator slowly.

$$p_i = \beta * p_{i-1} + (1 - \beta) * (\nabla w_i)^2 \tag{6}$$

$$w_{i+1} = w_i - \frac{\eta}{\sqrt{(p_i) + \epsilon}} \nabla w_i \tag{7}$$

- *Adam*

It combines the advantage of RMSProp and momentum based gradient descent

$$s_i = \beta_1 * p_{i-1} + (1 - \beta_1)(\nabla w_i) \tag{8}$$

$$p_i = \beta_2 * p_{i-1} + (1 - \beta_2)(\nabla w_i)^2 \quad (9)$$

$$w_{i+1} = w_i - \frac{\eta}{\sqrt{(p_i) + \epsilon}} s_i \quad (10)$$

2.3 Activation Functions

The common activation functions used in neural networks are [6, 12]:

- *Sigmoid Activation*

$$g(y) = \frac{1}{1 + e^{-y}} \quad (11)$$

$$g'(y) = f(y) * (1 - f(y)) \quad (12)$$

- *Tanh Activation:* The hyperbolic tangent function and its gradient is given by

$$g(y) = \tanh(y) = \frac{e^y - e^{-y}}{e^y + e^{-y}} \quad (13)$$

$$g'(y) = 1 - (g(y))^2 \quad (14)$$

- *ReLU Activation:* The Rectifier Linear Unit and its gradient is:

$$g(y) = \max(0, y) \quad (15)$$

$$\text{and } g'(y) = \begin{cases} 0 & \text{if } y \leq 0 \\ 1 & \text{if } y > 0 \end{cases} \quad (16)$$

- *Leaky ReLU Activation:*

To overcome the 'dying ReLU condition', Leaky ReLU is used which is given by the function:

$$g(y) = \max(\alpha y, y) \quad (17)$$

$$g'(y) = \begin{cases} \alpha & \text{if } y < 0 \\ 1 & \text{if } y \geq 0 \end{cases} \quad (18)$$

where the user defined parameter α is set to a small value like 0.01 (Table 1).

Table 1 Comparison of the activation functions

Activation functions	Neurons are not saturated	Zero centered	Less computation time
Sigmoid	No	No	No
tanh	No	Yes	No
ReLU	Yes	No	Yes
leaky ReLU	Yes	Yes	Yes

2.4 Hyper-Parameter Tuning

Many hyper parameters may be used in deep learning models, such as learning rate, activation functions, and weight parameter initialization [13].

2.5 Regularization

In order to avoid over fitting problem when the training set size is small, regularization method is used. Here L2 regularization is used. In case of over fitting the training error moves towards zero whereas the test error becomes more [14].

Data set augmentation makes the training error harder to move towards zero. After a certain number of epochs when the test error starts increasing, training of the data set can be stopped. This process is called as early stopping.

2.6 The Steps for Classification Using DNN

The implementation of DNN for multi-class classification consists of the following steps:

1. Initialize weights and biases.
2. Standardize the input features.
3. Apply neural network with back propagation [15]
4. Use the learning algorithms with different strategies like Batch, Mini-batch and Stochastic.
5. Apply the various Activation functions.
6. Use Regularizations and Hyper-parameter tuning.
7. Apply one-hot encoding for multi-class classification.
8. Compare the different combinations of the above methods on the given dataset to find the best one.
9. Check for training and test error and perform the following steps until the both training and test errors are low:

if (training error and test error are high)

- increase model complexity
- Train for more epochs

else

- perform regularization
- data set augmentation
- early stopping

3 Proposed Method

To overcome the local minima problem of Gradient Descent as well as the manual tuning of learning rate simulated annealing concept [16, 17] is used.

Algorithm

Simulated Annealing based Gradient Descent

1. **Initialize** weights (w) and biases (b)
2. **for** $i = \mu_{\max}$ To μ_{\min} **do** :
3. **for** $j = 0$ To No_of_iterations -1 **do**:
4. Compute y
5. Compute $\Delta\omega$ and Δb
6. **if** ($\Delta\omega > 0$) then :
7. $\omega = \omega - \mu \cdot (\Delta\omega)$
8. **else if** ($e^{\frac{\Delta\omega}{\mu}} > \text{rand}(0,1)$)**then** :
9. $\omega = \omega - \mu \cdot (\Delta\omega)$
10. **if** ($\Delta b > 0$) **then** :
11. $b = b - \mu \cdot (\Delta b)$
12. **else if** ($e^{\frac{\Delta b}{\mu}} > \text{rand}(0,1)$)**then** :
13. $b = b - \mu \cdot (\Delta b)$

For applying the concept of simulated annealing, the algorithm has to follow an annealing schedule of temperature. This involves the decision on initialization, update and also the amount by which the update to be done. Here an annealing schedule of the learning rate is used, where μ is the learning rate and it ranges from μ_{\max} to μ_{\min} . μ_{\max} is the maximum learning rate and μ_{\min} is the minimum learning rate. μ_{\max} is initialized to 0.01 and μ_{\min} is initialized to 0.0001. The $\Delta\omega$ and Δb is the derivative of the loss function with respect to w and b. Here the weights and

biases are updated based on the condition of simulated annealing which uses the gradients of weights and biases as parameters for checking the conditions.

The weights are updated if and only if either the gradient of loss function with respect to weights i.e. $\Delta\omega > 0$ or the simulated annealing condition of $e^{\frac{\Delta\omega}{\mu}} > rand(0, 1)$ is met. Similar condition is applicable for biases also. Otherwise the weights and biases are not updated. As in case of simulated annealing the temperature is decreased periodically, here the learning rate is decreased periodically. A new solution is evaluated only if any of the conditions is satisfied. The iterations continue until the stopping condition is met.

Initialization:

Weights and biases are initialized by using either random initialization, He initialization or Xavier initialization. It is seen that Xavier and He initialization gives better result than random initialization.

Weights and biases are optimized by updating them repeatedly for a particular number of epochs and also for different learning rates. The learning rate is updated automatically instead of manual tuning.

4 Experiment

The experiments are performed using Python–Scikit learn [18] and Pytorch [19].

4.1 Data and Evaluation Metric

Three datasets were used for the experiments, one csv dataset and the other two large image datasets. These datasets were Iris dataset [20], CIFAR10 [21] and MNIST [22] datasets. The datasets are explained in the subsections.

4.1.1 Iris Dataset

The standard Iris data set was downloaded from UCI machine learning repository. The data set contains total 150 instances and 4 features. It's a multi class classification problem and the number of classes is 3. Since the dataset is not linearly classifiable, deep DNN is used to classify the data non-linearly.

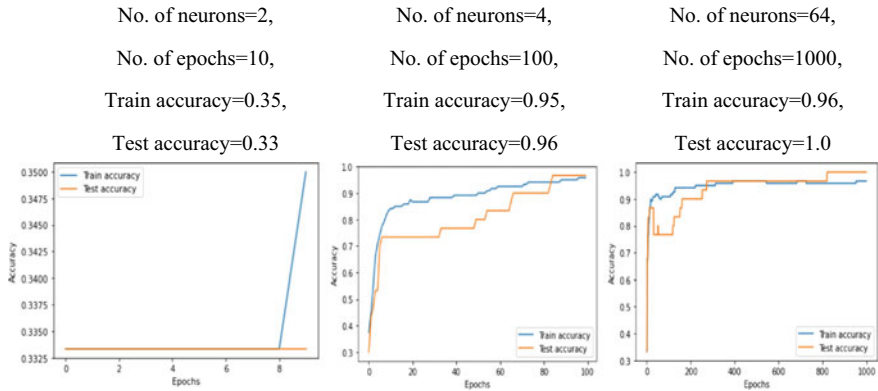


Fig. 1 Comparison of training and test accuracy for different number of neurons and epochs

This dataset is used to show the importance of hyper parameter tuning on the performance of DNN. The hyper parameter tuning involves varying the number of neurons in the network layers, number of epochs, regularization and early stopping. When our training error is high and our test error is also high, the model complexity is increased by adding more number of layers or more number of neurons in the layer and also the number of epochs, which is shown in Fig. 1. It shows that the training and test accuracy improves as the number of epochs and the number of neurons increases which indicates the importance of hyper parameter tuning in the performance of DNN.

On the other hand, if our training error is low and test error is high, i.e. to avoid over-fitting of the data, data set augmentation, Regularization and early stopping is performed which is shown in Fig. 2.

Various initialization methods are used with several activation functions to find the best suitable combinations. Table 2 gives the comparative performance of SAGD and GD for the combinations of different activation functions and initialization methods on Iris dataset. The SAGD method along with different activation functions like tanh, sigmoid, ReLU and Leaky ReLU is tested on Iris data set. The weights and biases are initialized using random initialization, He initialization and Xavier Initialization. Table 2 summarizes the results which shows that all the activation functions’ performance is remarkable. No. of iterations used for training is 1000.

The log loss for SAGD is shown in Fig. 3. It shows the convergence of SAGD with increase in epochs. The log loss comes to a saturation after around 400 epochs.

4.1.2 MNIST Dataset

The MNIST database is a collection of handwritten digits. It contains 60,000 training images and 10,000 test images. Fig. 4 shows the sample image realted to MNIST database.

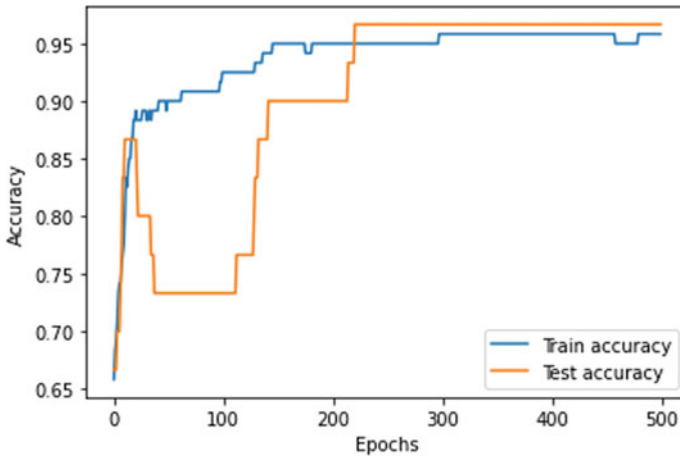


Fig. 2 Calculation of training and test accuracy on Iris dataset after applying regularization, dataset augmentation and early stopping

Table 2 Performance of simulated annealing based gradient descent (SAGD) and stochastic gradient descent (SGD) with different activation functions and Initialization methods

Model	Test accuracy for Iris data set	
	GD	SAGD
DNN + tanh + Random initialization	87.5	92.4
DNN + tanh + He initialization	88.3	97.3
DNN + tanh + Xavier initialization	93.1	100
DNN + sigmoid + Random Initialization	92.5	97.3
DNN + sigmoid + He Initialization	94.2	100
DNN + sigmoid + Xavier Initialization	85.2	86.8
DNN + ReLU + Random Initialization	88.3	96.3
DNN + ReLU + He Initialization	90.2	98.2
DNN + ReLU + Xavier Initialization	92.5	98.6
DNN + LeakyReLU + Random Initialization	92.7	100
DNN + LeakyReLU + He Initialization	93.0	100
DNN + LeakyReLU + Xavier Initialization	93.5	100

The x-train and x-test parts contain greyscale RGB codes. The y-train and y-test part contains the labels from 0 to 9 which represents the class to which the image belongs to. The size of the image is 28×28 pixels.

The VGG16, ResNet 18 and Inception v3 were trained for 20 epochs. The number of iterations in each epoch is 1500. The log loss is given in Fig. 5 for ResNet-18. After each epoch the models were tested on the test set. The best accuracy is shown in Table 3.

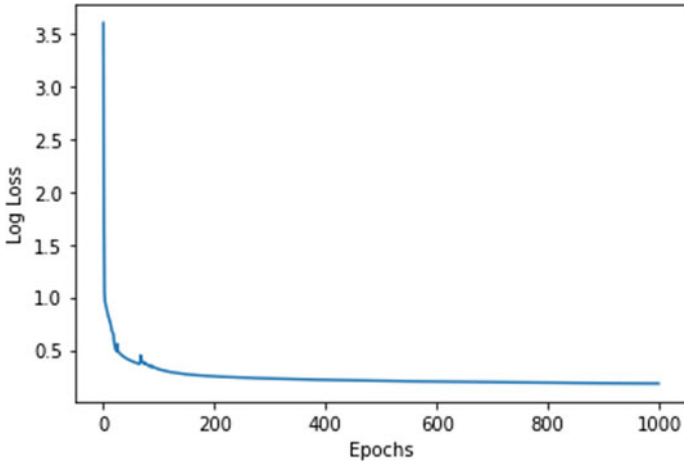


Fig. 3 Log loss for SAGD

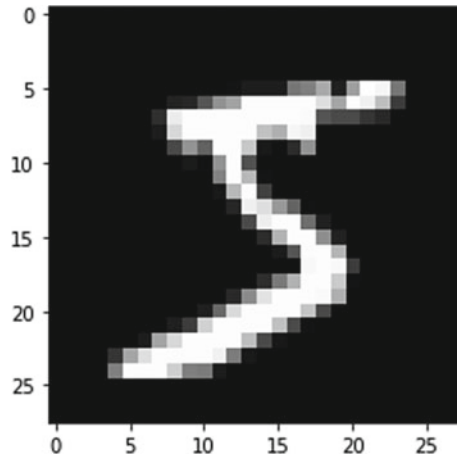


Fig. 4 Sample image from MNIST

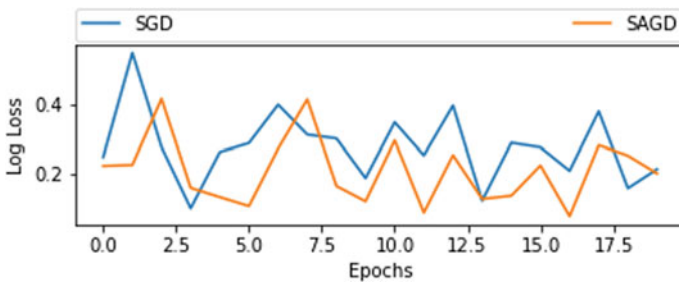


Fig. 5 Comparison of log loss for SGD and SAGD for MNIST dataset

Table 3 SAGD compared with SGD on MNIST and CIFAR-10 dataset. Improved performance is observed for SAGD for both the datasets on all the CNN networks

Model	MNIST		CIFAR10	
	Train accuracy	Test accuracy	Train accuracy	Test accuracy
VGG16 + SAGD	86.17	86.84	75.81	74.48
VGG16 + SGD	82.11	83.08	72.11	71.24
Resnet18 + SAGD	86.88	87.07	78.81	76.51
Resnet18 + SGD	83.06	82.07	78.09	76.02
InceptionV3 + SAGD	90.02	91.26	78.91	80.46
InceptionV3 + SGD	87.15	86.24	76.06	75.32

4.1.3 CIFAR-10 Dataset

CIFAR-10 dataset is a collection of 60,000 small RGB images. Out of which 50,000 are used as training images and 10,000 are used as test images. Each image is 32×32 pixels' color RGB image. The dataset contains images of frog, horse, cat, airplane, automobile, truck, deer, dog, bird, ship which falls into ten classes (Fig. 6).

The large Convolutional Neural Networks (CNN) like VGG 16, ResNet18 and InceptionV3 are used with a mini-batch size of 32. The models are trained for 20 epochs for both SGD and SAGD. Both the models are evaluated after each epoch and the comparative results are given in Table 3. The learning rate is 0.05 for both SGD and SAGD and the activation function is ReLu. The log loss values on ResNet-18 after each epoch is also shown in the Fig. 7. It shows that SAGD outperforms SGD in test accuracy.



Fig. 6 Sample images with labels generated from CIFAR-10 datasets

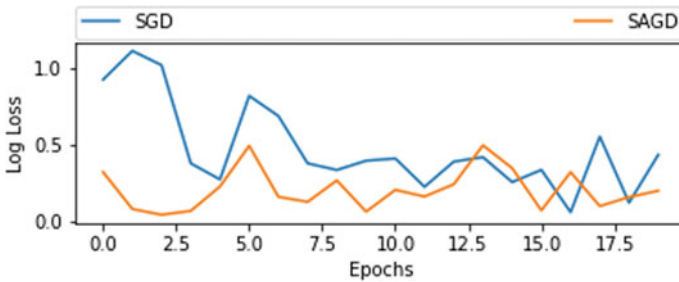


Fig. 7 Comparison of log loss for SGD and SAGD for CIFAR-10 dataset

5 Conclusion

In this article a better learning algorithm is introduced by using simulated annealing concept with gradient descent in optimizing deep neural networks for Multi class classification. The method showed promising results on the two datasets CIFAR-10 and MNIST. The comparison of the different combinations of learning algorithms and

activation functions is shown. Later the result is fine-tuned by applying regularizations which helps to avoid over-fitting. Our proposed method (SAGD) has performed better than the standard gradient descent methods with high accuracy for different initialization and activation functions.

References

1. El Afia A et al (2018) A self-tuned simulated annealing algorithm using hidden markov model. *Int J Electr Comput Eng* 8:291–298
2. Shrestha A, Mahmood A (2019) Review of deep learning algorithms and architectures. *IEEE Access* 7:53040–53065
3. Svozil D, Kvasnicka V, Pospichal J (1997) Introduction to multi-layer feed-forward neural networks. *Chemom Intell Lab Syst* 39(1):43–62
4. Er MJ, Venkatesan R, Wang N (2016) An online universal classifier for binary, multi-class and multi-label classification. In: *IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, 2016, pp 003701–003706. <https://doi.org/10.1109/SMC.2016.7844809>
5. Karlik B et al (2011) Performance analysis of various activation functions in generalized MLP architectures of neural networks. *Int J Artif Intell Expert Syst (IJAE)* 1(4)
6. Quoc VL et al (2011) On optimization methods for deep learning. In: *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA
7. Mohamed L et al (2018) *Int J Electr Comput Eng (IJECE)* 8(1):291–298
8. Kumar SK (2017) On weight initialization in deep neural networks. pdfs.semanticscholar.org
9. Goodfellow I et al: *Deep Learning*, book
10. Wibowo A et al (2019) Optimization of neural network for cancer microRNA biomarkers classification. *J Phys Conf Ser* 1217:012124
11. Ruder S (2017) An overview of gradient descent optimization algorithms. [arXiv:1609.04747](https://arxiv.org/abs/1609.04747)
12. Simone S et al (2017) Kafnets: kernel-based non-parametric activation functions for neural networks. [arXiv 1707.04035](https://arxiv.org/abs/1707.04035)
13. Yaseen MU et al (2018) Deep learning hyper-parameter optimization for video analytics in clouds. *IEEE Trans Syst Man Cybern Syst*. <https://doi.org/10.1109/TSMC.2018.284034>
14. Khan SH et al (2019) Regularization of deep neural networks with spectral dropout. *Neural Netw* 110:82–90. Elsevier
15. Hassan M, Hamada M (2017) Performance comparison of feed-forward neural networks trained with different learning algorithms for recommender systems. *Computation* 5:40
16. Bertsimas D et al (1993) Simulated annealing. *Stat Sci* VL–8:0883–4237. Institute of Mathematical Statistics
17. <https://www.sciencedirect.com/topics/engineering/simulated-annealing-algorithm>
18. Pedregosa F et al (2011) Scikit-learn: machine learning in python. *JMLR* 12:2825–2830
19. <https://pytorch.org/docs/stable/optimize.html>
20. <https://archive.ics.uci.edu/ml/datasets/iris>
21. Krizhevsky A, Hinton G (2009) Learning multiple layers of features from tiny images. University of Tront, Master's thesis
22. <http://yann.lecun.com/exdb/mnist/>
23. Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. *Aistats* 9:249–256
24. He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp 1026–1034

25. Yang L, Cai D (2021) AdaDB: an adaptive gradient method with data-dependent bound. *Neurocomputing* 419(1):183–189
26. Karabayir I, Akbilgic O, Tas N (2020) A novel learning algorithm to optimize deep neural networks: evolved gradient direction optimizer (EVGO). *IEEE Trans Neural Netw Learn Syst* 32(2):685–694

Lung Cancer Classification System for CT Images using Deep Convolutional Neural Network



A. Jayachandran  and N. Anisha

Abstract Lung cancer is one of the fatal cancer types, and its identification at early stages might increase the patients' survival rate up to 60–70%. Predicting patients' lung cancer survivability has become a trendy research topic by scholars from medicine and computer science domains. In this article, a novel approach to the deep convolutional neural network (DCNN) is proposed for the precise and automatic classification of lung cancer. Specifically, the auto weight dilated convolutional unit utilized multi-scale convolutional feature maps to acquire lung cancer features at different scales and employed a learnable set of parameters to fuse convolutional feature maps in encoding layers. The AD unit is an effective architecture for feature extraction in the encoding stage. We used the advantages of the U-Net network for deep and shallow features, combined with AD units to multimodal image classification. In this model, the four channel model inputs correspond to the CT images of four modes, respectively. The main body of the network is composed of auto-weight dilated (AD) unit, Residual (Res) unit, linear upsampling, and the first and last convolution units.. The network that applied Block-R3 had higher segmentation performance than the networks of Block-R1 and Block-R2 experimental results indicated that DCN had outperformed all the six classical machine learning algorithms in predicting the survival period of lung cancer patients with an accuracy of 88.58%. The results are believed to support healthcare professionals to manage costs and provide treatment at the appropriate time.

Keywords Lung cancer · CT · Segmentation · Deep CNN · Classification

A. Jayachandran (✉)
Department of CSE, Presidency University, Bangalore, India
e-mail: ajaya1675@gmail.com

N. Anisha
Department of CSE, PSN College of Engineering and Technology, Tirunelveli, Tamil Nadu, India

1 Introduction

Lung cancer patients' relative five-year survival rate is around 20%, making it the deadliest cancer type. A man's probability of lung cancer is 1/15, while it is 1/17 for a woman. This ratio is increased for smokers [1]. The major issue with lung cancer is the difficulty of its treatment because it is identified at later stages [2]. Identifying lung cancer at early stages might boost the patients' survival rate up to 60–70% [3]. Predicting the estimated survival period after identification improves the prognostic accuracy, which leads to physicians' and patients' families' better decision-making [4]. Therefore, predicting patients' lung cancer survivability has become a trendy research topic by scholars from medicine and computer science domains. The advent of artificial intelligence (AI) techniques, specifically machine learning algorithms, has improved the diagnosis and treatment of cancer patients.

Machine learning develops models capable of learning and providing informed decisions based on large amounts of historical data. The training data used by machine learning algorithms also enable professionals to get an insight into early cancer diagnosis, variation of treatments, and drug discovery [5]. This automatic process of identification and exploration has proved its performance and efficiency in classifying patients' lung cancer images using various machine learning algorithms [6]. The existing literature has extensively relied on computer tomography (CT) or X-ray images for identifying lung cancer patients. For instance, Vas and Dessai [7] used an artificial neural network (ANN) to categorize the cancer stages based on CT scan images, while Gang et al. [8] employed the parallel immune algorithm for cancer diagnosis based on X-ray images. However, these techniques were criticized for several reasons. First, the errors in the images taken through the CT and X-ray techniques lead to false-negative reports, causing delays in cancer treatment. Second, these techniques can sometimes be difficult for screening patients because of the low number of available devices, high costs, and radiation doses. Sathesh A 2020 proposed a lung nodule segmentation algorithm that uses adaptive weights as a feature for the recurrent neural network. The algorithm initially detects the lung parenchyma from which the background region is minimized. However, the boundaries of the obtained nodule candidate region are not accurate. The evaluation was done with the LIDC datasets using the metrics such as Hausdorff distance, probability rand index (PRI), accuracy, recall, and precision. The scheme provides accuracy, recall, and precision of 94.08%, 89.3%, and 94.1%, respectively [19, 20].

To overcome the limitations of CT and X-ray images, machine learning algorithms have been used to diagnose lung cancer patients by relying on patients' clinical features. However, there is a lack of research on predicting lung cancer patients' survival or death period, specifically with the use of deep learning techniques. Unlike classical machine learning, deep learning techniques learn by building a more abstract representation of data as the network expands deeper [10, 11]. This, in turn, helps maximizing the prediction accuracy of deep learning models compared to their counterparts of classical machine learning. Therefore, this research aims to predict lung cancer patients' survival or death period through a comparison between classical

machine learning and deep learning techniques using patients' demographic and clinical features. The early prediction of the survival period of lung cancer helps patients and healthcare professionals to manage costs better and provide treatment at the appropriate time [12–15].

2 Proposed Methodology

In recent years, deep learning methods gained significant interest in the segmentation of lung cancer. Deep learning method has gained a significant advantage compared to other approaches. Deep learning puts forward a way to let computers learn the features automatically based on data-driven to reduce the complexity of artificial design features. The deep learning model with essentially enlarged depth advances segmentation performance. Figure 1 depicts the summary of the proposed methodology.

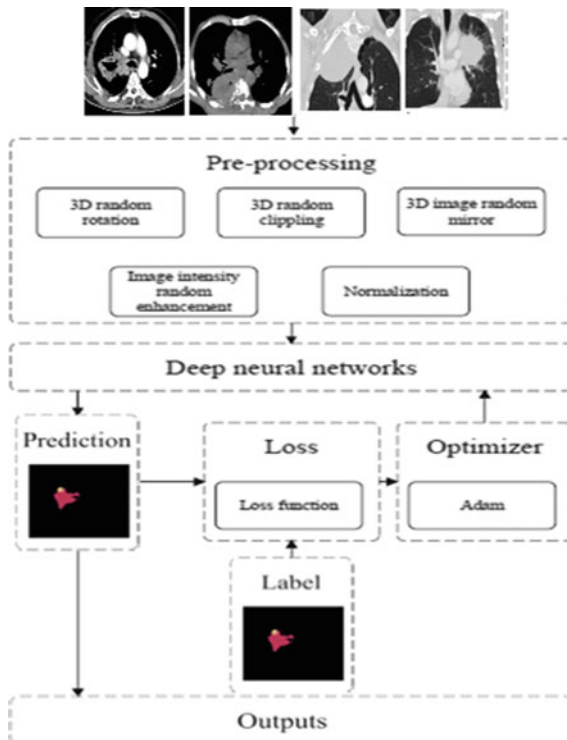


Fig. 1 Overall block diagram of proposed lung cancer segmentation model

2.1 Pre-processing

We used the randomization strategy as image preprocessing, which could ensure that the deep learning model still maintains strong generalization performance after a large number of repeated training. Multimodal brain images of the same patient use the same processing in one epoch training and different random measures in different epochs. It helps to learn the image features of different modes in the same brain while obtaining generalization. The figure shows the image preprocessing methods: 3D random clipping, 3D random rotation, 3D image intensity random enhancement, 3D image random mirror inversion, and normalization.

Image normalization is a widely used technique in computer vision, pattern recognition and other fields. The z-score normalization was applied in this work. It is defined as per Eq. (1):

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

where σ is the standard deviation, and μ is the mean value. Then, the 3D random clipping method randomly cuts the MRI image (240, 240, 155) into a matrix (144, 144, 128). The 3D random rotation method rotates the reduced image by the angle $U(-10, +10)$. The random intensity enhancement method of 3D image sets the image pixel value is defined as per Eq. (2):

$$x_{new} = x_{old} * U(0.9, 1.1) + U(-0.1, 0.1) \quad (2)$$

where U is the uniform distribution. The random mirror processing symmetrizes the image according to deep, height and width directions. We applied these image enhancement routines to extend the training data set to improve the performance and generalization ability of the deep neural network.

2.2 Deep Learning Model

Deep learning which is a subset of Artificial intelligence is gaining momentum each day by making different tasks much easier and more efficient. CNN which is a type of deep learning mechanism is an inevitable part of image vision problems. In recent years, deep learning methods gained significant interest in the segmentation of lung cancer. The performance of deep learning segmentation methods usually depends on the size of training data. However, it is always difficult to acquire a large number of image data with pixel-level annotation in clinical practice. In order to address the challenge of scarce annotation, studies [12–17] have applied few-shot learning on medical image analysis, where the labeled data is a small portion of the whole dataset. For these methods, it is often difficult to effectively utilize the unlabeled data. In recent years, unsupervised learning has been performed on medical

image, using unlabeled data for model optimization. The U-shaped model is an efficient and straightforward segmentation network in 3D medical images especially in lung cancer, learning features from deep and shallow neural units. The UNet model consists of four encoder layers and four decoder layers. The proposed model is shown in Fig. 2. In this model, the four channel model inputs correspond to the MRI images of four modes, respectively. The main body of the network is composed of auto-weight dilated (AD) unit, Residual (Res) unit, linear upsampling, and the first and last convolution units. In the downsampling stage (feature coding extraction), we use 8 AD units to obtain multi-scale feature maps. In the upsampling stage (feature decoding), we use the AD unit, Res unit and a linear upsampling layer to form a primary decoding layer. Finally, a convolution unit outputs the results of the network model. Moreover, each convolution unit, AD unit and Res unit contains batch normalization and ReLU functions. We used extended convolution to extract fine-grained and multi-scale glioma features, and employed residual structure to obtain long-dependent glioma features.

As for the Res Unit layer, we used two convolution units to reduce and then enlarge the number of convolution kernels so as to realize feature learning and feature map reorganization. From an experimental point of view, this is an efficient coding method. Then, we used two group convolution units with stripe 1 and group 16, and the kernel size is $3 \times 3 \times 3$. Finally, we used a convolution residual element to obtain the characteristic graph of long dependence.

As for the AD Unit layer, we used two convolution units firstly (like the Res unit). Then, we used three extended convolution units (the divided parameters are 1 and 2, respectively) and used two learnable parameters to adjust and fused the characteristics of the two group extended convolution units. Finally, a group convolution unit was used to output the result of the AD unit. We also set up residual calculations in the AD

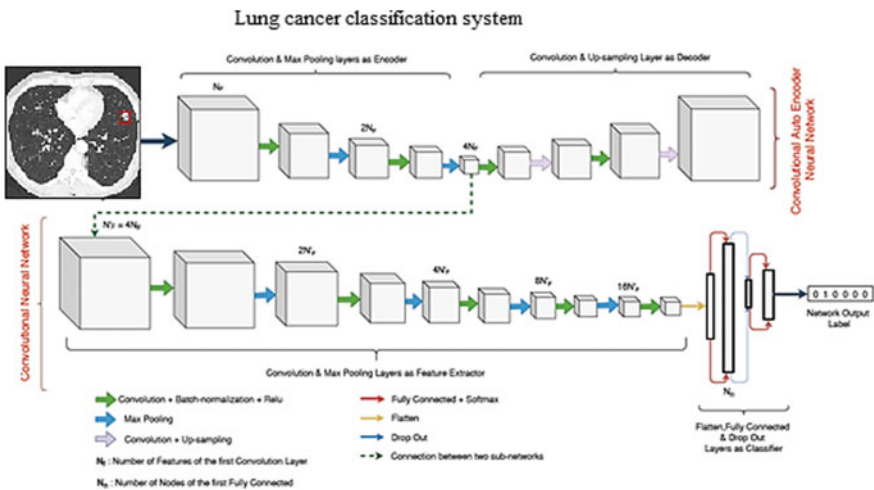


Fig. 2 An illustration of the proposed architecture for lung cancer classification system

unit. The dilated convolution could expand the receptive field of the convolutional kernel without sacrificing computational resources, while normal convolution could provide a more accurate feature map. The fusion of the two types of convolutions could strengthen the ability of the network to extract features.

In the encoder stage, each residual block is a dual-pathway structure. In this stage, we set channel depth to 32, 64, 128, and 256. The residual block is the critical structure of down-sampling. In the decoder stage, we connect a convolutional unit and a de-convolutional unit for upsampling. The stripe of the de-convolutional unit is $2 \times 2 \times 2$, and the kernel size is $3 \times 3 \times 3$. Batch Normalization and RReLU activation functions are connected behind the convolutional unit or the de-convolutional unit for all convolutional units and all de-convolutional units in this stage. Similarly, we set the channel depth in the decoder stage to $32 \times 64 \times 128 \times 256$. A combined deep neural network with the residual blocks enables the network to obtain more significant gradients in deep layers.

So, the phenomenon of gradient disappearance is relatively rare and gets more practical features of gliomas. The formula of the gradient propagation in the convolutional layer can be defined in Eq. (3),

$$\delta_l = \sigma'(O_l) \cdot (w_{l+1})^T \delta_{l+1} O_l \quad (3)$$

where σ' means the first derivative of the loss function, w describes the weight, O indicates the output matrix vector, and l is the layer l . Then, the gradient in Block-R1, Block-R2 and Block-R3 can be defined in Eq. (4),

$$\begin{aligned} O_{r1_{l+1}} &= f(\delta_2(f(\delta_1(O_{r1_l})) + O_{r1_l})) \\ O_{r2_{l+1}} &= f(\delta_2(f(\delta_1(O_{r2_l})) + O_{r2_l})) \\ O_{r3_{l+1}} &= \delta_2(f(\delta_1(f(O_{r3_l})) + O_{r3_l})) \end{aligned} \quad (4)$$

where f means the activation function, δ_1 and δ_2 represent the first and second convolution calculations, respectively. It is worth noting that the difference between Eq. (3) and Eq. (4) lies in the order of normalization, which is not reflected in the equation.

Multiplication is widely used in the calculation of series convolution, such as $\delta_2(f(\delta_1))$. The cumulative multiplication between $(-1, 1)$ makes it possible for the gradient to appear the result of approximate 0, so that the classical gradient disappears. The residual-connection weakens this problem through weight addition, and enhances the stability of the network. Obviously, it is a very effective way to use residual blocks to build architecture in very deep neural layers, especially in calculating the depth feature map.

We defined the convolution block (BN, RL, Conv) in AD unit as per Eq. (5),

$$\ell_{(c_i, k)} = w_{c_i}^T f(I_i) + b_{c_i} \quad (5)$$

where c_i is the convolution layer i , and k describes the kernel size, f is the activation function. The w and b represent the convolution weight and bias, I_i is the input data. Then, the AD block can be define as per Eq. (6)

$$\ell_{AD} = \ell_{(c_1,1)}\ell_{(c_2,1)}(a\ell_{(d_0,3)} + b\ell_{(c_3,3)})\ell_{(c_4,3)} + \ell_{(c_5,1)} \quad (6)$$

where ℓ_{d_0} is the dilated convolution. In the gradient back-propagation process, a and b can automatically adjust the weight ratio of the convolution integral branch in the main path. In addition, the channel parameters of the AD-Net were set to 32, 64, 128, 256, and the skip connection adopted the 3D matrix concatenate method. The residual structure was a necessary element. The residual calculation ensured the stability of gradient in deep feature calculation.

3 Experimental Results and Discussion

This research involved 10,001 subjects from the SEER database collected between 2000 and 2019 [16] (SEER, 2021). The age of the patients was ranged between 19 to 94 years old (Mean = 65.85, SD = 9.72). Further, 51.8% of the patients were males (n = 5181) and the remaining were females (n = 4820). There were 1214 (12.1%) alive cases and 8787 (87.9%) death cases. The age of alive cases ranged between 23 and 90 years old (Mean = 64.40, SD = 8.71), whereas the age of death cases ranged between 19 and 94 years old (Mean = 66.05, SD = 9.83). In terms of race, there were 86.3% white, 10.8% black, and 5.1% other nations (i.e., American Indian/AK Native, Asian/Pacific Islander). The alive cases included 571 (47%) males and 643 (53%) female patients. In addition, the death cases included 4610 (52.5%) male patients and 4177 (47.5%) females. An independent sample t-test was performed to test the difference between male and female patients. The outcomes pointed out that there is a significant difference between male and female subjects ($t(9999) = 3.55$, $p < 0.001$). This indicates that the probability of death was higher in males than females. The samples lung cancer CT images collected from SEER database are given in Fig. 3.

The following evaluation metrics are used for validating the classifier performance. It is given in Eq. (7)

$$\begin{aligned} Accuracy &= \frac{TP + TN}{Total\ no\ of\ images} \\ Precision &= \frac{TP}{FP + TP} \\ Recall &= \frac{TP}{FN + TP} \\ F1score &= 2 * \frac{Precision * Recall}{Precision + Recall} \end{aligned} \quad (7)$$

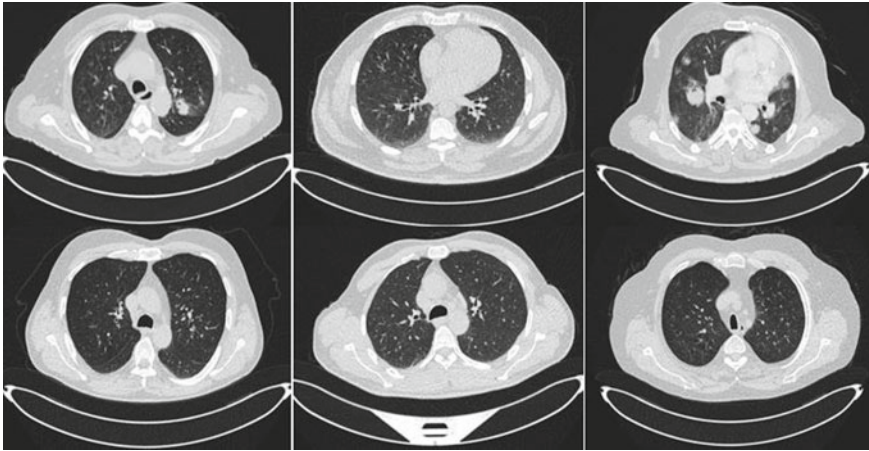


Fig. 3 Sample lung cancer CT images in SEER database

where accuracy explains how well the architecture can classify the images. It is given as the ratio of total correct prediction made to the total number of predictions made. Precision is given by the ratio of total number of correctly classified positive cases to the sum of total positive cases predicted. High value for precision is required in order to minimize the number of false positive classifications. Recall is defined as the ratio of total positive cases correctly classified to all correctly classified observations and F1 score gives the harmonic mean of precision and recall. It can be also defined as the weighted average of precision and recall [15–18]. The performance comparison results shown in Tables 1 and 2 suggest that all applied classifiers have a good performance. Experimental results of different models are visualized in Fig. 4 and the classification rate of various models using different statistics method is given in Fig. 5.

Table 1 Performance measures of different Lung cancer classification models

	Accuracy (%)	TP rate	FP rate	Precision	Recall	F-Measure
BayesNet	88.22	0.882	0.786	0.852	0.882	0.845
Logistic	88.50	0.885	0.824	0.884	0.885	0.839
LWL	88.08	0.881	0.768	0.848	0.881	0.847
ASC	88.51	0.885	0.817	0.876	0.885	0.840
OneR	88.50	0.885	0.807	0.869	0.885	0.842
J48	88.21	0.882	0.818	0.854	0.882	0.838
DCNN	0.8858	0.886	0.840	0.852	0.886	0.841

Table 2 Performance comparison of the classifiers

Classifier	Kappa statistic	MAE	RMSE	MCC	ROC area	PRC area
BayesNet	0.1486	0.1952	0.3159	0.213	0.679	0.856
Logistic	0.1009	0.1981	0.3144	0.214	0.660	0.852
LWL	0.1685	0.2032	0.3181	0.220	0.665	0.847
ASC	0.1128	0.2	0.3168	0.217	0.556	0.808
OneR	0.1268	0.115	0.3391	0.221	0.539	0.800
J48	0.1043	0.201	0.3202	0.182	0.551	0.805
DCNN	0.076	0.1902	0.3144	0.145	0.629	0.843

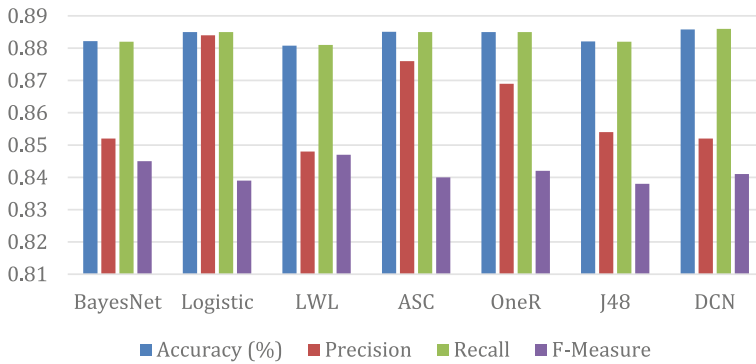


Fig. 4 Experimental results of different classification model

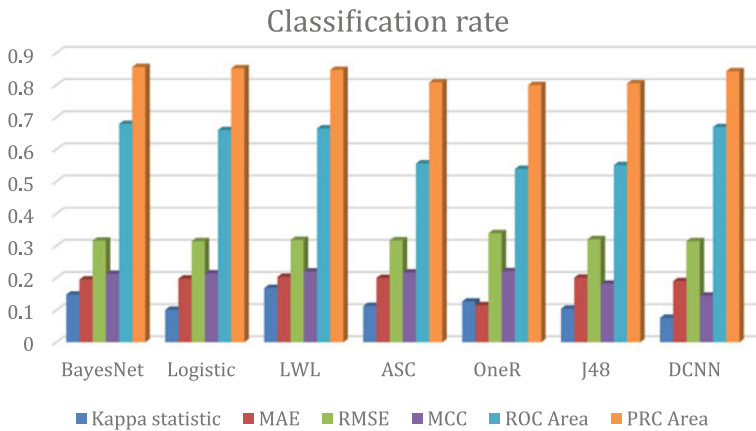


Fig. 5 Classification rate of different model using various statistics methods

4 Conclusion

In this paper, a new multi-scale approach to segment the lung cancer in CT image was described and evaluated in several publicly available databases. This paper also presents an assessment of the most appropriate scales for the lung cancer segmentation, complementing previous work that defines these scales empirically. Furthermore, it was also demonstrated that a multi-scale analysis can improve the lung cancer segmentation. Although recent research has been focusing on deep learning methods, rule-based methods can also be important for the definition of features, that can significantly improve the outcome of these methods. The achieved results show that the proposed approach is very competitive when compared with the current state of the art methods, particularly in high-resolution images. Our method still needs further improvement in the enhancing tumor region segmentation. It were practical tools in 3D lung cancer segmentation.

References

1. Adem K, Kiliçarslan S (2021) COVID-19 diagnosis prediction in emergency care patients using convolutional neural network. *Afyon Kocatepe Üniversitesi Fen Ve Mühendislik Bilimleri Dergisi* 300–309
2. Boddu RSK, Karmakar P, Bhaumik A, Nassa VK, Bhattacharya S (2021) Analyzing the impact of machine learning and artificial intelligence and its effect on management of lung cancer detection in covid-19 pandemic. *Mater Today Proc.* <https://doi.org/10.1016/J.MATPR.2021.11.549>
3. Jayachandran A, Dhanasekaran R (2013) Brain tumor detection using fuzzy support vector machine classification based on a texton co-occurrence matrix. *J Imag Sci Technol* 57(1):10507-1–10507-7(7)
4. Brinati D, Campagner A, Ferrari D, Locatelli M, Banfi G, Cabitza F (2020) Detection of COVID-19 infection from routine blood exams with machine learning: a feasibility study. *J Med Syst* 44(8):135. <https://doi.org/10.1007/s10916-020-01597-4>
5. Cai Z, Xu D, Zhang Q, Zhang J, Ngai SM, Shao J (2015) Classification of lung cancer using ensemble-based feature selection and machine learning methods. *Mol BioSyst* 11(3):791–800. <https://doi.org/10.1039/C4MB00659C>
6. Chen YC, Ke WC, Chiu HW (2014) Risk classification of cancer survival using ANN with gene expression data from multiple laboratories. *Comput Biol Med* 48(1):1–7. <https://doi.org/10.1016/J.COMPBIOMED.2014.02.006>
7. Jayachandran A, Kharmega Sundararaj G (2016) Abnormality segmentation and classification of multi model brain tumor in MR images using Fuzzy based hybrid kernel SVM. *Int J Fuzzy Syst* 17(3):434–443
8. Dass MV, Rasheed MA, Ali MM (2014) Classification of lung cancer subtypes by data mining technique. In: *Proceedings of the 2014 international conference on control, instrumentation, energy and communication (CIEC)*, pp 558–562. <https://doi.org/10.1109/CIEC.2014.6959151>
9. Doppalapudi S, Qiu RG, Badr Y (2021) Lung cancer survival period prediction and understanding: deep learning approaches. *Int J Med Inform* 148:104371. <https://doi.org/10.1016/J.IJMEDINF.2020.104371>
10. Dutta AK (2022) Detecting lung cancer using machine learning techniques. *Intell Autom Soft Comput* 31(2):1007–1023

11. Friedman N, Geiger D, Goldszmidt M (1997) Bayesian network classifiers. *Mach Learn*. <https://doi.org/10.1002/9780470400531.eorms0099>
12. Peng G, Yang X, Liu L (2011) Parallel immune algorithm for lung cancer detection in X-ray images based on object shared space. In: 2011 12th international conference on parallel and distributed computing, applications and technologies, pp 197–200. <https://doi.org/10.1109/PDCAT.2011.64>
13. Gao Y, Lyu Q, Luo P, Li M, Zhou R, Zhang J, Lyu Q (2021) Applications of machine learning to predict cisplatin resistance in lung cancer. *Int J Gener Med* 14. <https://doi.org/10.2147/IJGM.S329644>
14. Hsu CH, Manogaran G, Panchatcharam P, Vivekanandan S (2018) A new approach for prediction of lung carcinoma using back propagation neural network with decision tree classifiers. In: 2018 IEEE 8th international symposium on cloud and service computing (SC2), pp 111–115. <https://doi.org/10.1109/SC2.2018.00023>
15. Ibrahim DM, Elshennawy NM, Sarhan AM (2021) Deep-chest: Multi-classification deep learning model for diagnosing COVID-19, pneumonia, and lung cancer chest diseases. *Comput Biol Med* 132:104348. <https://doi.org/10.1016/J.COMPBIOMED.2021.104348>
16. Prabhu AJ, Jayachandran A (2018) Mixture model segmentation system for parasagittal meningioma brain tumor classification based on hybrid feature vector. *J Med Syst* 42(12)
17. SEER (2021) SEER Incidence Data, 1975 - 2018. National Cancer Institute (NCI). <https://seer.cancer.gov/data/>
18. Namboodiri S, Jayachandran A (2020) Multi-class skin lesions classification system using probability map based region growing and DCNN. *Int J Comput Intell Syst* 13(1):77–84
19. Sathish A (2020) Adaptive shape based interactive approach to segmentation for nodule in lung CT scans. *J Soft Comput Paradigm* 2(4):216–225
20. Sunghheetha A, Rajesh SR (2020) Comparative study: statistical approach and deep learning method for automatic segmentation methods for lung CT image segmentation. *J Innov Image Process* 2:187–193

Performance Analysis of Machine Learning Algorithms in Heart Diseases Prediction



K. Nanthini, M. Pyingkodi, D. Sivabalaselvamani, Shweta Kumari, and Tarun Kumar

Abstract It is critical to discover aberrant heart conditions early in order to identify heart abnormalities and prevent sudden cardiac death. Cardiovascular infections (CVDs) are the leading cause of death worldwide, killing 17.9 million people year and accounting for 31% of all fatalities. Four out of every five CVD deaths are caused by coronary illnesses and strokes, with 33% of these deaths happening unexpectedly in adults under the age of 70. CVDs are known to cause cardiovascular breakdown, and this dataset comprises 12 elements that can be used to predict a probable coronary sickness. People who have cardiovascular disease or are at high cardiovascular risk (due to the presence of at least one risk factor such as hypertension, diabetes, hyperlipidemia, or a pre-existing illness) require early identification and treatment, which an AI model may give. According to WHO figures, heart disease is the leading cause of non-communicable illness death in India, accounting for 24% of all fatalities. Heart illnesses account for one- third of all deaths worldwide. Heart disease accounts for half of all deaths in the United States and other industrialized countries. Every year, almost 17 million people die from cardiovascular diseases (CVD), and the condition is particularly widespread in Asia. In this post, we will attempt to construct a machine learning model and use various machine learning algorithms. We are comparing their performance and then will suggest which algorithm is going to be model for this specific task. We are going to doing lots of task feature selection, model evaluation and will building another next set of model based on selected model and then we will decide which machine learning model is more generalize. **Purpose:** The study of a dataset (ECG dataset) and its ability to predict whether or not a person has cardiac disease. A person with a heart condition is represented by 1 and a person without a heart condition is represented by 0. The system will use machine learning to anticipate cardiac disease and get the best result. **Conclusion:** Nine-classification methods are used to detect the heart disease and evaluate the performance of each model for the given dataset. Logistic Regression has high accuracy among all.

K. Nanthini · M. Pyingkodi · D. Sivabalaselvamani (✉) · S. Kumari · T. Kumar
Department of Computer Applications, Kongu Engineering College, Perundurai, Erode,
TamilNadu, India
e-mail: sivabalaselvamani@gmail.com

Keywords Random Forest · SVM · KNN · Adaboost · XGboost · Catboost · LR · LGBM · Naïve Bayes · Heart diseases prediction · Machine learning

1 Introduction

The work proposed during this paper focus essentially around different information mining rehearses that are utilized in coronary illness expectation. Human heart is that the chief piece of the body. Essentially, it manages blood stream during our body. Any anomaly to heart can cause trouble in several pieces of body. Any reasonably aggravation to ordinary working of the guts is delegated a Heart sickness. In today's modern times, coronary illness is one in all the essential explanations behind event of most passing. Coronary illness might happen due to unfortunate way of life, smoking, liquor and high admission of fat which could cause hypertension. As indicated by the planet Health Organization in way over 10 million kick the bucket thanks to Heart illnesses each and each year everywhere the earth. A solid way of life and earliest location are just ways of stopping the guts related in factions. Coronary illness is that the normal reasons of individuals' passing in India and in several countries. Inconstancy a critical attributes is concluding the guts condition. Electrocardiogram (ECG) signs and pulse reflects the cardiovascular soundness of human heart. Pulse changeability is employed to evaluate the distinctions within the heart signs and every one the more especially contrasts per unit season of the amount of pulses. ECG is one in every of the strategies to differentiate the center throbs. ECG is that the electrical development of heart it produces electrical signs which are called as PQRSTU waves. The most indispensable wave is QRS complex. Electrocardiogram may be a illustration of the extent and heading of the electrical movement that's created by depolarization and depolarization of the ventricles and atria.

Types of Heart Failure

Left-sided cardiovascular breakdown

- Most normal type of cardiovascular breakdown.
- Liquids may develop in the lungs, causing windedness - pneumonic edema. Right-sided cardiovascular breakdown
- Frequently happens with left-sided cardiovascular breakdown.
- Liquid develops in the mid-region, legs and feet causing expanding fringe edema.

Systolic cardiovascular breakdown

- The left ventricular constriction is strange low left ventricular launch part. Diastolic cardiovascular breakdown
- The left ventricle neglects to unwind or fill completely, showing unusual diastolic capacity

The Risk Factor for Developing Heart Diseases

According to the Centers for Disease Control and Prevention, approximately half of all men and women in the United States have three or more heart disease risk factors. Let's take a look at few these risks, so that you can modify and control them if possible.

1. **Age:** Growing older increases your risk of damage and narrowed artery and weakened your heart muscles.
2. **Gender:** Men are more likely than women to develop heart disease, and the risk of heart disease increases after menopause.
3. A family history of heart diseases increases your risk of coronary artery.
4. **Smoking:-**
 - Coronary attack is over two times as high as somebody who doesn't smoke
 - Coronary illness is fundamentally expanded on the off chance that you are a lady utilizing the oral preventative pill
 - Stroke is over two times as high as somebody who doesn't smoke Fringe blood vessel sickness, which can prompt gangrene, is expanded by in excess of multiple times.

Types of Heart Diseases

Coronary Artery Diseases: It is the most well-known cardiovascular problem. Your coronary arteries, which provide blood to your heart, may become blocked if you have CAD. This can cause a halt in the flow of blood to your heart muscle, preventing it from receiving the oxygen it requires. The disease is usually caused by atherosclerosis, also known as artery solidification.

Heart Arrhythmias: Cardiac arrhythmia defined as when heart deviated from normal rhythm, for the normal cardiac rhythm the heart beat rate should be between 60 and 100.

There are two types of arrhythmias tachycardia and bradycardia.

Tachycardia: If the heart rate becomes more than 100 then it is called as tachy cardiac.

Bradycardia: If the heart rate becomes less than 60 then it is called as Brady cardiac.

2 Literature Review

Jie Zhang et al. [1] they offer a novel technique for predicting cardiac illnesses from ECG signals using cardiology, signal processing technologies, and a deep learning model based on the MIT-BIH dataset. The author of this paper employed Deep Learning, Neural Networks, and Wavelet Transform to predict an 87% accuracy

rate in diagnosing heart illness. Saravana Kumar [2] characteristics extracted from the QTDB dataset the effectiveness of the various categories was determined, and the Radial Basis function classifier outperformed the others. The created method should be applied to improve health care. Concerned individuals the author of this paper employed a neural network and a radial basis function classifier to predict heart disease detection. T.V.N. et al. [3] for all instances, we used the MIT-BIH dataset, as well as features collected from the frequency, time, and time-frequency domains. They found the accuracy by using an improved neural network to forecast heart attacks. Sharmila Rengasamy et al. [4] from the UCI repository dataset they proposed an approach for prediction of heart diseases using SVM from Spark MLlib has been proposed in order to apply of in-memory concept of Spark, rate of 83%.

M. Durairaj et al. [5] suggested a methodology for Intelligent Prediction Methods and Techniques Employing Illness Diagnosis in Medical Database using KNN Classifier processing and cleansing of disease data from the CSEDB dataset, which aids in accurate disease diagnosis and device therapy methods. Data from the MIT-BIH dataset is used by Jiaming Chen et al. [6] to forecast & Smart Heart Monitoring: Predictive Analysis of ECG Signals for Early Detection of Heart Problems They employed spatial transformation and a customised classifier to predict cardiac issues with a sensitivity rating of 90%.

The CSEDB dataset was analysed by K. Butchi Raju et al. [7], who utilised the information to predict Smart Heart Disease Prediction System with IoT and Fog Computing Sectors Enabled by Cascaded Deep Learning Model with a 95% accuracy rate. The data is processed to the prediction of heart disease system is proposed using discrete wavelet transform and support vector machine classification by Sana S. Zadawale et al. [8] from the MIT-BIH database. M.H. Vafaie et al. [9] the purposed method uses a genetic-fuzzy classifier to forecast cardiac illnesses based on ECG signal at the rate of 98.67%. Tim Smole et al. [10] The suggested risk-stratification model, which is based on real-time clinical data, shows good accuracy in predicting events in patients with hypertrophic cardiomyopathy using random forest, svm, and boosted tree neural networks, with the greatest accuracy for boosted tree being 75%.

In this research from MIT-BIH, Ahmed I. Taloba et al. [11] employed discrete wavelet transform (DWT) to recover features such as the R peak and RR interval, and multilayer perceptron (MLP) in the techniques of classification using SVM and MLP rate of SVM and MLP. R. Valarmathi et al. [12] they employed hyper parameter optimization (HPO) tuning and machine learning algorithms such as random forest and XG boost to predict heart disorders from the Cleveland heart diseases dataset, with success rates of 83.96 and 82.7%, respectively. Perna Sharma et al. [13] Noise is filtered out of the video data before MAPO is used to estimate heart rate, with a Pearson correlation of 0.9541 and a Standard Error Estimate of 2.418, respectively. With the aid of nave Bayes and XG boost, the accuracy was determined to be 95% for nave Bayes and 85% for XG boost. After creating the labelled dataset, Yasser Zeinali et al. [14] from the MFCC (Mel Frequency Cepstral Coefficients) used classification algorithms before and after executing the dimension reduction and feature selection algorithms and found accuracy of 81.25 and 75% using GBC, RFC, and SVC.

Ibomoiye Domor Mienye [15] from cleveland and Framingham datasets, they proposed an ensemble learning for the prediction of heart disease risk using gradient boosting and random forest and found the accuracy 91 and 93% respectively. Asif Newaz et al. [16] from UCI repository released by kaggle they predict heart failure for comparison they used four machine learning algorithm for comparison purpose and achieved 67.93% accuracy. Dineo Mpanya et al. [17] taken dataset from MIT-BIH and Predicting mortality and hospitalization in heart failure using nine different machine learning and predict a model and find best accuracy as 63%. Anping Cai et al. [18] They employed real-time data and three types of machine learning, including random forest, support vector, and artificial neural network, to try to discover the best accuracy in predicting hypertension and heart failure treatment.

S.P. Patro et al. [19] collaborated with the UCI Machine Repository to predict heart disorders using a unique optimization technique. They discovered that the confusion matrix plots optimization approach is particularly effective in predicting heart diseases with a 93.3% accuracy rate. Md Mamun Ali et al. [20] used a dataset from kaggle to predict cardiac disease using supervised machine learning algorithms and three classification-based performance analyses with a rate of 100% accuracy and sensitivity. Farman Ali et al [21] from real time dataset they predict smart health caring diseases using ensemble deep learning with maximum accuracy is 98.5%. Sushmita Roy Tithi et al [22] from UCI repository they found ECG dataset and they analyses ECG signal and predict heart diseases by using different machine learning and such as RF, SVM, LR, KNN and ANN and find found logistic regression as best result. T.V.N. PREETAM [23] from MIT-BIH they collected ECG dataset and analyze ECG signal and predict heart disease using optimized neural network, the ECG is mainly used for diagnosis of heart disease, and found best accuracy. Subashini et al. [24] The UCI repository dataset is utilized for prediction. They discovered ECG signal and improved cardiac disease classification accuracy by applying image denoising technology from ECG signal and Bayes' shrink algorithm.

Deepika D et al. [25] The UCI repository dataset is utilized for prediction. In this study, a Multi-Layer Perceptron (MLP-EBMDA) with 94.28% accuracy was combined with enhanced Brownian motion on the basis of the Dragonfly Algorithm. Victor Chang et al. [26] For prediction, the CHB—MIT dataset is employed. They predict an artificial intelligence model for heart disease diagnosis using machine learning techniques, with 83% accuracy over training data, utilizing random forest. Ashley N. Beecy et al. [27] utilized a dataset from kaggle and XGboost for the best results in predicting 30-day unplanned readmission or all-cause death in heart failure. The CHB—MIT dataset is used for prediction by Luiz Eduardo Virgilio Silva et al. [28]. They used heart rate variability and machine learning to predict echocardiographic characteristics in Changes illness using random forest, neural network, support vector, and k-closest neighbor, and found the best result among them.

3 Materials and Methods

Dataset Description

We used the Kaggle-released UCI dataset in our recommended approach. In this table, there are 12 columns and 918 rows. Patient information in the dataset includes age, gender, kind of chest discomfort, resting blood pressure, serum cholesterol, fasting blood sugar, resting ECG findings, maximum heart rate reached, output class, and output class.

Type of Chest Pain

Typical Angina: Typical angina (TA) is characterized as sub sternal chest torment encouraged by actual effort or enthusiastic pressure and assuage with rest or dynamite. Ladies and older patients are generally have abnormal side effects both very still and during stress, frequently in the setting of no obstructive coronary course illness (CAD).

Atypical Angina: Abnormal chest anguish occurs when one experiences chest pain that does not fit the criteria for angina. Angina chest pain is a strain or crushing sensation that occurs when your heart muscle does not receive enough oxygenated blood.

NON-Aginal Pain: Non-cardiovascular chest anguish is a word used to describe. Non-cardiovascular chest pain is frequently described as feeling like angina, the chest pain caused by coronary artery disease. Behind the bosom bone, the patient feels a strain or crushing discomfort. The pain can also move to the neck, left arm, or back, according to some people. The annoyance can last for a few moments or for a long time.

Asymptomatic Pain: Asymptomatic means causing neither nor showing side effects of illness. Suggestive means an actual sign (rash, torment, uneasiness and so on) of infection or turmoil. For instance, red spots are indicative of measles and chest torment is suggestive of a respiratory failure (myocardial localized necrosis).

ST Segment Depression

The ST segment of an electrocardiogram (ECG) often addresses an electrically unbiased section of the complex between ventricular depolarization (QRS complex) and repolarization as seen in Fig. 1. (T-wave). If the ST segment is up sloping after exercise, the person is okay; however, if the ST segment is down sloping or horizontal after exercise, the person may be suffering from heart disease.

Methodology

We utilized the nine machine learning algorithms to process the dataset and predict cardiac illnesses. Such as 1. Random Forest, 2. Support vector machine 3. Catboost, 4. XGboost, 5. K-nearest Neighbour, 6. Naïve Bayes, 7. Logistic Regression, 8. Adaboost, 9. LGBM. We will study the algorithms with the highest accuracy and undertake a performance analysis of machine learning algorithm to predict cardiac illnesses using the provided method.

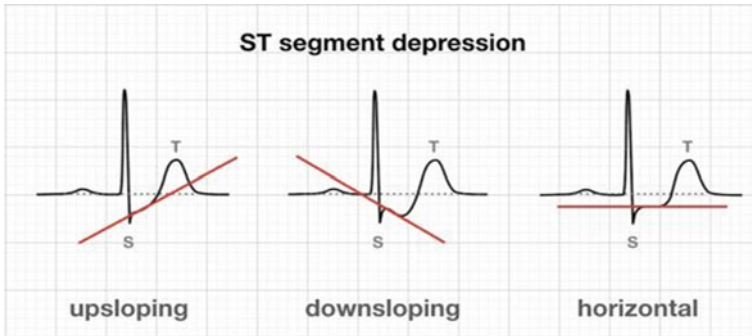


Fig. 1 ST segment depression

4 Implementation

We created a machine learning system to predict whether or not a person has heart disease in the proposed work. I used the UCI repository dataset from kaggle for this project. This study is based on nine machine learning models, including 1. Random Forest and 2. Support vector machine. 3. K-nearest Neighbour, 4. XGboost, 5. Catboost 6. Bayesian Inference, 7. Regression Logistic, 8. Adaboost, 9. LGBM.

Steps for Implementing the Model Step1:—Selection of Dataset

Data Overview: In this section dataset is collected from UCI repository having 918 row and 12 attribute.

Detect and Remove Outliers: Removal of unexpected data from the dataset.

Detect and Impute Missing Data: Get the data and find the missing value present in the dataset.

Applying Suitable Normalization Techniques: The data collected from UCI repository released by kaggle that consist of 12 column i.e. attribute and 918 row of data.

The data type of dataset is in six of them are in integer value, four of them are string value, 1 decimal value and 1 is other. By using python all the data type is converted into Nominal and Numeric format and use for the visualization purpose. The nominal value used in terms of 0 and 1.

Data Visualization: A graph chart or other graphic style is used to portray data or information. In machine learning, data visualization is essential because it highlights trends and patterns. We'll need to be able to grasp more and larger amounts of data as big data becomes more common. Machine learning makes it easier to analyze and forecast data, which may subsequently be used as a presenting tool.

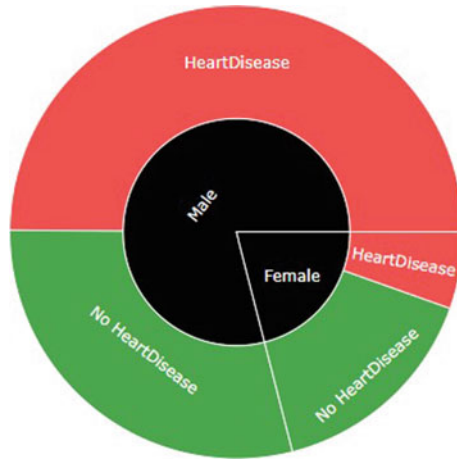


Fig. 2 ST Pie chart of Gender

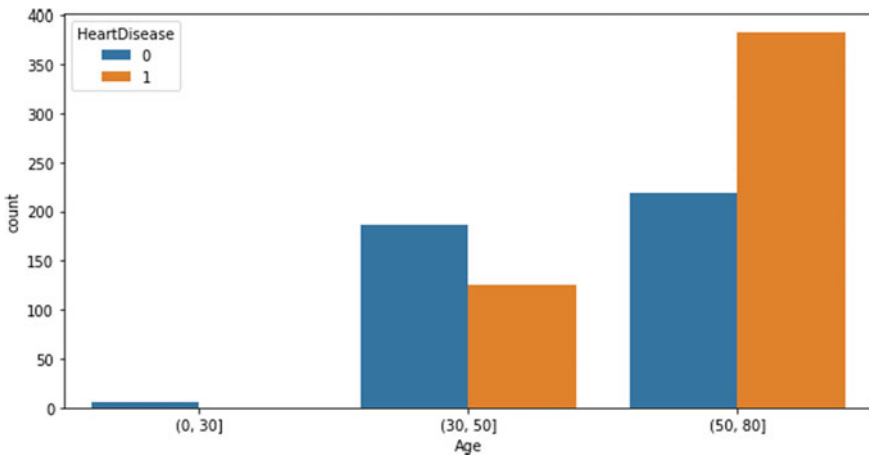


Fig. 3 Age

In the above Fig. 2 given dataset has total no of patient is 918 in which no of male is 725 and no of female is 193. In 725 male 458 has heart diseases and 267 male does not have heart diseases. Whereas in 192 female 50 female has heart diseases and 143 does not have heart diseases.

In above Fig. 3, the minimum age group is 28 and the maximum age group is 77. By visualizing the age group found that adult age 50 and older is more likely than younger people to have heart diseases.

Atypical Angina (ATA), Typical Angina (TA), Non-Angina Pain (NAP), and Asymptomatic Angina are the four types of chest pain shown in Fig. 4 (ASY). Asymptomatic people are more likely to get heart disease.

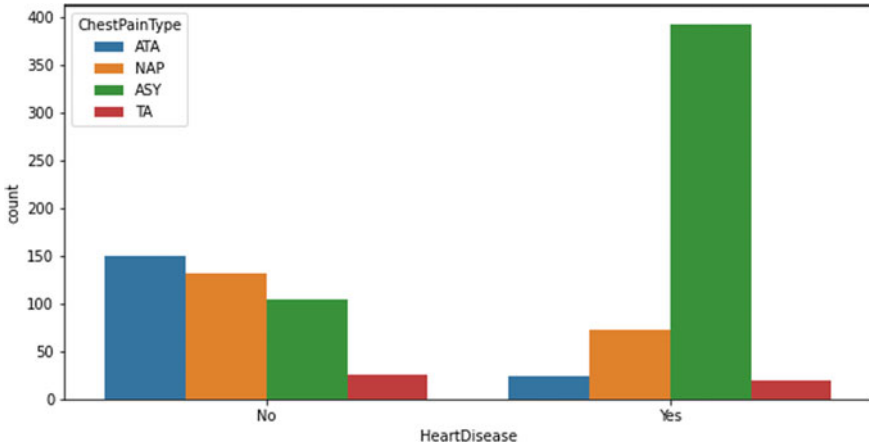


Fig. 4 Chest pain type

The above Fig. 5 represents the Resting ECG. Resting ECG means the ECG signal of the person who is in the rest. In this graph there are three types of ECG signals, Normal, ST and LVH. The slope of the peak exercise ST segment.

Flat is the most common type of ST_Slope among individuals with heart disease, as seen in Fig. 6, the ST segment shift relative to exercise-induced increases in heart rate.

Figure 7 Workout Angina is a sort of pain that develops after engaging in any type of physical activity. Strenuous exercise narrows the airways in the lungs, causing exercise-induced asthma. During or after exercise, it produces shortness of breath, wheezing, coughing, and other symptoms

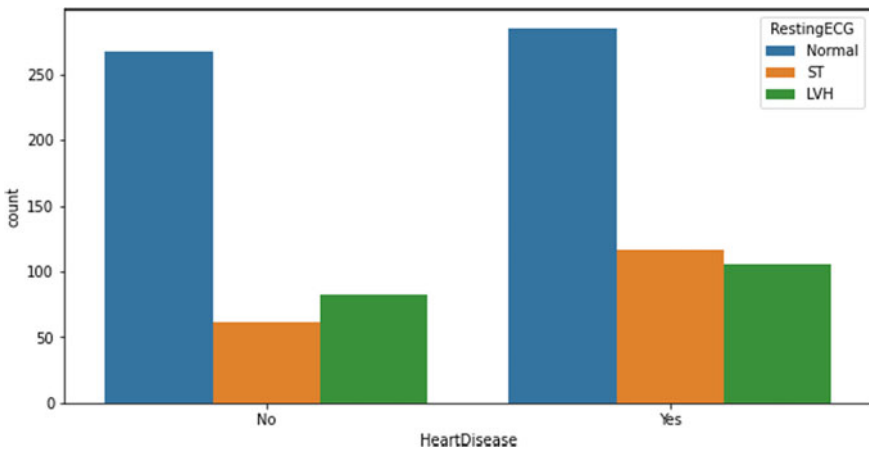


Fig. 5 Resting ECG

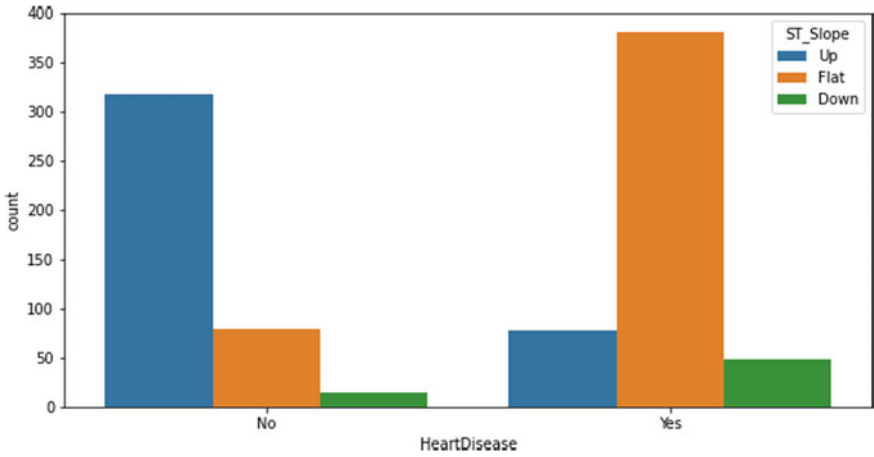


Fig. 6 ST_Slope

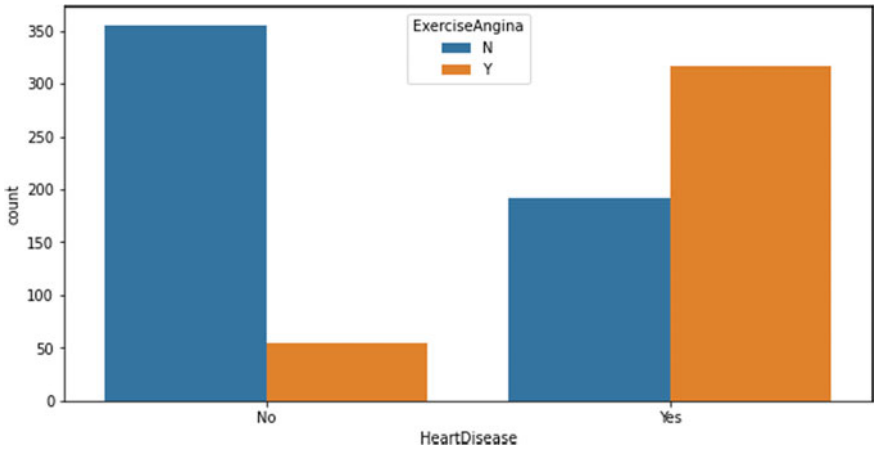


Fig. 7 Exercise Angina

The Fig. 8 represents Correlation Map when it ensures the existence of relationship between two variables then it is called as correlation. It could be either positive, negative or zero. When independent variable goes up and due this dependent variable also goes up, then it shows positive correlation. When one variable move upward and other downward then it is called as negative correlation. The value of correlation lies in between -1 to 1 .

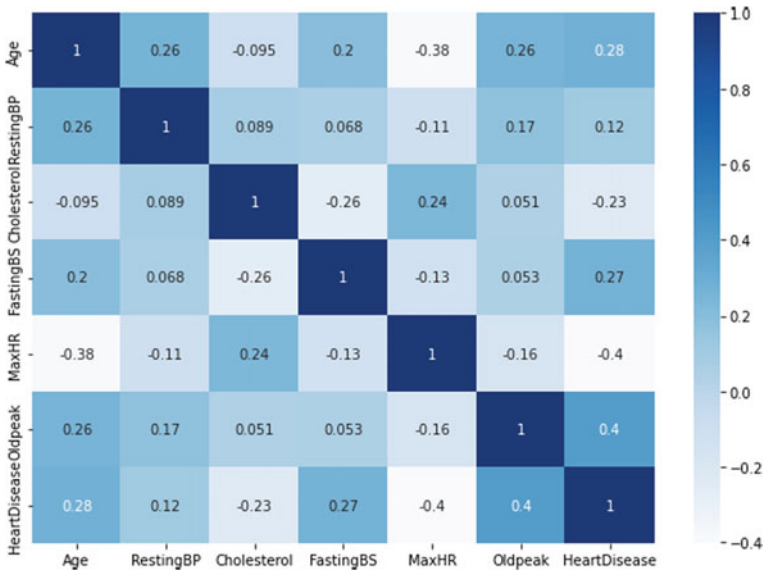


Fig. 8 Correlation map

Model Selection

Understanding Data Value (Classes):

In the dataset 12 attribute values are given one is target variable and other 11 is feature variable, the target variable is health diseases which is nominal value if diseases is there it is written as 1 and if the person is normal then written as 0.

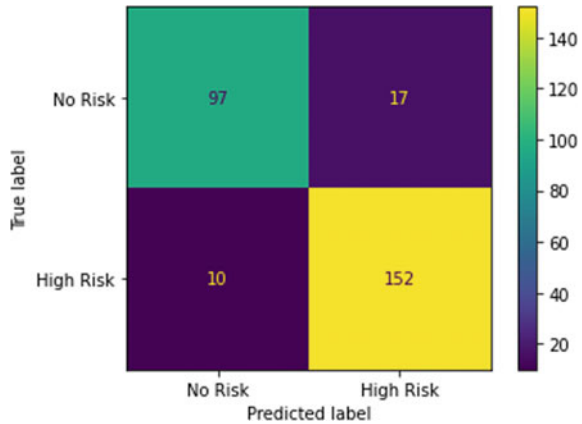
Machine Learning Model Selection:

As previously mentioned, nine machine learning algorithms are used to forecast heart problems in patients and to measure their effectiveness. This section contains a brief description of each model.

Logistic Regression

Logistic regression is a type of “supervised machine learning” in which regression analysis is used to learn and forecast the parameters in a dataset. Calculating binary classification probability underpins the learning and prediction activities. In a logistic regression model, the class variable must be binary classified. In this dataset, the goal column comprises two sorts of binary numbers: “0” for patients who are not at danger of heart failure and “1” for patients who are. Independent variables, on the other hand, can be classified as binary, nominal, or polynomial.

Fig. 9 Confusion matrix of logistic regression



Accuracy: For classification algorithms, it is the most popular performance metric. It’s the number of right guesses divided by the total number of predictions.

When Logistic Regression applied for predicting heart diseases a confusion matrix is prepared. The Fig. 9 show a Logistic Regression’s confusion matrix of two rows and two columns. Where the value TP is 97, meaning that there is no chance of heart diseases and model is correctly classified, FP value is 17, meaning there is chance of heart diseases but the model is incorrectly classified as no chance of heart diseases. FN value is 10, there is chance of heart diseases but the model is incorrectly classified. TN values is 152 means there is high chance of heart diseases and the model is correctly classified.

B. Naive Bayes

By calculating the probability of independent variables, Nave Bayes, a supervised learning classification model, achieves the same result. After calculating the likelihood of each class, the high probability class is assigned to the whole transaction. In a number of datasets, including educational and medical data mining, Nave Bayes is a prominent method for predicting classes. This method can be applied to a variety of datasets, including sentiment analysis and virus identification. Based on the values of independent variables, it predicts a pre-defined class for each record.

C. Random Forest

Random forest is the next model that was chosen and used in this investigation. Because it belongs to the classification family, this model is also known as supervised learning algorithm. During the learning phase, this model creates a forest of many random trees. When a dataset has “x” number of characteristics, it selects a feature called “y” at random. It creates nodes using the best rift strategy possible, taking into account all characteristics (i.e. “y”). Furthermore, the algorithm will be able to produce a full forest by repeating the preceding procedures. During the prediction phase, the software tries to link the trees using the expected outcome and voting

technique. The purpose of voting to merge random trees in a forest is to eliminate the tree with the most votes.

D. SVM

Support Vector is a widely used supervised learning approach for both classification and regression. The support vector machine approach aims to find a hyperplane in an N-dimensional space (N-number of features) that unambiguously categorises a data point. It is, however, largely used to solve categorization problems. Each data item is represented as a point in n-dimensional space, and the value of each feature is the value of a specific coordinate in the SVM algorithm. Then, to complete classification, we find the hyper-plane that clearly divides the two groups. Scikit-svm Multiple SVM algorithms are included in Learn's library as built-in classes.

E. XGBoost

XGBoost (Extreme Gradient Boosting) is a gradient boosting-based decision-tree-based ensemble Machine Learning technique. Regression, classification, ranking, and user-defined prediction problems can all be solved using it. Decision-tree-based algorithms are now considered best-in-class for small-to-medium structured data. Boosting iteratively constructs models from individual "weak learners". Unlike Random Forest, in a Boosting individual model are not completely built on random subsets of features. Sequentially puts more weight on instances with wrong prediction learns from past mistakes Gradient boosting uses gradient descent to minimize the loss function. Speed and performance makes popular to the XGBoost. In the proposed work XGBoost is used to predict heart diseases and find the accuracy, precision and recall.

F. CatBoost

CatBoost is the second model employed in the planned research. CatBoost is a new open-source machine learning algorithm developed by Yandex. The term "CatBoost" comes from the words "Category" and "Boosting," and it works with a variety of data types including audio, text, and images, as well as historical data. The term "Boost" is derived from the gradient boosting machine learning method. Gradient Boosting is a sophisticated machine learning technique that is frequently utilized in a variety of datasets, including the prediction of heart illnesses, market prediction, and forecasting. The primary motivation for adopting CatBoost is to improve performance by automatically addressing categorical information. CatBoost is robust enough to utilize in the proposed study since it reduces the amount of hyper-parameter adjustment required and reduces the risk of over-fitting, resulting in a more generic model. For classification, we can utilize the "CatBoostClassifier" model, which is used for both classifier and regression.

G. KNN

One of the most extensively used machine learning techniques is the K-Nearest-Neighbor algorithm. It can be used for both classification and regression issues, however we will just use it for classification in our proposed study. To utilize the KNN model, first pick the number “K” of neighbors, then compute their Euclidean distance, and then choose the K closest neighbors based on the calculated Euclidean distance. Count the data points in each category and assign them to the K value with the greatest value.

H. LGBM

Light GBM is a type of ensemble machine learning technique used in the proposed study for binary classification. It's a decision tree-based gradient boosting framework that's quick, distributed, and high-performing. It divided the tree into leaves. It can result in over fitting, which can be avoided by setting the splitting depth. There are numerous hyper-parameters that can be tweaked. In the proposed work, a light gradient boosting model is used to forecast a person's heart illness using two techniques. Gradient-based Onside Sampling and Exclusive Feature Bundling (EFB).

I. AdaBoost

AdaBoost is an ensemble learning. Ensemble learning is a learning technique in which multiple individual model combine together to create a master model, the process is called ensembling. It is implemented in proposed work with the use of python. Let's understand how Adaboosting work. Adaboosting work sequentially. It is computational scalability, uses dependency between models. It can handle missing value, outliers and mixed predictor as well (quantitive and qualitative).The technique used in AdaBoost is SAMME (Statewise Additive Multi-Modeling using Multiclass Exponential Loss Function).It is also called weak learner because its depth is always 1 and it takes the decision on only one feature.

The Nine model used for performance analysis to predict heart diseases such as 1. Random Forest, 2. Support vector machine 3. Catboost, 4. XGboost, 5. K-nearest Neighbor, 6. Naïve Bayes, 7. Logistic Regression, 8. Adaboost, 9. LGBM. In this study, the cross-validation method was used to test all model prediction outputs. The dataset's performance is measured using training and testing data. Heart illnesses have been discovered and predicted with the best accuracy of 90% using this recorded dataset and Logistic Regression.

5 Conclusion

The project “PERFORMANCE ANALYSIS OF MACHINE LEARNING ALGORITHMS FOR HEART DISEASES PREDICTION” assists us in predicting heart disease using human datasets such as age, gender, cholesterol, blood pressure, and other factors, hence lowering the mortality rate. In this work, several machine learning algorithms for classification were applied, and an optimal supervised approach for feature selection for heart disease prediction was created. The suggested system was assessed using certain criteria including precision, recall, and accuracy. F1-score, accuracy, and support the proposed system was compared to the existing system in terms of the supplied parameters and the testing data obtained to determine its efficiency in heart disease prediction. The results of the analysis revealed that the proposed approach outperformed established methods in terms of accuracy in predicting cardiac disease. The proposed approach outperformed the Logistic Regression by 90% in terms of prediction accuracy.

In the proposed work dataset is collected from UCI Repository, and applied ML in the Python. It is one of the popular programming language Machine Learning algorithm used in this proposed work to predict the Heart diseases. The evaluated results are compared on basis of validation score, cross validation score, precision, accuracy, recall and F1 score.

The below Fig. 10 is a compression chart which is created after applying all dataset to the model although dataset is same for each classifier but Logistic Regression has the best result at the rate of 90% and AdaBoost has worst among all at the rate of 85%.

In Fig. 11 Methods for determining a score for each of the input characteristics in a model are referred to as feature importance. The scores merely indicate how important each element is. A higher score suggests that the feature will have a bigger impact on the model used to anticipate a particular variable.

	Model	Validation Score	Cross_Validation Score
0	LogisticRegression	0.902174	0.864800
1	K-NN	0.873188	0.860203
2	Ada	0.858696	0.855627
3	Naive Bayes	0.876812	0.856500
4	Random Forest	0.880435	0.874167
5	SVC	0.880435	0.865217
6	CatBoost	0.884058	0.879843
7	LGBM	0.873188	0.865668

Fig. 10 Classification results

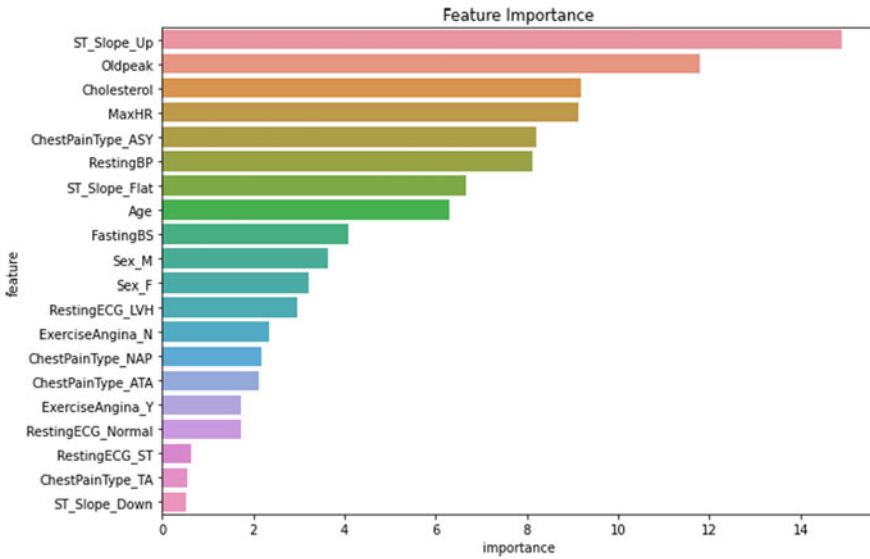


Fig. 11 Feature importance

6 Future Enhancement

The focus of future research will be on early detection of cardiac problems. It is advantageous for the patient to be treated before the condition becomes severe. This project will be processed in the future utilizing a significant quantity of data that can be obtained from any clinical organization, and this research can be carried out using a variety of machine learning approaches to produce improved multi-classification prediction. ECG signal processing for heart disease prediction with real time advancement is going to predict the disease earlier.

References

1. Zhang J et al Method of diagnosing heart disease based on deep learning ECG sign. School of Electronic Information, Wuhan University, vol 6141B0801010
2. Kumar S (2020) Heart disease detection using radial basis function classifier. Department of Information Technology, Ponnaiyah Ramajayam Institute of Science and Technology, India, vol 01, September 2020
3. Preetam TVN et al (2020) ECG signal analysis and prediction of heart attack with the help of optimized neural network, ISSN 2231-3990, April 2020
4. Rengasamy S, Surianarayanan C, Chellaih PR (2020) Machine learning based method for prediction of heart disease in big data environment 9(6). ISSN 2278-3075
5. Durairaj M, Ramasamy N (2015) Intelligent prediction methods and techniques using disease diagnosis in medical database, pp 2153–2160

6. Chen J et al (2019) Smart heart monitoring: early prediction of heart problems through predictive analysis of ECG signals, vol 7. IEEE
7. Butchi Raju K, et al (2022) Smart heart disease prediction system with IoT and fog computing sectors enabled by cascaded deep learning model 2022:22, Article ID 1070697
8. Zadawale SS et al (2017) ECG signal based heart disease prediction system using DWT and SVM. IJARCCCE 6(7)
9. Vafaie MH et al (2014) Heart diseases prediction based on ECG signals' classification using a genetic-fuzzy system and dynamical model of ECG signals, pp 291–296, August 2014
10. Smole T et al (2021) A machine learning-based risk stratification model for ventricular tachycardia and heart failure in hypertrophic cardiomyopathy, vol 135
11. Taloba AI et al (2021) Machine algorithm for heartbeat monitoring and arrhythmia detection based on ECG systems 2021:9, Article ID 7677568
12. Valarmathi R et al (2021) Heart disease prediction using hyper parameter optimization (HPO) tuning, vol 30
13. Sharmaa P et al (2020) Artificial plant optimization algorithm to detect heart rate and presence of heart disease using machine learning, vol 102
14. Zeinali Y et al (2022) Heart sound classification using signal processing and machine learning algorithms, vol 7
15. Newaz A et al (2021) Survival prediction of heart failure patients using machine learning techniques, vol 26
16. Mpanya D et al (2021) Predicting mortality and hospitalization in heart failure using machine learning: a systematic literature review, vol 34
17. Cai A et al (2021) The use of machine learning for the care of hypertension and heart failure. JACC ASIA 1(2)
18. Patro SP et al (2021) Heart disease prediction by using novel optimization algorithm: a supervised learning prospective, vol 26
19. Ali MdM et al (2021) Heart disease prediction using supervised machine learning algorithms: performance analysis and comparison, vol 136
20. Ali F et al (2020) A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion 23:208–222
21. Tithi SR et al (2019) ECG data analysis and heart diseases prediction using machine learning algorithms. IEEE. ISBN 978-1-72-810297-9
22. Preetam TVN (2020) ECG signal analysis and prediction of heart attack with the help of optimized neural network. Alochana Chakra J 9(4). ISSN 2231-3990
23. Subashini et al (2019) Enhancing the Classification Accuracy of Cardiac Diseases using Image Denoising Technique from ECG signal. IEEE. ISBN 978-1-53-869471-8
24. Beecy AN, Gummala M et al (2020) Utilizing electronic health data and machine learning for the prediction of 30-day unplanned readmission or all-cause mortality in heart failure, Elsevier Inc., pp 2666–6936
25. Silva LEV et al (2021) Prediction of echocardiographic parameters in Chagas disease using heart rate variability and machine learning, pp 1746–8094. Elsevier Ltd.
26. Sivabalaselvamani D, Selvakarthi D, Yogapriya J, Thiruvenkatasuresh MP, Maruthappa M, Chandra AS (2021) Artificial intelligence in data-driven analytics for the personalized healthcare. In: 2021 international conference on computer communication and informatics (ICCCI), pp 1–5. IEEE, January 2021
27. Hemalatha S, Tamilarasi A, Kavitha T, Sivabalaselvamani D, Raj MK (2022) A crossbreed framework for heart disease prediction using SVM and rough set techniques. In: 2022 international conference on computer communication and informatics (ICCCI), pp 1–5. IEEE, January 2022
28. Pandian AP (2019) Review of machine learning techniques for voluminous information management. J Soft Comput Paradigm 1(2):103–112

Fluorescence Microscopic Image Reconstruction Using Variational Autoencoder and CycleGAN



Marrivada Gopala Krishna Sai Charan, S. S. Poorna, K. Anuraj,
Choragudi Sai Praneeth, P. G. Sai Sumanth,
Chekka Venkata Sai Phaneendra Gupta, and Kota Srikar

Abstract Many times, noise corrupts images, especially fluorescence microscopic data. Conventional methods for increasing the Signal to Noise ratio (SNR) of corrupted images, such as deconvolution frequently fail to achieve a high SNR since only an estimate of the point spread function is available due to modelling deficiencies or complications. In comparison to statistical approaches, deep learning methods significantly enhanced the SNR of reconstructed images. Deep learning algorithms are computationally simpler while still outperforming approaches that are computationally intensive. In this work, we attempt to reconstruct images using Variational Autoencoders and CycleGAN. Metrics like Peak Signal to Noise Ratio (PSNR) and Mean Square Error (MSE) are used for evaluating the quality of reconstruction.

Keywords VAE · CycleGAN · Image reconstruction · Fluorescence microscopic data

1 Introduction

Deep learning (DL) has forayed into many fields which were earlier dominated by conventional methods. Corruption of Fluorescence microscopic images due to various non-idealities will degrade the usefulness of those images to the biologists. Until recently, reconstruction of images were done by conventional approaches like Block Matching and 3D Filtering (BM3D) which is a Gaussian denoising technique, Poisson denoising techniques like Variance-Stabilizing Transformation (VST) or a combination of these etc. Deep learning methods superseded these approaches by providing better SNR for almost every image reconstruction task. Autoencoders try to convert the higher dimensional data to a lower dimension and later reconstruct the input from the lower dimensional data. By imposing a reconstruction loss for

M. G. K. S. Charan · S. S. Poorna (✉) · K. Anuraj · C. S. Praneeth · P. G. Sai Sumanth ·
C. V. S. P. Gupta · K. Srikar
Department of Electronics and Communication Engineering, Amrita Vishwa Vidyapeetham,
Amritapuri, India
e-mail: poornass@am.amrita.edu

uncorrupted image and corrupted copy of the image, using autoencoder, one can force the it to learn the latent structure of data, which will help in denoising tasks. Variational Autoencoders are widely used for image reconstruction tasks. Generative Adversarial Networks are widely used for sampling and synthesis tasks. CycleGAN tries to map one distribution to the other. In this work, we attempt to show that the output of the autoencoder which is an approximate version to the uncorrupted data or the ground truth can be fed to CycleGAN to further improve the reconstruction of image.

2 Literature Review

In [1] authors compared conventional approaches used in image reconstruction like BM3D, VST with DL based reconstruction techniques, employing Convolutional Neural Networks (CNN). They used a CNN with two branches each containing Convolutional and Deconvolutional layers. Convolutional layers were used for encoding/compressing whereas deconvolutional layers, for decoding/decompressing. Skipped connections or residual connections were also added for capturing non linear characteristics. Skipped connections also make back propagation easy by alleviating vanishing gradients. One major drawback of using deconvolutions is block artifacts, which make upsampling difficult. CNNs are not particularly good at learning the latent representation of data.

A survey of various methods used for medical image reconstruction is given in [2]. They also compared the weakness and strengths of various approaches. They mentioned that Deep learning techniques lack proper theory about their inner working and large amount of data is required for training them to a reasonable accuracy. This work also reviewed various DL applications for image reconstruction like Magnetic Resonance Imaging (MRI), Computed Tomography (CT), Positron Emission Tomography (PET) etc. Various deep learning frameworks which are used for Image reconstruction task are also listed here in this work.

In [3] the authors implemented various autoencoders for medical image reconstruction. They implemented Deep autoencoders along with Artificial Neural Network (ANN), CNN and Deep Boltzmann machines. Deep autoencoders can help networks learn highly non linear functions which would otherwise be difficult/impossible to learn with shallow autoencoder networks. Compression will be higher for deeper networks which results in efficient encoding. They achieved MSE of 5.31 and a PSNR of 35 dB on NIH clinical chest X-ray image dataset.

Multi-spectral images for image reconstruction task is used in [4]. They proposed Variational Auto Encoder-Generative Adversarial Network (VAE-GAN) architecture for reconstruction of image. Color images/RGB images are limited in amount of data they contain compared to Multi-Spectral Images (MSI). The paper proposed reconstruction of RGB images from MSI. They defined the total loss function by adding mean square error loss of input image and reconstructed image and weighted loss of KL-divergence and Cross entropy loss for GAN. Cross entropy loss used by

the authors is not directly related to the quality of image generated by the GAN, it also suffers from Discriminator saturation problem which creates a trouble for back-propagating gradients.

A new unsupervised learning algorithm GANs for generating images from noise was introduced in [5]. Authors used Artificial Neural Networks for generating gray scale images. In [6], they replaced Artificial Neural Network with CNN and also introduced latent vector algebra and latent vector interpolation which give semantically-correct outputs. In [7] the authors introduced Cycle-consistent Generative Adversarial Network (CycleGAN) for mapping of image from one domain to other domain, this GAN can be used for image enhancement, Neural-style transfer etc. Works [8–11] improved GANs and made them suitable for hi-res generative modeling and handling images with large sizes like 512×512 , 1024×1024 etc. In [12], authors compared traditional machine learning methods with Deep Convolutional neural networks for melanoma detection. In [13], the authors compared various state-of-the-art Convolutional Neural Network architectures like ResNet-152, ResNet101, VGG16, VGG19, AlexNet for image classification problem. In [14], the discussed about various aspects of Convolutional Neural Networks like convolutional layer, various types of pooling, various types of activations, various popular frame works available for efficient implementation of Neural Networks. In [15], the authors used Microsoft Kinect for Three-dimensional(3D) geometric processing for Indian Sign Language Recognition.

3 Preliminaries

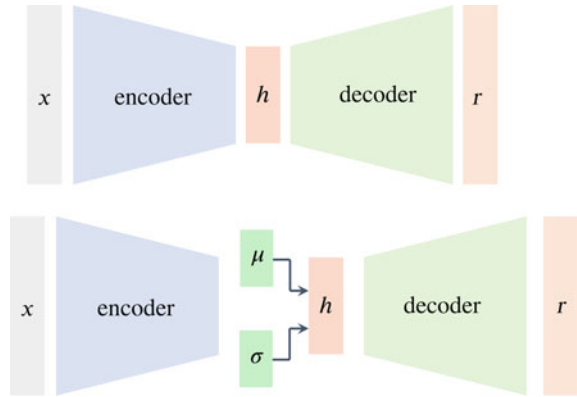
3.1 Variational Autoencoder

Autoencoder [16] is an unsupervised learning algorithm which tries to output a faithful copy of the input. This faithful copy is called *Reconstruction* and it contains two networks *encoder* and *decoder*. Encoder maps input to a latent space and decoder maps latent space to output. Encoder can be consider as a mapping f and Decoder can be considered as a mapping g . Latent space vector can be denoted by h and r is the reconstruction. This is summarised in Eq. (1) If the $dim(f) > dim(h)$, where dim represents dimension of the space, then the mapping cannot be one-one and it would help capture the commonalities of data, mapping similar features to same feature in the latent space. Such kind of autoencoder is called *Undercomplete*.

$$\begin{aligned} f : x &\xrightarrow{\text{encoder}} h \\ g : h &\xrightarrow{\text{decoder}} r \end{aligned} \tag{1}$$

Our objective will be to minimize the reconstruction loss for making output to be a faithful copy of the input.

Fig. 1 (up) Autoencoder with encoder and decoder, (down) Variational Autoencoder with a prior probability distribution on latent space/code



Objective to be minimized : $\mathcal{L}(x, g(f(x)))$ (2)

In Eq. (2), \mathcal{L} represents the Reconstruction loss function that needs to be minimized. Reconstruction loss can be chosen to be *Mean Square loss*.

$$\text{Reconstruction loss} = \|x - r\|^2 \tag{3}$$

$f(g(x))$ can never be an identity function because of *undercompleteness* and hence this network can be used to learn useful features of the data. If $\text{dim}(f) < \text{dim}(h)$, then such a decoder is more prone to making $f(g(x))$ an identity function which is not a useful thing. Such type of autoencoder is called *Overcomplete* (Fig. 1).

Variational autoencoders assume a prior on the latent space, hence they differ from normal autoencoders. Usually the prior probability distribution for h is chosen to be Normal distribution $\mathcal{N}(\mu, \sigma^2)$. We have two probability distributions for h , one prior probability distribution that is assumed, other the inferred probability distribution from the encoder. Hence, we can define an objective to minimize the distance between the two probability distributions. One such distance metric is *Kullback-Leibler Divergence*. We define the following terms:

1. $p(h|x; \theta)$ is the probability distribution inferred from encoder.
2. $p(h)$ is the prior probability distribution assumed for the latent space
3. $p(r|h; \psi)$ is the probability distribution of the reconstructed output inferred from the decoder

Regularization loss can be defined between inferred distribution for h and prior probability distribution for h as follows:

$$\begin{aligned} \text{Regularization loss} &= D_{KL}(p(h|x; \theta) \| p(h)) \\ \text{where, } D_{KL} &\text{ is Kullback-Leibler Divergence} \end{aligned} \tag{4}$$

Total loss can be computed by adding Eqs. (3) and (4) as follows:

$$\begin{aligned} \text{Total loss}_{VAE} &= \text{Reconstruction loss} + \text{Regularization loss} \\ &= \|x - r\|^2 + \beta D_{KL}(p(h|x; \theta) \| p(h)) \end{aligned} \tag{5}$$

In Eq. (5), β is the *Regularization factor*. For sampling of h , Normal Distribution $\mathcal{N}(\mu, \sigma^2)$ can be used it can be obtained from a mean vector μ which is fixed and sampling ϵ from $\mathcal{N}(0, 1)$ as follows:

$$h = \mu + \sigma \odot \epsilon \tag{6}$$

In Eq. (6), \odot represents *Hadamard product* or element wise multiplication of 2 vectors.

3.2 Generative Adversarial Network

Instead of using *Maximum Likelihood* for learning the representation of data, GANs directly sample the data. GANs [5] consists of 2 networks *Generator* and *Discriminator* each with a different objective (Fig. 2).

Discriminator of an unconditional GAN is a binary classifier which outputs 0/1 for real/fake. If we use Binary Cross Entropy (BCE) as loss function for Generator and Discriminator, then Discriminator loss is defined in Eq. (7)

$$\begin{aligned} \text{Loss}_{DIS} &= \text{Real loss} + \text{fake loss} \\ &= BCE(D(x_{real}), y_{real}) + BCE(D(x_{fake}), \mathbf{0}_{1 \times n_{fake}}) \end{aligned} \tag{7}$$

where,

Loss_{DIS} is the loss of the Discriminator,

x_{real} are a batch of real images,

y_{real} are labels of batch of real images,

x_{fake} are a batch of images generated by the Generator,

$D(x_{real})$ are the outputs of discriminator for real images $D(x_{real}) \in (0, 1)$

$D(x_{fake})$ are the outputs of Discriminator for fake images $D(x_{fake}) \in (0, 1)$

$\mathbf{0}_{1 \times n_{fake}}$ is a zero vector of dimensions $0 \times$ number of fake images

Fig. 2 Forward pass of a GAN, which involves Generator transforming noise into image and Discriminator taking both real and fake images as inputs and classifies them as real or fake (opaque lines show the forward path)

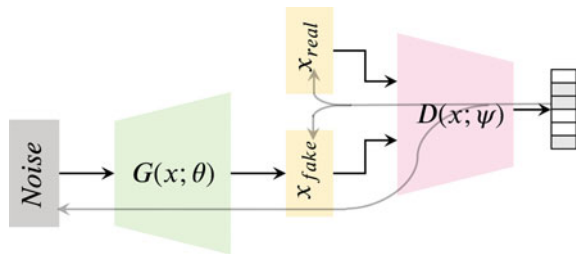


Fig. 3 Backpropagating outputs of discriminator to the inputs of the discriminator for computing gradients which are used to update the weights of the discriminator

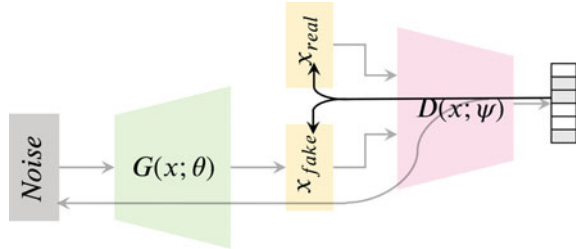
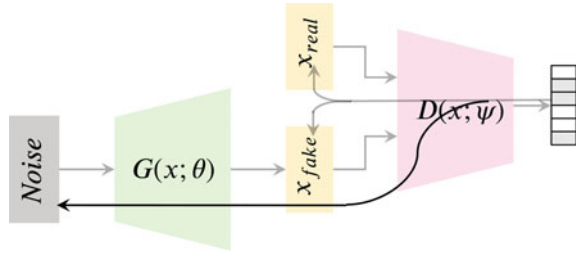


Fig. 4 Backpropagating outputs of discriminator to the inputs of the generator for computing gradients which are used to update the weights of the generator. Discriminator also propagates gradients which is like a feedback discriminator gives for generator



Generator loss can be defined as follows

$$\text{Loss}_{GEN} = BCE(D(x_{fake}), \mathbf{1}_{1 \times n_{fake}}) \tag{8}$$

where,

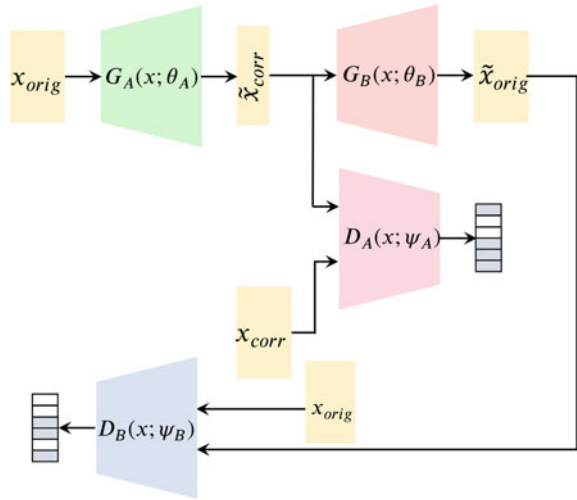
Loss_{GEN} is the loss of the generator,

$\mathbf{1}_{1 \times n_{fake}}$ is the unit vector of length $1 \times$ number of fake images (Figs. 3 and 4)

We can also choose different loss functions for generator and discriminator. Better feedback to generator can be given by training the discriminator more times compared to generator. In [7], the authors proposed a new method for mapping images from one domain to another domain. They also defined *Cyclic Consistency Loss* for reconstructed outputs of both the generators to be closer in space to the corresponding inputs of Generators. This objective has to be minimized for getting points closer in space. It consists of 2 Generators and 2 discriminator for giving feedback to both the generators. In this work we provide the roles of 2 generators and 2 discriminator as follows:

1. GeneratorA will take original data as input
2. GeneratorB receives the compressed data as output and tries to reconstruct the input of Generator A.
3. DiscriminatorA receives corrupted data and compressed output by GeneratorA as input.
4. DiscriminatorB receives original data and reconstructed data by GeneratorB as input.

Fig. 5 CycleGAN: \tilde{x}_{corr} is the reconstructed version of corrupted image, \tilde{x}_{orig} is the reconstructed version of original image, All other symbols have same meaning as defined for Eq. (9)



New Objective for CycleGAN becomes:

$$\begin{aligned} \mathcal{L}(G_A, G_B, D_A, D_B) = & \mathcal{L}_A(G_A, D_A, x_{corr}, x_{orig}) \\ & + \mathcal{L}_B(G_B, D_B, x_{corr}, x_{orig}) \\ & + \gamma \mathcal{L}_{cyc}(G_A, G_B) \end{aligned} \quad (9)$$

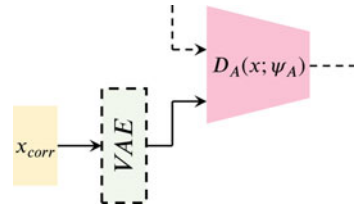
where,

- G_A -GeneratorA, G_B -GeneratorB,
- D_A -DiscriminatorA, D_B -DiscriminatorB,
- \mathcal{L}_A is the objective for GeneratorA,
- \mathcal{L}_B is the objective for GeneratorB,
- \mathcal{L}_{cyc} is the cyclic consistency loss as defined in [7]
- γ is the regularization factor,
- x_{corr} are corrupted images,
- x_{orig} are original images

3.3 Training of CycleGAN and VAE

For training CycleGAN and VAE, Confocal microscopic images of Fluorescence Microscopy Denoising (FMD) [17] dataset is used. This dataset consists of 256×256 images which are corrupted by Poisson-Gaussian noise. Instead of using x_{corr} as an input to $D_A(x; \psi_A)$, the output of the Variational autoencoder $VAE(x_{corr})$ is given as input to $D_A(x; \psi_A)$, shown in Fig. 5. We call this as REFINEMENT. This REFINEMENT

Fig. 6 $VAE(x_{corr})$ is applied to D_A instead of directly applying x_{corr} to D_A for CycleGAN shown in Fig. 5



will help in better reconstruction of corrupted images as the underlying representation of data is learned by the VAE (Variational autoencoder), which is a *refined* version of corrupted images. ResNet-18 encoder and ResNet-18 decoder used in [18] is adopted in this work for implementation of VAE. CycleGAN architecture in [7] is adopted with minor modifications. For cleaning of images, averaging is done as noise is a high frequency signal and imposes an effect of low pass filtering (Fig. 6).

$$\text{Total loss function} = \mathcal{L}(G_A, G_B, D_A, D_B) + \lambda \mathcal{L}_{VAE} \tag{10}$$

Where,

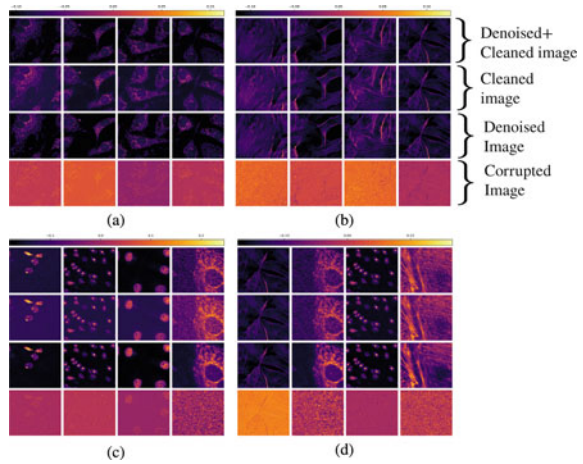
- \mathcal{L}_{VAE} is the objective of Variational Autoencoder,
- $\mathcal{L}(G_A, G_B, D_A, D_B)$ is the objective of CycleGAN,
- λ is the regularization factor

4 Results of CycleGAN and VAE on FMD Dataset

As mentioned earlier all the images are corrupted with Poisson-Gaussian noise. All models are implemented in PyTorch framework and trained on Nvidia P100 GPU. Corrupted images, denoised images, cleaned images, cleaned+denoised images are shown in Fig. 7. We can observe that denoised images and denoised+cleaned images have significantly better SNR (Signal to Noise ratio) compared to original corrupted image. Metrics such as PSNR, Root Mean Square Error (RMSE) are used for evaluating the quality of reconstructions. PSNR and RMSE are defined as follows in Eq. (11).

$$\begin{aligned} MSE &= \|\tilde{x}_{orig} - x_{orig}\|_F^2, \\ PSNR &= 10 \log_{10} \left(\frac{255^2}{MSE} \right), \\ RMSE &= \sqrt{MSE} \end{aligned} \tag{11}$$

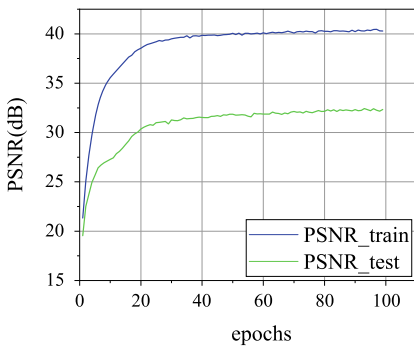
Fig. 7 In **a, b, c, d** Grid of images correspond to color maps of (corrupted, denoised, clean, denoised+clean) from lowest row to highest row



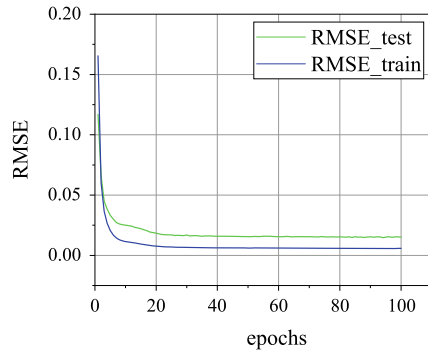
where,

- \tilde{x}_{orig} is the reconstructed version of tensor of corrupted images,
- x_{orig} is the tensor of original images,
- $RMSE$ is the Root Mean Square Loss ,
- $\| \cdot \|_F$ is Frobenius norm

This model is trained for 100 epochs. PSNR for test and train sets vs No. of epochs and RMSE for test and train sets vs No. of epochs are shown in Fig. 8. At the end of the training, PSNR, RMSE of the model for test set and train set respectively are (32.33,40.28 dB) and (1.51e−2, 5.8e−3). Initial PSNR of images in the test set is around 15 dB. At the end of the training PSNR improved by roughly 17 dB, which shows significant improvement.



(a) PSNR vs No.of Epochs



(b) RMSE vs No.of Epochs

Fig. 8 PSNR, RMSE vs No. of epochs for train and test sets

5 Conclusion

The paper proposes a Variational Autoencoder combined with CycleGAN for image reconstruction for Fluorescence Microscopic Image. The corrupted image quality after reconstruction is evaluated using performance metrics viz. PSNR is significantly high and RMSE is significantly low, which indicates the effectiveness of this method.

References

1. Liu P-Y, Lam EY (2018) Image reconstruction using deep learning. arXiv preprint [arXiv:1809.10410](https://arxiv.org/abs/1809.10410)
2. Ahishakiye E, Van Gijzen MB, Tumwiine J, Wario R, Obungoloch J (2021) A survey on deep learning in medical image reconstruction. *Intell Med* 1(03):118–127
3. Saravanan S, Sujitha J (2020) Deep medical image reconstruction with autoencoders using deep Boltzmann machine training. *EAI Endorsed Trans Pervasive Health Technol* 6(24):e2
4. Liu X, Gherbi A, Wei Z, Li W, Cheriet M (2020) Multispectral image reconstruction from color images using enhanced variational autoencoder and generative adversarial network. *IEEE Access* 9:1666–1679
5. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. *Adv Neural Inf Process Syst* 27
6. Radford A, Metz L, Chintala S (2015) Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint [arXiv:1511.06434](https://arxiv.org/abs/1511.06434)
7. Unpaired image-to-image translation using cycle-consistent adversarial networks
8. Brock A, Donahue J, Simonyan K (2018) Large scale GAN training for high fidelity natural image synthesis. arXiv preprint [arXiv:1809.11096](https://arxiv.org/abs/1809.11096)
9. Kavalerov I, Czaja W, Chellappa R (2021) A multi-class hinge loss for conditional GANs. In: *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp 1290–1299
10. Zhao S, Liu Z, Lin J, Zhu J-Y, Han S (2020) Differentiable augmentation for data-efficient GAN training. *Adv Neural Inf Process Syst* 33:7559–7570
11. Karras T, Aittala M, Laine S, Härkönen E, Hellsten J, Lehtinen J, Aila T (2021) Alias-free generative adversarial networks. *Adv Neural Inf Process Syst* 34
12. Poorna SS, Ravi Kiran Reddy M, Akhil N, Kamath S, Mohan L, Anuraj K, Pradeep HS (2020) Computer vision aided study for melanoma detection: a deep learning versus conventional supervised learning approach. In: *Advanced computing and intelligent engineering*. Springer, Singapore, pp 75–83
13. Bharath Chandra BV, Naveen C, Sampath Kumar MM, Sai Bhargav MS, Poorna SS, Anuraj K (2021) A Comparative study of drowsiness detection from EEG signals using pretrained CNN models. In: *2021 12th international conference on computing communication and networking technologies (ICCCNT)*, pp 1–3. <https://doi.org/10.1109/ICCCNT51525.2021.9579555>.
14. Aloysius N, Geetha M (2017) A review on deep convolutional neural networks. In: *International conference on communication and signal processing (ICCS)*, pp 0588–0592. <https://doi.org/10.1109/ICCS2017.8286426>
15. Geetha M, Manjusha C, Unnikrishnan P, Harikrishnan R (2013) A vision based dynamic gesture recognition of Indian Sign Language on Kinect based depth images. In: *2013 international conference on emerging trends in communication, control, signal processing and computing applications (C2SPCA)*, pp 1–7. <https://doi.org/10.1109/C2SPCA.2013.6749448>
16. Goodfellow I, Bengio Y, Courville A (2016) *Deep learning*. MIT Press
17. Mannam V, Zhang Y, Zhu Y, Howard S (2019) Fluorescence microscopy denoising (FMD) dataset. Notre Dame. <https://doi.org/10.7274/r0-ed2r-4052>

18. Ou X, Yan P, Zhang Y, Bing T, Zhang G, Jianhui W, Li W (2019) Moving object detection method via ResNet-18 with encoder-decoder structure in complex scenes. *IEEE Access* 7:108152–108160
19. Liao Y, Xiong P, Min W, Min W, Lu J (2019) Dynamic sign language recognition based on video sequence with BLSTM-3D residual networks. *IEEE Access* 7:38 044-38 054
20. Miyato T, Kataoka T, Koyama M, Yoshida Y (2018) Spectral normalization for generative adversarial networks, arXiv preprint [arXiv:1802.05957](https://arxiv.org/abs/1802.05957)
21. Ioffe S, Szegedy C (2015) Batch normalization: accelerating deep network training by reducing internal covariate shift. In: *International conference on machine learning*, pp 448–456
22. Lee KS, Town C (2020) Mimicry: towards the reproducibility of GAN research. arXiv preprint [arXiv:2005.02494](https://arxiv.org/abs/2005.02494) (2020)
23. Miyato T, Koyama M (2018) cGANs with projection discriminator. arXiv preprint [arXiv:1802.05637](https://arxiv.org/abs/1802.05637)
24. Odena A, Dumoulin V, Olah C (2016) Deconvolution and checkerboard artifacts. *Distill* 1(10):e3
25. Nguyen A, Clune J, Bengio Y, Dosovitskiy A, Yosinski J (2017) Plug & play generative networks: Conditional iterative generation of images in latent space. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp 4467–4477

Tomato Leaf Disease Detection Based on Convolutional Neural Network



Jagmohan Sahu and Pavan Kumar Mishra

Abstract Tomatoes are a popular and important crop in India, with a large economic price and great production capacity. Diseases harm the health of the plant, which has an impact on its growth. It is critical to monitor the progress of the farmed crop to guarantee minimal losses. There are a slew of tomato diseases that are wreaking havoc on the crop's leaves. One of the major linkages in the avoidance and control of crop diseases is the identification of infections in the leaf portions during the planting phase. Tomato leaves, including six popular species (Bacterial Spot, Black Mold, Early Blight, Late Blight, Mosaic Virus, and Septoria Spot), are used as experimental objects in this work to extract disease features from the leaf surface. Deep learning-based disease identification might help prevent such a catastrophe. A Convolutional Neural Network (CNN) is a type of deep learning algorithm that is currently commonly used for image categorization. In our studies, we used the CNN architecture to identify diseases in tomato leaves. This data set contains 2800 pictures of plant diseases. The Convolutional Neural Network was used in our proposed system to detect plant leaf diseases in seven categories, comprising six classes for diseases found in various plants and one class for healthy leaves. As a result, we were able to attain remarkable training and testing accuracy, with a training accuracy of 97.190% and a testing accuracy of 96.607% for all data sets used.

Keywords Deep learning · Leaf diseases · Convolutional neural network · Transfer learning

J. Sahu (✉) · P. K. Mishra
Department of Information Technology, National Institute of Technology, Raipur, India
e-mail: jagmohansahu199@gmail.com

P. K. Mishra
e-mail: pavanmishra.it@nitrr.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
P. P. Joby et al. (eds.), *IoT Based Control Networks and Intelligent Systems*,
Lecture Notes in Networks and Systems 528,
https://doi.org/10.1007/978-981-19-5845-8_31

437

1 Introduction

In India the agriculture industry employs the bulk of the citizens. The most favored vegetable in India is tomatoes. The three most important antioxidants found in tomatoes are vitamin E, vitamin C, and beta-carotene. They're also high in potassium, a mineral that's essential for overall health. India's tomato-growing region is estimated to be at 3,50,000 hectares, with output volumes totaling around 53,00,000 tonnes. India is now the third massive tomato grower in the world. Diseases are widespread in the tomato crop at all stages of its growth because of crop reactivity and environmental circumstances. Plants infected with disease account for 10–30% of overall crop losses [1]. The detection of such diseases in plants is critical for avoiding significant agricultural production losses in terms of yield and volume. Because of its complexity, manually monitoring plant diseases is a tough and time-consuming undertaking. As a result, there is a need to decrease the amount of human labor required for this work while still providing correct predictions and ensuring the lives of farmers are as stress-free as possible. Many farmers make incorrect conclusions about the diseases because visually evident trends are tough to discern at a quick glance [2]. As a result, farmers' preventative measures may be ineffectual and, in some cases, detrimental. They don't have access to professional guidance on how to manage their crop infection since they don't have it. Farmers frequently band together and apply common disease preventive techniques. There have been instances when overdosing or underdosing of the insecticide resulted in crop damage owing to a lack of understanding or misreading of the disease's severity. This is the driving force behind the suggested technology, which attempts to precisely identify and categorize diseases in tomato plants. The methods proposed in the research apply to the most frequent diseases that affect tomato plants, such as bacterial leaf spot and septoria leaf spot, as well as early blight and other infections. Any leaf picture can be categorized into one of the disease classifications, or it can be declared healthy. We suggested a unique method for detecting diseases in tomato plants. Farmers will be capable of addressing their detection of plant diseases without the need to chase down plant experts. It will therefore assist them in curing the plant's sickness in a timely manner, enhancing the quality and quantity of food crops produced, as well as the farmer's profit. We used kaggle to get the tomato leaf dataset for this experiment. Subsequently downloading the dataset, we design a Convolution Neural Network model to categorize the photographs. The performance of a pre-trained model was assessed using a range of criteria, including training accuracy, validation accuracy, and the number of trainable and testable parameters.

The following is how the paper is organized: The literature review of the available techniques is discussed in Sect. 2. Section 3 explains the recommended methodology and technique, as well as the steps performed to accomplish the intended goals. Section 4 deals with the results and interpretation of the proposed technique. The conclusion of the article is presented in Sect. 5, as is the potential for further research.

2 Literature Survey

It's critical to understand prior research on this topic in order to go forward in the proper way. Image processing and deep learning methods came to be extensively employed for reliable categorization of plant leaf diseases, which has been a major study topic. We review the most widely used strategies in the literature on the relevant topic in this study. Manually monitoring a vast field of crops is a time-intensive operation. It is vital to reduce the amount of human labor required for plant oversight. As a result, this is a prominent study topic that draws a large number of scholars. Several studies on plant diseases have been found in the literature.

Ding Jiang et al. [3], the symptoms of diseases over the leaf surface are extracted using the deep learning approach on tomato leaves just as testing objects. The network can conclude the division of each disease image afterwards through continual repeat learning. The underlying network model used in the experiment is Resnet-50. For comparison, the network's activation function was changed to Leaky-ReLU, and the first convolutional layer's kernel size was changed to 11*11. The training set accuracy rose by 0.6 to 98.3%, while the test accuracy increased by 2.3%.

Mohit Agarwal et al. [4], a Convolutional Neural Network based technique used for disease recognition and categorization. In this model, two completely associated layers come after three convolutional layers and three max pooling layers. The proposed model outperforms pre-trained models like VGG16, InceptionV3, and MobileNet in terms of experimental detection. The classification accuracy varies from 76 to 100% conditional on the class, and the suggested model's mean accuracy is 91.2% for the 9 diseases and 1 healthy class.

Marwan Adnan Jasim et al. [5], apply the convolutional neural network (CNN) in his suggested system to classify plant leaf diseases into 15 categories, comprising 12 classes for diseases found in distinct plants, such as bacteria, fungus, and others, and three classes for healthy leaves. As a result, he obtained remarkable training and testing accuracy, with 98.29% training accuracy and 98.029% testing accuracy for all data sets tested.

Akshay Kumar et al. [6], used the CNN architecture to identify disease in tomato leaves in his studies. To identify the type of leaves, he used the PlantVillage dataset of 14,903 photos of sick and healthy plant leaves to train a deep convolutional neural network. He achieved test accuracy for the trained model of 99.25%.

Huiqun Hong et al. [7], transfer learning is employed to minimize the bulk of the training data, as well as the time and computing expenses associated with deep learning. The feature extraction was carried out using five deep network structures: Resnet50, Xception, MobileNet, ShuffleNet, and Densenet121 Xception. The parameters and average accuracy of the five convolutional neural networks are varied. Densenet Xception has the greatest recognition accuracy at 97.10%, but the parameters are tiny. ShuffleNet has the best recognition accuracy at 83.68%, and the parameters are modest.

Prajwala TM et al. [8], uses a small version of the LeNet convolutional neural network model. Automatic feature extraction is used in neural network models to

help in the categorization of input images into disease classifications. The proposed system achieved an average accuracy of 94–95%, indicating the neural network approach's feasibility even under difficult circumstances.

Surampalli Ashok et al. [9], to diagnose Tomato Plant Leaf Disease, researchers propose integrating CNN with image processing methodologies based on picture segmentation, clustering, and open-source algorithms. The suggested technique is a hierarchical feature extraction CNN algorithm that outlines input picture pixel intensities and compares them to a training dataset image. As an image classifier approach, the comparison picture is categorized into diseased afflicted and normal leaves. Artificial neural networks, fuzzy logic, and hybrid methods may additionally be used. The efficacy of the suggested approach of tomato leaf disease recognition provided a huge accuracy figure of 98.12%, according to the research.

Nithish kannan E et al. [10], describes how to use Convolutional Neural Networks (CNNs), a category of deep neural network, to identify diseases in a tomato leaf. A pre-trained model (ResNet-50) is imported and tweaked according to his classification challenge using the transfer learning principle. Data augmentation has been implemented to improve the quality of the ResNet model. Taking all of this into account, a tomato leaf disease detection model employing deep - CNNs has been created using PyTorch. Data augmentation was used to enhance the data set to four times the original data, and the model was found to be 97% accurate.

Manpreet Kaur et al. [11], for recognition and regulation, a pre-trained Deep Learning Convolution Neural Network model is employed. The RESnet 101 is a model that is based on the DAG (Residual network). The characteristics were cited using the Convolution layer's filters, and the completely associated layer was trained for seven classes. To train the optimum model for dealing with the noisy channel, the Error Correcting Output code is utilized. For categorization, the ECOC is utilized. Accuracy, Specificity, Sensitivity, F-Score, and True Negative Rate are the parameters utilized to compute the classification results. The Accuracy of the trained model is 98.8%.

Ayesha Batool et al. [12], proposes an improved classification approach for detecting and classifying tomato leaf disease. The training dataset is made up of 450 photos. He extracted picture functions using many pre-trained Deep Neural Networks (DNN) models, then utilized a k-NN classifier to identify sick leaves. Several models are used to extract visual features, and kNN is used to classify them. AlexNet model has the greatest classification accuracy of 76.1% when compared to other models.

Robert G. de Luna et al. [13], employed a Convolutional Neural Network to determine whether tomato illnesses were present on the plants being monitored. The F-RCNN trained anomaly detection model has an accuracy of 80%, whereas the Transfer Learning illness recognition model has a 95.75% accuracy. The automatic picture capture system was tested in the field and found to be 91.67% accurate in identifying tomato plant leaf diseases.

3 Methodology

3.1 Dataset

Tomatoes are widely grown in India and are one of the most prevalent agricultural crops. Higher standards in order to diagnose and prevent crop diseases have been put in place as ecological agriculture has progressed. 2800 photos of the six major leaf diseases of tomatoes (Bacterial spot, Black mold, Early blight, Late blight, Mosaic virus, and Septoria spot) and one healthy class, were chosen for the experiment. In total, 2240 training photographs and 560 images for testing were distributed in a 4:1 ratio across the training and test sets [14]. To produce a better disease feature extraction impact, the picture is dominated by leaves, avoiding the cluttered quality backdrop. The information was gathered through Kaggle, and several examples are provided. Figure 1 shows samples of different tomato disease pictures and Table 1 shows the tabulation of the Dataset.

To assure the reasonableness of network input, the picture resolution is evenly scaled to 224×224 before training. To avoid the over-fitting phenomenon in the training process, random data augmentation was performed on the selected tomato leaf photos because of the minimum number of training samples. Also dropout function was used six times in order to avoid the overfitting problem.

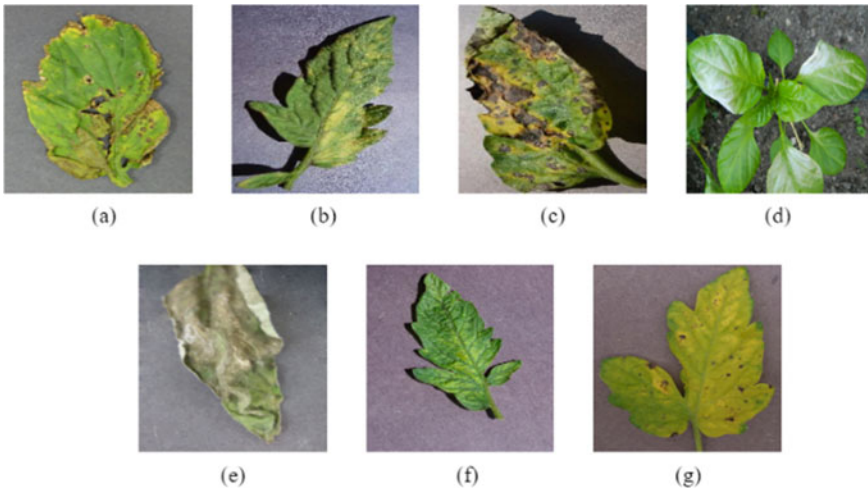


Fig. 1 Tomato leaf diseases namely, **a** Bacterial spot **b** Black mold **c** Early blight **d** Healthy **e** Late blight **f** Mosaic virus **g** Septoria spot

Table 1 Tabulation of the dataset

Class name	No. of images in training	No. of images in testing
Bacterial spot	320	80
Black mold	320	80
Early blight	320	80
Healthy	320	80
Late blight	320	80
Mosaic virus	320	80
Septoria spot	320	80
Total	2240	560

3.2 Data Preprocessing

The Deep CNN needs a large quantity of training data in order to deliver improved outcomes. In order to increase the model’s performance, image augmentation is frequently required when there is inadequate training data [15]. Image augmentation increases the number of pictures in the data set and reduces overfitting by including a few distorted photos in the training data. Image augmentation creates training images artificially using a variety of filtering processes or a mix of approaches, as an example, flipping the image, rotating it, blurring, relighting, and random cropping [16].

Due to the photos having different widths, they include a large amount of duplicated data, which is bad for the classification of tomato diseases. Image samples are normalized to 224 * 224 * 3 pixels, and the refined picture is deaveraged, as it where the mean value is removed from the G, R, and B grayscale values. Figure 2 shows the block diagram of the proposed work.

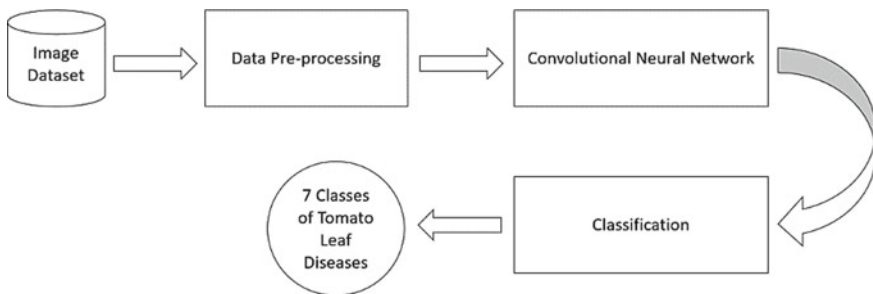


Fig. 2 Block diagram of the proposed methodology

3.3 Convolutional Neural Network (CNN)

Deep learning is a type of machine learning method that has several layers. The product of the previous layer is used as input for each subsequent layer. Unsupervised, supervised, or semi-supervised learning can all take place. Deep learning is a representation learning approach [17]. Optimizations are made by representation learning algorithms to determine the best practical method to represent data [18]. Because the attributes are extracted accordingly when the model is trained, deep learning eliminates the need for independent feature extraction and categorization. It's used in image processing, picture restoration, voice identification, natural language processing, and bioinformatics, to name a few.

In this research, CNN was used as the deep learning approach. CNN is outstanding in testing optical pictures and can smoothly detach the needed characteristics because of its multi-layered architecture, which can simply recognise and categorize objects with minimum preprocessing. The convolutional layer, pooling layer, activation function layer, and completely linked layer are the four primary layers.

3.3.1 Input Layer

The input layer has the photos as well as their pixel values.

3.3.2 Convolutional Layer

CNN is named after the convolution layer. The feature map of the input picture is extracted using a series of mathematical procedures in this layer [19]. A filter is used to minimize the size of the input picture. Starting at the uppermost left edge of the image, the filter is gradually moved. At each stage, the image's values are multiplied by the filter's values, and the result is added together. The supplied picture is used to produce a recent matrix along a reduced length.

3.3.3 Pooling Layer

After the convolution layer, the pooling layer is frequently enforced. This layer reduces the length of the output matrix derived from the convolution layer. Although other sizes of filters could be used in the pooling layer, the most common length is 2×2 . In this layer, you may utilize functions like maximum pooling, average pooling, and L2-norm pooling [20]. The maximum pooling filter with stride 2 was used in the research. The greatest value in each of the sub-windows is chosen and transferred to a new matrix to achieve maximum pooling.

3.3.4 Activation Layer

The activation function in artificial neural networks ensures that the input and output layers have a curvilinear connection. The performance of the network is harmed as a result of this. The activation function is responsible for non-linear network learning. There are other activation functions available, in addition to linear, sigmoid, and hyperbolic tangents, but CNN commonly uses the nonlinear ReLU (Rectified Linear Unit) activation function [21]. Values lower than zero are fixed to zero in ReLU, but those larger than zero are left alone.

3.3.5 Fully Connected Layer

Subsequently, after extensive iteration of the past levels, the data reached the ending layer of the CNN, which is the fully linked nodes [22]. The neurons of the two neighboring layers are precisely linked to the neurons in the completely associated network.

3.3.6 Softmax Layer

The performance of a network might be difficult to assess. In classification problems, it's common to conclude that the CNN uses a softmax function. After extracting values for 7 plant disease classes in the absolutely linked stage, a Softmax would be created considering them, allowing the class to be picked in all processes as well as based on the characteristics derived from the earlier layers, such as the pictures of plant diseases passed over [23]. The Softmax function is used to find the right disease class in this layer. Softmax function provides probabilities of all the classes but gives the highest probability of that class that represents the input image.

3.4 Training

The process of acquiring kernels in convolutional layers and weights in totally associated layers to decrease variations among output forecasting and given ground truth labels about a training dataset is known as training a network. We used 80% of the data for training in our research, so that the network that had been set up learned from obtaining characteristics coming from plant leaf diseases photos in the direction of learning from the indicated characteristics for every image to be well known on its own ground.

3.5 Testing

The testing dataset is used to offer the unbiased last design fit assessment through the training set of data. At this point, we utilize the groups that were trained within the former phase of CNN, and the characteristics were taken out through learning the network while the datasets were passed through plant leaf diseases on the present network, and we use 20% of the data for testing.

3.6 Experimental Settings

The dataset was used to test the suggested technique. It contains over 2800 photos of tomato leaf diseases from 7 distinct classifications. The model was implemented using Keras, a Python-based neural network API. Of the total of 2800 images, 560 photos were put separately for testing, and 2240 images were employed as training. Automatic data augmentation methods including random 20-degree rotation, horizontal flipping, and vertical and horizontal shifting of photos were utilized to enlarge the dataset. The Adam optimizer was used to optimize the loss function, which was category cross entropy. Adaptive Moment Estimation is a approach for optimising gradient descent algorithms. When working with vast complication with a lot of data or parameters, the approach is quite productive. It is effective and takes minimum memory. It's essentially a hybrid of the 'gradient descent with momentum' and the 'RMSP' algorithms. The Adam optimizer surpasses other optimizers by a huge margin in terms of providing an optimised gradient descent. The model was trained over 200 epochs with a batch size of 32. The learning rate is set at 0.001. Early halting has also been utilized to track validation loss and halt the training process if it exceeds a certain threshold. An Intel Core i3-1115G4U CPU was used for all of the testing.

4 Result Analysis

The accuracy of the suggested model was utilized to assess its performance. Over 200 training epochs, the best validation accuracy was 96.607%, while the maximum training accuracy was 97.19%. Validation accuracy would be determined to be 96% on average. This is a good indicator of how well the deep learning model classified the data. Figures 3 and 4 show a visual depiction of model convergence rate by plotting train and test accuracy and loss vs epochs of the best Pre-Trained model i.e. DenseNet201. Apart from that, the accuracy and loss graphs of the Pre-Trained models are shown in Figs. 5, 6, 7, 8, 9, 10, 11 and 12. Also Figs. 13 and 14 show the same for the proposed CNN model. The model seems to have stalled at 200 epochs,

and metrics had not risen remarkably over the previous 56 epochs. The model works well on the dataset, according to the results.

Contrary to big neural networks, that sometimes need a huge amount of computer support or the need of a graphics processing unit (GPU), the implementation technique necessitates very small scale hardware. Because there are limited layers with lower filter sizes and lesser train sizes, there are fewer training parameters. As a consequence, the model gives an easy and successful solution for the issue of plant diseases identification. The model allows approximate outcomes to standard state of the art approaches with less resource limitations and little data.

From the above Table 2 we observe that the Pre-Trained model DenseNet201 is giving us the highest accuracy. The accuracy and loss graph of the same is given below:

The graphs of the remaining Pre-Trained models are given below:

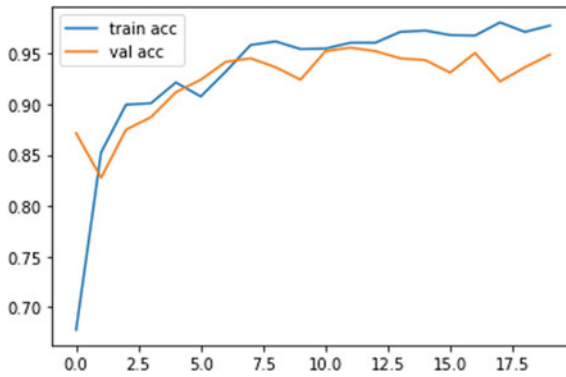


Fig. 3 DenseNet201 accuracy graph

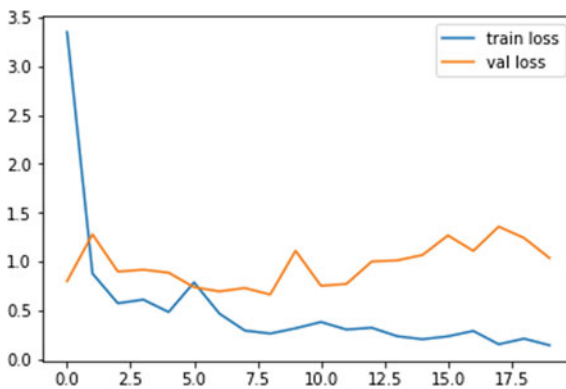


Fig. 4 DenseNet201 loss graph

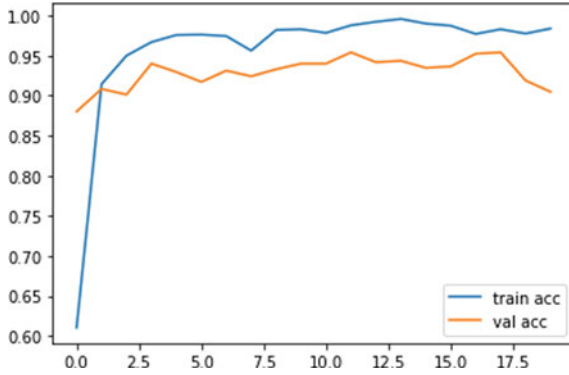


Fig. 5 InceptionV3 accuracy graph

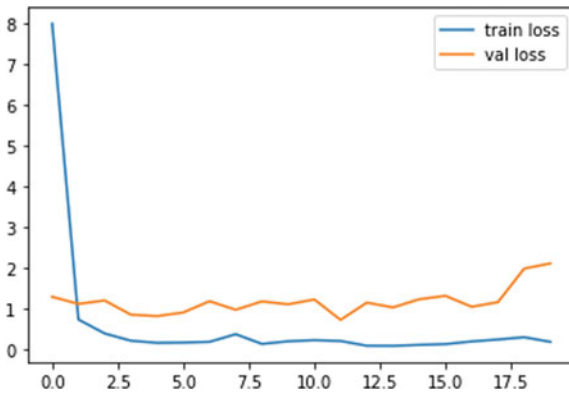


Fig. 6 InceptionV3 loss graph

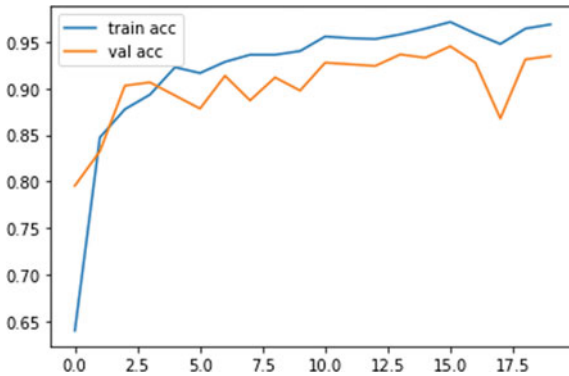


Fig. 7 ResNet152V2 accuracy graph

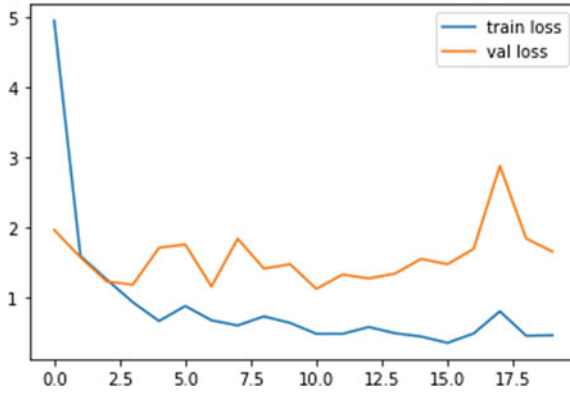


Fig. 8 ResNet152V2 loss graph

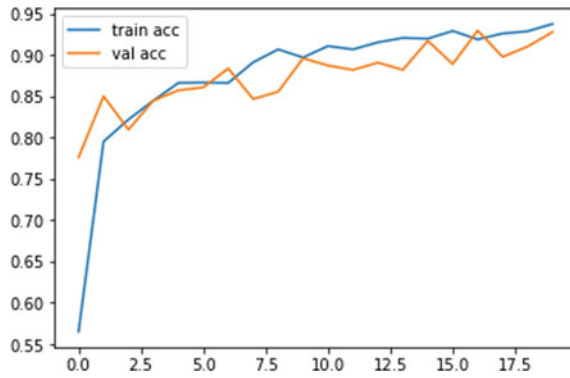


Fig. 9 NASNetMobile accuracy graph

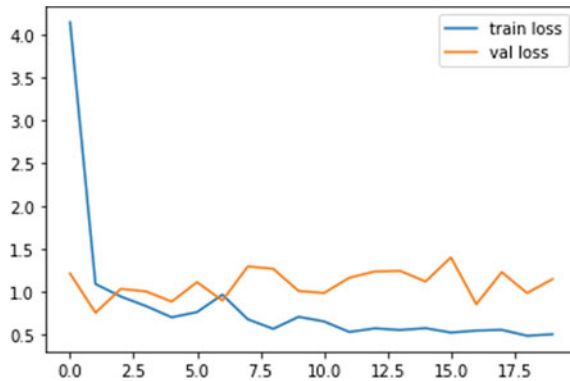


Fig. 10 NASNetMobile loss graph

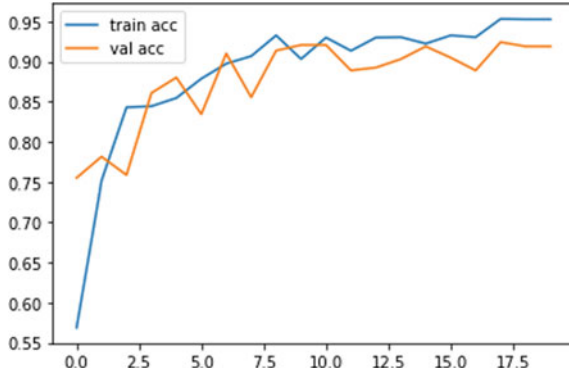


Fig. 11 InceptionResNetV2 accuracy graph

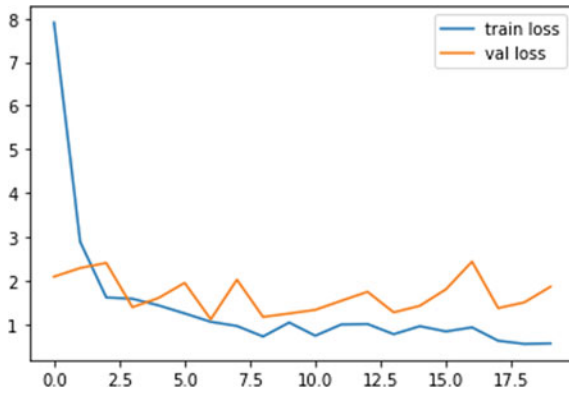


Fig. 12 InceptionResNetV2 loss graph

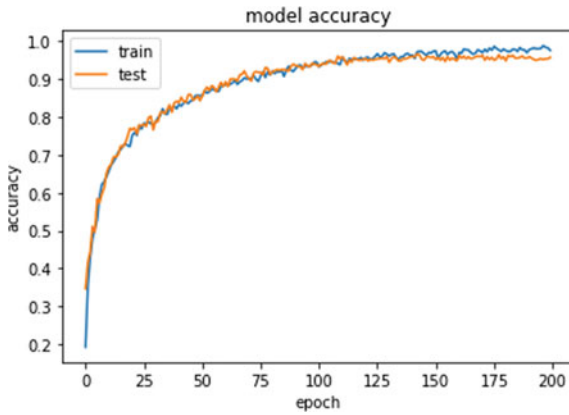


Fig. 13 CNN model accuracy graph

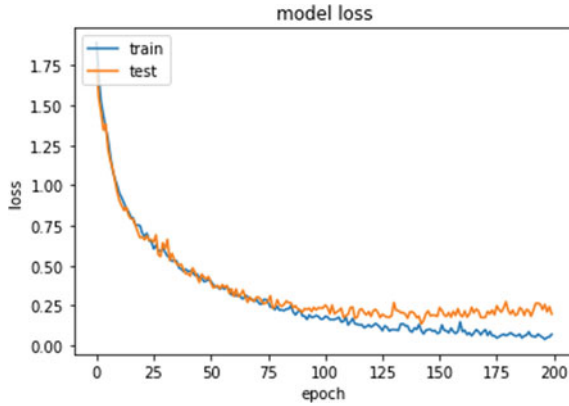


Fig. 14 CNN model loss graph

Table 2 Training & testing accuracies of the pre-trained model used

Pre-Trained model used	Training accuracy (%)	Testing accuracy (%)
DenseNet201	96.07	95.59
InceptionV3	98.79	95.41
ResNet152V2	95.80	93.65
NASNetMobile	91.87	92.95
InceptionResNetV2	95.31	92.42

Table 3 Training and testing accuracy of the CNN model used

Model used	Training accuracy (%)	Testing accuracy (%)
CNN	97.190	96.607

In order to improve the accuracy further, we have used CNN and built our own architecture. The number of epochs were taken as 200 and the batch size was 32. Six convolutional and six max pooling layers were used alternatively followed by six fully connected layers. The results obtained are shown above in Table 3.

5 Conclusion and Future Work

The agricultural industry continues to be one of the most significant sectors in India, with the bulk of the population relying on it. The identification of diseases in such crops is thus crucial for the economy’s rise. Tomatoes are a staple crop that is grown in vast numbers. As a result, the goal of this article is to find out and distinguish 6 distinct diseases in the tomato crop. To categorize tomato leaf diseases out of the

dataset, the suggested technique employs a convolutional neural network model. To identify tomato leaf diseases into six distinct classes, we used a simple convolutional neural network within a small number of layers. As part of a future study, other learning rates and optimizers might be utilized to experiment with the suggested model. It might also entail testing with the latest architectures in order to improve the model's efficiency against the training set. Therefore, the foregoing model can be drawn on as a conclusion gadget to assist and back farmers in recognising diseases that particularly affect tomato plants. The proposed technology could accurately identify leaf diseases while requiring little computational effort and gave testing accuracy of 96.607%.

References

1. Park H, Eun JS, Kim SH (2017) Image-based disease diagnosing and predicting of the crops through the deep learning mechanism. In: 2017 International Conference on Information and Communication Technology Convergence (ICTC). IEEE
2. Narvekar P, Patil SN (2015) Novel algorithm for grape leaf disease detection. *Int J Eng Res Gen Sci* 3(1):1240–1244
3. Jiang D et al (2020) A tomato leaf diseases classification method based on deep learning. In: 2020 Chinese Control and Decision Conference (CCDC). IEEE
4. Agarwal M et al (2020) ToLeD: tomato leaf disease detection using convolution neural network. *Procedia Comput Sci* 167: 293–301
5. Jasim MA, Al-Tuwaijari JM (2020) Plant leaf diseases detection and classification using image processing and deep learning techniques. In: 2020 International Conference on Computer Science and Software Engineering (CSASE). IEEE
6. Kumar A, Vani M (2019) Image based tomato leaf disease detection. In: 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT). IEEE
7. Hong H, Lin J, Huang F (2020) Tomato disease detection and classification by deep learning. In: 2020 International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE). IEEE
8. Prajwala TM et al (2018) Tomato leaf disease detection using convolutional neural networks. In: 2018 Eleventh International Conference on Contemporary Computing (IC3). IEEE
9. Ashok S et al (2020) Tomato leaf disease detection using deep learning techniques. In: 2020 5th International Conference on Communication and Electronics Systems (ICCES). IEEE
10. Kaushik M et al (2020) Tomato leaf disease detection using convolutional neural network with data augmentation. In: 2020 5th International Conference on Communication and Electronics Systems (ICCES). IEEE
11. Kaur M, Bhatia R (2019) Development of an improved tomato leaf disease detection and classification method. In: 2019 IEEE Conference on Information and Communication Technology. IEEE
12. Batool A et al (2020) Classification and identification of tomato leaf disease using deep neural network. In: 2020 International Conference on Engineering and Emerging Technologies (ICEET). IEEE
13. De Luna RG, Dadios EP, Bandala AA (2018) Automated image capturing system for deep learning-based tomato plant leaf disease detection and recognition. In: TENCON 2018–2018 IEEE Region 10 Conference. IEEE
14. Saleem MH, Potgieter J, Arif KM (2019) Plant disease detection and classification by deep learning. *Plants* 8(11):468

15. Geetharamani G, Pandian A (2019) Identification of plant leaf diseases using a nine-layer deep convolutional neural network. *Comput Electr Eng* 76:323–338
16. Zhang Y-D et al (2019) Image based fruit category classification by 13-layer deep convolutional neural network and data augmentation. *Multimed Tools Appl* 78(3):3613–3632
17. Guo W, Wang J, Wang S (2019) Deep multimodal representation learning: a survey. *IEEE Access* 7:63373–63394
18. Durmuş H, Güneş EO, Kırıcı M (2017) Disease detection on the leaves of the tomato plants by using deep learning. In: 2017 6th International Conference on Agro-Geoinformatics. IEEE
19. Tümen V, Söylemez ÖF, Ergen B (2017) Facial emotion recognition on a dataset using convolutional neural networks. In: 2017 International Artificial Intelligence and Data Processing Symposium (IDAP). IEEE
20. Giusti A et al (2013) Fast image scanning with deep max-pooling convolutional neural networks. In: 2013 IEEE International Conference on Image Processing. IEEE
21. Lin G, Shen W (2018) Research on convolutional neural networks based on improved Relu piecewise activation function. *Procedia Comput Sci* 131:977–984
22. Kui L et al (2018) Breast cancer classification based on fully-connected layer first convolutional neural networks. *IEEE Access* 6:23722–23732
23. Alabassy B, Safar M, El-Kharashi MW (2020) A high-accuracy implementation for softmax layer in deep neural networks. In: 2020 15th Design & Technology of Integrated Systems in Nanoscale Era (DTIS). IEEE

PDF Steganography Using Hybrid Crypto Encryption Technique



Sunil Kumar Patel  and Saravanan Chandran 

Abstract Recently, for all script and content writing, and transferring purposes, the Portable Document Format (PDF) has been widely used. This article presents PDF steganography based on a hybrid crypto encryption technique. The proposed method does not modify the portable document structure or content while concealing the secret data at any stage. Before the stego operation is performed, the confidential data undergo hybrid crypto encryption. Encryption consists of an advanced encryption standard 256-bit key and RSA encryption algorithm. As the document structure is not modified throughout this technique, it is not suspicious to technocrats during communication. Only the size of the PDF document increases as the size of the secret data increases. The human vision system cannot differentiate between the stego and a standard document, as the content and structure are unaltered.

Keywords Data hiding · AES and RSA algorithm · PDF steganography · Hybrid cryptography

1 Introduction

Over the past years, technology has been rapidly increasing for communicating information over the internet. Information hiding [1] comprises of many approaches such as copyright preservation for digital data, watermarking, information embedding, and steganography [2]. Steganography is a data concealing procedure that utilises images, audio files, video streams, and portable text documents as a medium to transfer secure data from one end to another.

Steganography is classified into different types based on the carrier object. It's a technique of hiding a secret image, file, message, audio, or video within another

S. K. Patel (✉) · S. Chandran

Department of Computer Science and Engineering, National Institute of Technology,
Durgapur 713209, India

e-mail: sunilpatel.bsb@gmail.com

S. Chandran

e-mail: cs@ieee.org

image, file, message, audio, or video. The word steganography is extracted from a work by Johannes Trithemus (1462–1516 AD) entitled *Steganographia*. It combines two Greek words; *steganos* meaning “concealed or covered,” and *graphy* meaning “drawing or writing.”

Steganography gains the advantage over cryptography alone because it intentionally conceals confidential information. Now, the personal information does not attract concentration to oneself as an entity of inspection. The encrypted messages in cryptography are directly visible to anyone. Although breaking the encrypted ciphertext is not an easy task in real-time, the visibility of encrypted confidential information arouses the interest of spammers. Therefore, cryptography is the mechanism of securing personal data, i.e., the contents only. At the same time, steganography covers both hiding the truth that confidential content is being communicated and the information content held or included in the communication.

The wide use of portable documents for writing and sharing the contents from one to another pull out the attention for PDF steganography. Based on the format and structure of the portable documents, text steganography [3, 4] is classified into three categories, as shown in Fig. 1. The three classifications are based on their (i) format, (ii) random and statistical generation, and (iii) linguistic method for performing text steganography.

The PDF layout: The use of PDF documents [5] is extensive, and it’s known as a portable document format. The foremost known PDF semantics [6] is divided into four segments as follows:

- Entity
- File structure
- Document structure
- Content streams

A data structure of a PDF document is collected from a compact set of primary types of data entities. The file structure of a PDF document decides how entities are stored, retrieved, and upgraded in a PDF file. The structure of the PDF file is free from the definition of the entity. The PDF file document structure identifies how the primary entity is used to characterize PDF document components. The components are the number of pages, style of fonts, annotations, etc. A PDF file

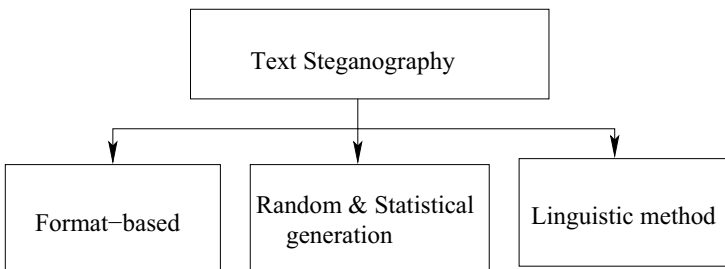


Fig. 1 Categories of text steganography

content stream includes instructions in a sequence that express the visual appearance of pages or additional symbolic and diagrammatic structures. These specifications, also constituted as an entity, are analytically unconnected from the primary entity that constitutes the document's structure.

This article proposes a PDF steganography technique using a hybrid crypto encryption algorithm. The random and statistical generation text steganography technique has been considered to conceal confidential data in portable documents. The proposed model is checked with various sizes of confidential input data for stego and its steganalysis is confirmed against visual detection. The stego is so subtle that someone explicitly looking for it is unlikely to notice the difference between original and stego PDF documents. It produces an efficient, robust, and imperceptible stego technique with PDF documents as a carrier.

The application area of the proposed model is to secure data transmission through PDF steganography. PDF documents have extensive use in recent times in medical domain, banking, educational institute, etc. The primary purpose of this model is to provide security to the data from the unintended recipient over a public communication channel. It offers efficient storing capacity and securely transmits data by various means of communication.

The rest of the research article is organised as follows. Section 2 illuminates previous related works on PDF document steganography. The proposed model is described in Sect. 3. Experimental discussion and analysis are explained in Sect. 4. Finally, in Sect. 5 conclusion is drawn for this research article.

2 Related Work

This work concentrates on the steganography technique [7] and PDF structure. The PDF structure described is concerned while designing the proposed PDF steganography technique. The state-of-the art techniques in PDF steganography till now have been presented in this section.

Liu and Tsai [8] proposed a steganographic technique in 2007 for concealing secret data in Microsoft word documents. It used the change tracking technique to hide confidential data in Microsoft word documents. It modified the structure of the document to disguise the secret data.

In 2008, Por and Delina [3] proposed an approach for text steganography. It used inter-word and inter-paragraph spacing for hiding confidential information in text documents. The technique had limitations; the stego data were lost and not recovered if spaces were removed. Li Lingjun et al., [9] proposed a steganography technique in PDF documents. It used the word shift technique for stego operation. The blind steganalysis method is used to detect the stego PDF documents. The concept of "environment equal" and "neighbor difference" was used for detecting the spaces to perform word shifts in PDF documents.

In 2016, Stephane and Rene [10] proposed a Chinese remainder theorem approach for PDF steganography. In this approach, ASCII code A0, which is unnecessary, was

released from the PDF documents. The released ASCII code A0 place was used for hiding confidential information in PDF documents. R. Vinothkanna [11] proposed a secure steganography technique used for various format for files. It used dual-RSA based cryptography algorithm for steganography to provide security.

In 2019, N.R. Zaynalov et al., [4] proposed three techniques for text steganography. The methods were based on a PowerPoint presentation of Microsoft, feature coding, and line shift operation. Behrooz Khosravi et al., [12] gave a new PDF steganography method using justified texts in the documents. Justifying was used for unformatted text that is not aligned along the left or right side of the document page. However, justified formatted the ragged edges of the text in the document. In this, private data were first compressed by Huffman coding and then the place to hide was identified in the *PDF* document.

In 2020, Katarzyna and Marek [13] proposed a steganography technique for PDF documents using a distributed approach. This technique used de-referenced entities and splitting secret or sharing algorithms for concealing the data in modified pages of the portable document. The page that conceals the secret data was hidden by manipulating the document's structure.

Thomas Sloan and Julio [14] proposed a technique that accurately detects the PDF steganography performed using the OpenPuff tool. The OpenPuff tool detected the stego documents if the structure or contents are modified in the documents. Istteffanny and Hassan [15] in 2020 proposed a technique to detect the steganography in PDF files which was used for malevolent activities.

3 Proposed Model

This section presents the PDF steganography proposed model with the hybrid encryption algorithm. This stego model conceals encrypted confidential information in PDF documents. The portable document is used as a carrier to transmit personal data from sender to receiver. This proposed model consists of two stages of security for transferring information. It uses cryptography encryption power to encrypt the plain text to ciphertext twice through a hybrid encryption algorithm. Encrypted ciphertext is converted into bit-stream to generate a fake image with stego into the PDF document. Figure 2 represents the complete flow diagram of the proposed model.

Cryptography encryption algorithms are categorised into symmetric key cryptography and asymmetric key cryptography. The symmetric key cryptography algorithm uses the same key for encryption and decryption. Asymmetric key cryptography uses pair of keys, a public key for encryption and a private key for decryption. The public key is known by all the parties in communication, as it's shared among the participating parties, and the private key is only known to the particular user.

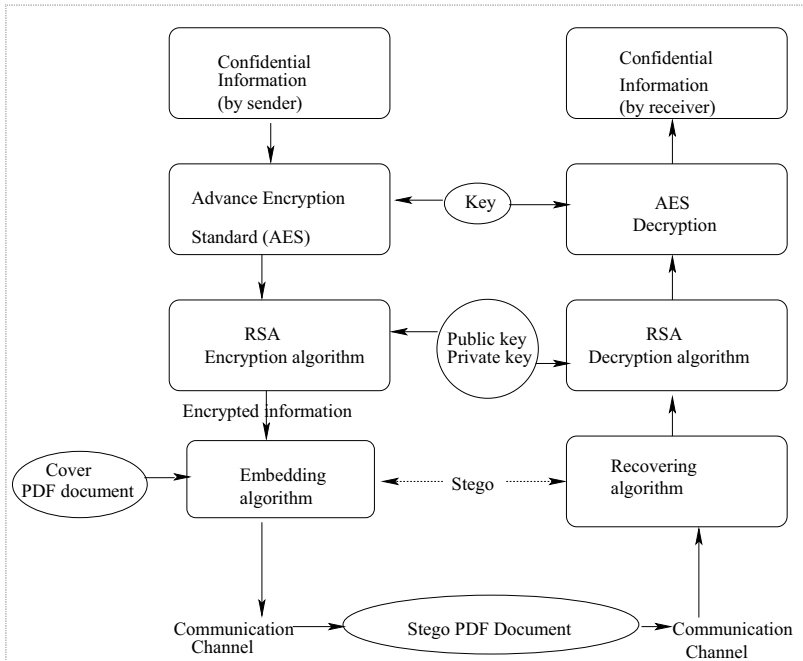


Fig. 2 Flow diagram of proposed text steganography model

3.1 Advanced Encryption Standard (AES)

National Institute of Standards and Technology (NIST) 2001 introduced AES [16, 17] to replace Data Encryption Standard (DES). It is a popular and widely used symmetric-key cryptography. It’s an upgraded version of the DES [18]. DES key is very small, so AES replaces it. Although triple DES is reliable, it requires higher computational potential and a prolonged execution process. The US government used the AES algorithm to secure crucial records and applied information encryption all around the globe [19, 20]. The features of AES are as follows:

- 128/192/256-bit keys and 128-bit data
- Symmetric key symmetric block cipher

The AES symmetric encryption key generation was given by Behrouz A. Forouzan [21]. AES goes through several rounds for different keys, which are multiple of 32; 10 rounds for 128-bit keys, 12 rounds for 192-bit keys, and 14 rounds for 256-bit keys. In the first round’s add-round key, the output goes through middle rounds before the final round. Four transmutations are performed during every one of these rounds; they are: (i) Sub-bytes, (ii) Shift-rows, (iii) Mix-columns, and (iv) Add round keys. In the final round of AES, mix-column transmutation is not performed.

3.2 *RSA (Rivest, Shamir, and Adleman) Encryption Technique*

RSA encryption technique is a public key cryptography encryption algorithm and is considered one of the most reliable ways of encryption. Rivest, Shamir, and Adleman invented it in 1978, and consequently, it was named the RSA encryption technique [22]. It utilises the idea of two keys. One is private, and another one is a public key. The ciphertext is generated from plain text using the public key, and to decrypt the ciphertext private key is used. RSA algorithm uses the concept of large prime number factorisation. The security and efficiency of the algorithms depend on the difficulty of factorisation of the selected prime numbers. RSA algorithm consists of four steps: selection of prime numbers, public and private key generation, encryption, and decryption.

The following steps explain the public and private key pair generation RSA encryption technique:

1. Two large random prime numbers p_1 and p_2 are chosen such that $p_1 = 6 p_2$
2. $i = (p_1 - 1) * (p_2 - 1)$
3. Encryption key m is determined randomly where $(1 < m < \varphi(i))$ and m and $\varphi(i)$ are coprime
4. Decryption key n is computed from the equation, $mn = 1 \text{ mod } \varphi(i)$ and $0 \leq n \leq i$
5. *Public key* = (m, i)
6. *Private key* = (n, i)

3.3 *Hybrid Encryption*

The confidential information was first encrypted through a hybrid crypto technique. Algorithm 1 shows the steps for encrypting the secret message using a hybrid encryption algorithm. The output of this encryption is further used as confidential information for performing the stego operation in PDF documents.

Algorithm 1: Hybrid Encryption

1 Input: Plain text file, two prime number p_1, p_2 and key for AES algorithm**2 Output:** Encrypted text file

```

3      begin
file1 = Plaintext file
message = Read(file1)
close(file1)
AES cipher = AES.Encrypt(message, AES key)
private, public = RSA.generate_keypair(p1, p2)
Encrypted message = RSA.encrypt(public, AES cipher)
file2 = Encrypted text file write(Encrypted message, file2)
close(file2)
4      end

```

3.4 Steganography

In this section, the encrypted secret data are embedded into the PDF document file. This process conceals the encrypted text data in a PDF file such that the file looks unchanged by visual steganalysis.

Steps for stego the encrypted secret data into a PDF file:

1. Take encrypted secret data as an input message.
2. Create a null image at a random location whose size is the square root of the length of the message.
3. Store the ASCII values of all characters of the message in an array.
4. Replace the pixel values of the image with the ASCII values.
5. Image stored in PDF at a random location such that HVS cannot detect it.
6. Save the embedded PDF file as a new stego PDF file.

Algorithm 2: Steganography

```

1 Input: Encrypted text data and PDF file.
2 Output: Embedded stego PDF file.
3 begin
  file1 = Encrypted text file
  message = read(file1) close(file1)
  l = length(message)
  size = sqrt(l)
  PDF = Input pdf file
  image = 2-D np array of dimension (size * size)
  ascii = an array
  counter = 0
  for i = 0 to l do
    4 |       ascii[i] = ascii value(message[i]);
  5 end for
  6 for i = 0 to size do
  7 |   for j = 0 to size do
  8 |   |   if counter < l then
  9 |   |   |   image[i][j] = ascii[counter];
  10 |   |   |   else
  11 |   |   |   |   break;
  12 |   |   |   |   end if
  13 |   |   end for
  14 |   |   counter = counter + 1;
  15 end for
16 output = insert_image(PDF, image)
  File2 = open(stego.pdf)
  save(File2)
17 end

```

Algorithm 2 represents the stego operation for the encrypted secret message. The private message is converted into a bit-stream and embedded into the PDF file. The human vision system cannot detect the embedded location of the secret data within the PDF document.

To retrieve the confidential message concealed in the PDF document with the above procedure, at first, the stego PDF record is identified and the destego operation is performed to retrieve the encrypted confidential information. Hybrid decryption performs over the encrypted confidential information retrieved from the destego process.

3.5 De-steganography

This step aims to retrieve the message hidden at a random location in the embedded PDF document file.

Steps for destego the encrypted secret data:

1. Read the embedded *PDF* file.
2. Retrieve the image written on the file in stego step.
3. Read the pixel values and convert them to corresponding character of the ASCII value.
4. Store the message in a character array.
5. Show the message.

Algorithm 3: Destego operation

Input: Stego PDF file.

Output: Encrypted text file.

begin

stego = StegoPDF file

s = string

l = length of the hidden message

image = extract_image(stego)

size = shape(image)

for (*i = 0 to size*) **do**

for (*j = 0 to size*) **do**

if *counter < l* **then**

s=s+char(image[i][j]);

else

break;

end if

end for

end for

file = open('embd.txt') write(file,s) close(file)

end

Algorithm 3 represents the destego procedure for the encrypted text image from the stego *PDF* file. Input to destego algorithm is stego portable document. The output of the algorithm is encrypted secret information used for decryption to retrieve the original secret data used for communication.

3.6 Decryption

Algorithm 4 takes encrypted text data, private key for RSA and AES decryption key as input for the decryption process. The output is secret information in original format used for communication.

Algorithm 4: Decryption

1 Input: Encrypted text file, private key for RSA and AES decryption key.

2 Output: Plain text file.

```

3      begin
file1 = open(embd.txt)
encrypted message = read(file1)
close(file1)
dcrypt = RSA.decrypt(RSA private key,encrypted message)
plain = AES.decrypt(dcrypt,AES key)
file2 = open('out.txt')
write(plain,file2)
close(file2)
4      end

```

4 Results and Discussion

This section presents the experimental results for the proposed model with some sample PDF documents discussed. The experiment is carried out on a personal computer with Ubuntu 20.04.4 LTS version having a 64-bits type operating system. The sample PDF is taken from different sources for testing and validating the model. Table 1 contains different amounts of data used for testing the model. It includes the number of characters used for testing, its size in bits and the time taken for stego and destego operations. The maximum capacity which can be stego in the PDF document maintaining imperceptibility is shown in Table 1.

Table 1 displays the variation of stego and destego time taken in milliseconds (msec). Table 1 contains the permissible amount of data stego within a single PDF document page. The number of characters stego inside the PDF maintains imperceptibility by HVS. The model is tested with the maximum amount of data that can be

Table 1 Different sizes of secret messages used for testing the model

No. of characters	Size (bits)	Stego time (msec)	Destego time (msec)
640	5,120	5.976	2.249
1,428	11,424	4.269	3.84
2,042	16,336	5.984	5.488
4,899	39,192	11.485	6.437
14,280	114,240	26.150	13.687
142,800	1,142,400	214.728	101.354
14,28,000	11,424,000	2,353.701	1,090.104

Table 2 The size of PDF after stego

No. of characters	Size of stego PDF (kB)
640	2.16
1,428	2.92
2,042	3.57
4,899	6.29
14,280	15.50
142,800	141.00
14,28,000	1360.00

stego effectively without detection. The original size of the PDF taken for stego operation is 1.15 KB. Table 2 illustrates the PDF size after the stego operation performed with different embedding sizes of the characters.

Comparative Analysis

Hybrid encryption has been applied on the secret data before stego to equip more security to stego data. The hybrid encryption consists of AES 256-bit keys with 128-bit data and RSA public and private key encryption. RSA encryption has two large random prime numbers; p_1 and p_2 are chosen such that $p_1 = 6 p_2$. Selecting the place for performing the stego is random in this model, and so this model is robust in the case of statistical steganalysis.

AES have larger size key of 128, 196, and 256 bits, so it is more secure as compared to DES. For example, DES with a 56-bit cipher key requires 2^{56} tests and needs t seconds to break the cipher. AES with 128-bit cipher key requires 2^{128} number of tests and requires $(2^{128} \times t)$ seconds to break the cipher. This would be almost impossible for real-time computation. AES also have two higher versions having 196 and 256 bits cipher keys. No differential and linear attacks are possible on AES yet.

RSA security is based on the larger factorisation of the modulus, which is infeasible to factor in a reasonable time. To secure the RSA encryption algorithm i should be more than 300 decimal digits, where $i = (p_1 - 1) * (p_2 - 1)$. It means the modulus must be at least 1024 bits. With the fastest computer available today, it is

impossible to factor an integer of 1024 bits size. It is secure till an efficient algorithm for factorisation has not been found.

The visual steganalysis cannot detect the existence of stego data in the proposed PDF steganography model. While in other methods [12, 23] the secret data were destroyed by retyping [8] and derived suspicions to technocrat. Katarzyna and Marek [13] concealed the private data by manipulating the document's structure, and the stego page in the document is hidden from the PDF file. N.R. Zaynalov et al., [4] used distance between words using spaces, paragraph marks, tabs, etc. If the document's structure is modified, then the secret data will be destroyed. The technique given by Stephane and Rene [10] for PDF steganography is that, the number of pages grows exponentially based on the amount of stego data.

The primary reason to apply this technique on a cover PDF document is to ensure that a steganographic approach has not modified the file. The PDF document structure and contents are not altered at any level in the proposed model. After embedding, only the size of the portable document increases with an increase in the size of secret data. The stego page, which contains confidential data, is not hidden from the PDF file while transferring over a public communication channel. The AES decryption and RSA public key are communicated with the involving trusted party through a secure medium. In the proposed steganographic model for communication, the secret message existence is not detected by visual steganalysis, so it does not derive suspicions to technocrats. The confidential data can only be destroyed when the complete PDF document is lost or destroyed over the communication.

The limitation of the proposed model is the steganographic portable document file size increases with the increase in the size of the secret data. 11,424,000 bits size of data can be hidden without any modification in the document structure and its contents has been achieved, so HVS will not detect any alteration. If the data size is increased, the image in which data is stego visible in the document structure will be detected by visual steganalysis. Hybrid encryption performs over confidential data before the stego operation to achieve higher security. But it arises, the challenge of communicating the private and public keys with the party involved in communication. This model application is limited only to PDF as a carrier for steganography.

5 Conclusion

PDF steganography has been presented in this article. PDF steganography has a broad application value because of the potential use of portable documents in the present digital world. There are two aspects of privacy. Firstly, the secret information is encrypted with a hybrid encryption technique having AES and RSA encryption algorithms. Secondly, the encrypted confidential data are stego into the PDF document, imperceptible by visual steganalysis. The method presented in this article does

not modify the content of the existing PDF. Only the size of the portable document increases with the increase in the size of the embedded message. The document contained are not modified in this method, so it does not look suspicious to technocrats. The secret message is lost only in case the PDF is destroyed or lost.

References

1. Fridrich J (2009) *Steganography in Digital Media: Principles, Algorithms, and Applications*. Cambridge University Press, Cambridge
2. Moulin P, O'Sullivan JA (2003) Information-theoretic analysis of information hiding. *IEEE Trans Inf Theory* 49(3):563–593
3. Por LY, Delina B (2008) Information hiding: a new approach in text steganography. In: WSEAS International Conference Proceedings. Mathematics and Computers in Science and Engineering, vol 7. World Scientific and Engineering Academy and Society
4. Zaynalov NR, Aliev SA, Muhamadiev AN, Qilichev D, Rahmatullaev IR (2019) Classification and ways of development of text steganography methods. *ISJ Theor Appl Sci* 10(78):228–232
5. Stevens D (2011) Malicious pdf documents explained. *IEEE Secur Priv* 9(1):80–82
6. Adobe Inc. acrobat developer resources. <https://opensource.adobe.com/dcacrobat-sdk-docs/index.html>. Accessed 08 Dec 2021
7. Patel SK, Saravanan C, Patel VK (2021) Cloud-based reversible dynamic secure steganography model for embedding pathological report in medical images. *Int J Comput Appl* 43(10):1002–1010
8. Liu T-Y, Tsai W-H (2007) A new steganographic method for data hiding in microsoft word documents by a change tracking technique. *IEEE Trans Inf Forensics Secur* 2(1):24–30
9. Li Lingjun L, Liusheng H, Wei Y, Xinxin Z, Zhenshan Y, Zhili C (2008) Detection of word shift steganography in pdf document. In: *Proceedings of the 4th International Conference on Security and Privacy in Communication Networks*, pp 1–8
10. Ekodeck SGR, Ndoundam R (2016) Pdf steganography based on Chinese remainder theorem. *J Inf Secur Appl* 29:1–15
11. Vinothkanna R (2019) A secure steganography creation algorithm for multiple file formats. *J Innov Image Process (JIIP)* 1(01):20–30
12. Khosravi B, Khosravi B, Khosravi B, Nazarkardeh K (2019) A new method for pdf steganography in justified texts. *J Inf Secur Appl* 45:61–70
13. Koptyra K, Ogiela MR (2020) Distributed steganography in pdf files—secrets hidden in modified pages. *Entropy* 22(6):600
14. Sloan T, Hernandez-Castro J (2018) Dismantling openpuff pdf steganography. *Digit Investig* 25:90–96
15. Kazemian Araujo II, Kazemian H et al (2020) Vulnerability exploitations using steganography in pdf files. *Int J Comput Netw Appl (IJCNA)* 7(1):10–18
16. Heron S (2009) Advanced encryption standard (aes). *Netw Secur* 2009(12):8–12
17. Osvik DA, Bos JW, Stefan D, Canright D (2010). Fast software AES encryption. In: Hong S, Iwata T (eds.) *Fast Software Encryption, FSE 2010. Lecture Notes in Computer Science*, vol 6147, pp 75–93. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-13858-4_5
18. Coppersmith D (1994) The data encryption standard (des) and its strength against attacks. *IBM J Res Dev* 38(3):243–250
19. Abd Elminaam DS, Abdual-Kader HM, Hadhoud MM (2010) Evaluating the performance of symmetric encryption algorithms. *Int J Netw Secur* 10(3):216–222
20. Albahar MA, Olawumi O, Haataja K, Toivanen P (2018) Novel hybrid encryption algorithm based on AES, RSA, and twofish for bluetooth encryption. Scientific Research Publishing, Inc.
21. Forouzan BA, Mukhopadhyay D (2015) *Cryptography and Network Security*, vol. 12. Mc Graw Hill Education (India) Private Limited New York, NY, USA

22. Rivest RL, Shamir A, Adleman L (1978) A method for obtaining digital signatures and public-key cryptosystems. *Commun ACM* 21(2):120–126
23. Lee I-S, Tsai W-H (2010) A new approach to covert communication via pdf files. *Signal Process* 90(2):557–565

An Efficient Classification Algorithm for Employee Well-Being Prediction Using Deep Learning



S. Sunandha Shri and M. Ezhilarasan

Abstract The impact of COVID-19 has changed the way work is being done especially in the IT sector. The emergence of work from home as an option has resulted in the evolution of hybrid work culture going forward as the world is moving towards endemic. On these circumstances there has been drastic change in work pattern of employees which clearly impacted the efficiency levels and their wellbeing (both physical and mental). It has also become imperative for the employers to track the efficiency of employees during their working hours in order to ensure maximum productivity in hybrid working model. This paper proposes a system that can detect and track the employee efficiency through facial landmarks by assessing the parameters like drowsiness and stress using deep learning techniques and hybridization of classification algorithms.

Keywords Employee efficiency · Drowsiness · Stress detection · CCNN · HCCNN · Feature extraction · Hybridization algorithm · Random forest · Radial basis function

1 Introduction

The way offices have operated and the employees have worked may not be going back to the pre-covid times as the COVID-19 pandemic is approaching endemic. Employees especially those belonging to the white-collar sector have accustomed themselves to work remotely and manage time between office and home. As per a PwC report surveyed with a sample size of 1200 workers, more than 55% of them preferred working remotely three days a week. Among the 133 executives surveyed, 68% of them said, employees may be present in office at least three days a week in

S. Sunandha Shri (✉) · M. Ezhilarasan
Department of Information Technology, Puducherry Technological University,
Puducherry 605014, India
e-mail: sunandharish@pec.edu

M. Ezhilarasan
e-mail: mrezhil@ptuuniv.edu.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
P. P. Joby et al. (eds.), *IoT Based Control Networks and Intelligent Systems*,
Lecture Notes in Networks and Systems 528,
https://doi.org/10.1007/978-981-19-5845-8_33

467

order to sustain the company's culture as they thought remote working will have an impact on the culture of the organization. In a survey conducted by Gartner in 2020 involving 127 leaders of various companies found that only 30% of the leaders were concerned about maintaining the corporate culture in hybrid model of working.

McKinsey conducted a survey involving 100 executives from different industries and geographies on the post-pandemic future of work. The survey report suggests that nine out of ten organizations will prefer a combined remote and in-office working model. The survey also suggests that during the pandemic many organizations has seen a phenomenal increase in individual and team productivity along with employee engagement due to an increase in focus and energy. There is also a rise in customer satisfaction. But not all organizations have experienced this positive result. For example, taking individual productivity into consideration, 58% of executives felt that there is a increase in productivity, where as a third of them were of the opinion that the productivity hasn't changed. Companies that are lagging accounting for the remaining 10% were of the opinion that the individual productivity has declined during the pandemic. There is a correlation between individual and team productivity. Executives who reported an increase in the individual productivity are more likely to report an increase in team productivity and vice-versa. Given this background, it has become extremely important to have a system/methodology in place to track the efficiency of the employee and one important parameter in estimating the efficiency is fatigue. Using the existing models of drowsiness detection in drivers [1], Haar Cascade and Cascaded Convolution Neural Network is proposed to predict the efficiency of the employees using facial landmarks and deep learning. Deep Learning algorithms will be a key in employee fatigue or drowsiness and stress detection as there are multiple works conducted using various deep learning techniques especially the hybridization of classification algorithm in order to deduce lacking of employee performance due to their well-being as the system is trained with existing datasets and real-time inputs [6].

2 Background and Related Work

Extensive research has been done for calculating the drowsiness, stress, emotions parameters in order to increase the performance of a person.

2.1 Detection of Facial Motions Through Deep Learning Techniques

Based on this study, the system will be trained and evaluated on the largest publicly available dataset UTA-RLDD dataset [1]. The samples of this dataset consist of numerous videos taken from different participants across various demographics and

are classified into three classes namely: Alert, Low Vigilance and Drowsy. The design of the model discussed in this work perform binary classification for drowsy and alert samples [5]. As a result, the videos used from this dataset represent alert and drowsy behaviours. Videos in this dataset were captured in various light conditions: very low light, normal light, and very bright light. The participants of the dataset were asked to record three different types of sleepy videos for better variety of sleepy faces [7] closed eyes with head leaning downward and upward, closed eyes and open eyes with head straight, and yawning.

2.2 Facial Motion Analysis Using Sensor

This work discusses an extremely successful and efficient approach to recognize and localize the eyes of the person and motions of the mouth from video streams [2]. Deep learning algorithms like Spiking Neural Networks (SNN) and Convolved Neural Networks (CNN) serves as a base to drowsiness [8] of the employees by tweaking the behavioral pattern of the person to that of the employees. In this study a neuromorphic vision sensor is introduced which is a type of alternative vision sensor signal acquisition-based paradigm [2]. This is inspired by the human visual design, and helps to suppress redundancy and low latency through asynchronous and temporal level crossing sampling. This is an alternative approach to the classical spatially dense sampling at fixed frequency implemented in traditional frame-based cameras [3].

2.3 Expression Detection Using Various Deep Learning Classifiers

Because of the historic strategies supported clinical identification, there is a need for associate degree automated detection system of the depression. An absolute particular audio- primarily based method to routinely find out depression mistreatment hybrid method. This model combines CNN and SVM [17], wherein SVM takes truly linked connected layers in CNN. SVM Classifiers have a high level of accuracy and can anticipate events quickly. The SVM algorithm's purpose is to find the optimum line or decision boundary for categorizing n-dimensional space into classes so that additional data points can be readily placed in the correct category in the future [9]. The choice squares measures routinely extracted CNN and consequently the classification is completed mistreatment SVM classifier. This method is goes to be evaluated mistreatment DAIC-WOZ dataset supplied via way of means AVEC 2016 depression evaluation sub-challenge. Experimental consequences confirmed that hybrid model achieved associate degree accuracy of that beats CNN method [15].

2.4 Modelling of Depression Technique from Social Network

This study aims to categorize users with depression through more than one instance of learning by expanding a deep learning model, that may research from user-level labels to identify post-level labels [11]. Users with depression are categorized by combining each opportunity of posts label category which generates temporal posting profiles [17]. Diagnosis of disturbance may be a tough task which could solely be done by health specialist for his/her disorder to be properly discovered, patients got to recall however they felt and what happened to them the previous essential quantity, that facilitates the clinical acquire comprehensive background info. Statistics of Facebook users is collected from certain year to examine the depression of people in social network and build a predictive model using more than one instance of learning neural networks to detect signs and symptoms of depression [16]. This was crucial to expand their method with the help of some labelled bags instead of requiring all the labels of the times used [15]. The depression level of a person will be further classified by multiple instances learning to identify user label levels to post label category [14].

2.5 Limitations of Existing Work

The existing works has high computational complexity and require specialized hardware which prevent their usage in real-time [4]. Performance of few methods suffers with data shortage and accuracy while maximum of the works is centered only on drowsiness detection in driving [5]. There is not more examination that empathies employee performance and well-being especially concerned to the personnel whose mode of work is via desktop/Laptop (E.g., IT personnel). Given this background, the onset of COVID-19 has impacted the work habits of the employees as well. With the evolution of hybrid work culture, there are not tons of works present currently that takes a check on the efficiency and wellbeing of the employees.

3 Employee Efficiency Prediction System (EEPS)

3.1 EEPS Architecture Design

3.2 Face Detection Using Cascaded Convolution Neural Network

Face detection, which is entirely dependent on computer vision, is one of the most important technologies for drowsiness detection [8]. A drowsiness detection

model not only requires high approximation, but also additional excessive speed. Deploying deep learning models, specifically the convolutional neural network technique greatly enhances the accuracy of picture identification [12]. A CNN model (ConvNet/CNN) is an amalgamation of Deep Learning rules that takes an image as input and give importance (learnable weights and biases) to various aspects/objects in the image, allowing them to be distinguished from one another. The major significance of cascaded convolutional neural network is that will deal with addressing problems and hence the performance is increased. It will be more against false positives. Hence the exact face of the employee will be detected and false prediction will get removed. In a ConvNet, the pre-processing is required to splendid down in exam to different kind of classification algorithms. While in old techniques filters are hand-engineered, ConvNets with adequate training, have the potential to research those filters/characteristics [2]. As per Fig. 1, once the image is obtained from the live camera feed, Eye Aspect Ratio (EAR) is inferred to deduce whether the eye is open or not. Based on the EAR [5], the system will be able to detect whether the eyes of the employee are in an open state or closed state. It will also check if an employee face is present or not in front of the camera. There are two major tasks involved here. The first venture is used to identify whether a valid face is present or not (face/no face), which is a classification process. To attain this, we use EEPS loss function for training process. For any sample x_i , EEPS loss function is given by,

$$p_i^1 = -(a_i^1 \log(c_i) + (1 - a_i^1)(1 - \log(c_i))) \tag{1}$$

where c_i is outcome of the network, a_i^1 is genuine label of x_i (With face/Without face). The second goal is to forecast the facial region box boundary coordinates. Because this is a regression problem, Euclidean loss functions are used in the training procedure. The EEPS loss function is,

$$p_i^2 = \|c_i - a_i^2\|_2^2 \tag{2}$$

where c_i is the coordinates of facial region and a_i^2 is the exact coordinate of the face area in the image.

The face of the employee can be accurately retrieved by training the Cascaded Convolutional Neural Network as in Fig. 2, which gives a stable face picture to apply the algorithm to be developed [8]. On the other side, all the system is interconnected hence the cascading failure in one system can lead to a failure of the other therefore will increase the error rate and reduction in capacity.

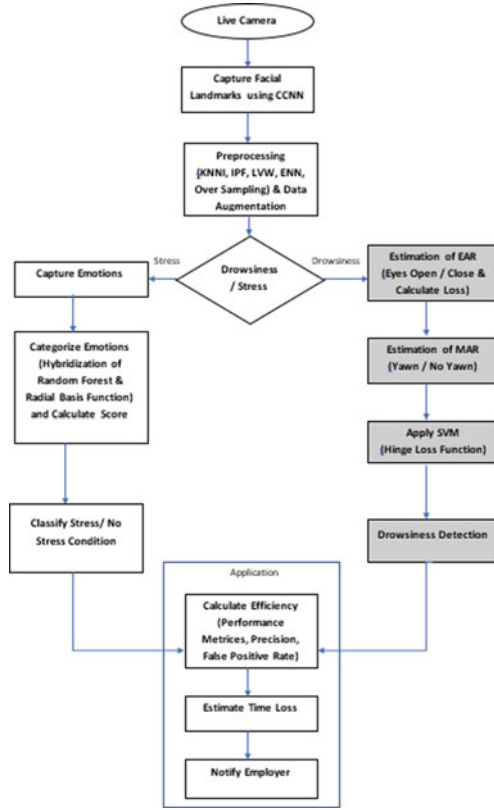


Fig. 1 EEPS architecture design

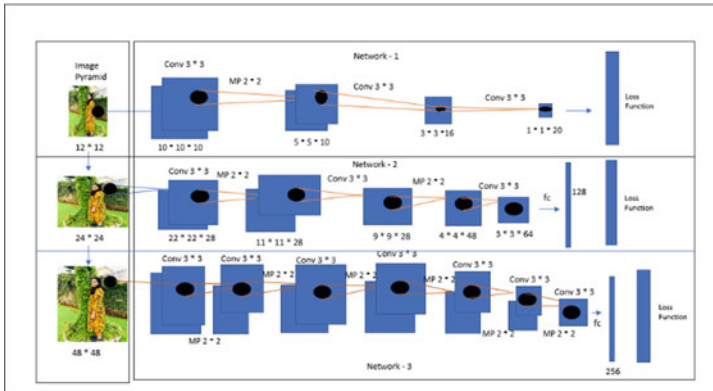


Fig. 2 Cascaded convolution neural network for EEPS design

3.3 Face Recognition Using Haar Cascaded Convolution Neural Network

Face recognition is particularly useful for identifying a person using a biometric method that relies on an image of their face. Biological characteristics are used to identify a person. People can be easily recognized by simply looking at them, however human eyes have a limited focus period. As a result, automatic facial recognition algorithms have been developed. [10]. Face recognition entails certain basic processes such as detecting faces by default, as well as confirming personnel from each image or video. Despite the fact that facial recognition has received a lot of attention, there are still a lot of difficult scenarios to overcome, such as:

- Misalignment
- Pose Variation
- Luminescence Variation
- Expression Variation

Face landmark detection is a computer vision task where we want to detect and track key points from a human face [3]. This task applies to many problems. In the EEPS model proposed in this research work, Haar cascade classifier is used to classify the face of the employee. It detects the face of the person from multiple system trainings using high dimensional pictures and number of training stages which would yield a better system result. The library used is TensorFlow which runs faster and produce accurate results. The maximum training period is three days. The more of the training period the less amount of loss with better accuracy. The system's accuracy will be tested by identifying three persons (most of the times in different places) to see how light intensity impacts the system's performance. The accuracy can be checked by using a confusion matrix. The following is the basis for the calculation:

$$Accuracy = ((Sum\ of\ True\ Negative\ and\ True\ Positive)/Total) \times 100 \quad (3)$$

3.4 Head Position Analysis

The position was initially estimated using KNN algorithm which analyses with many numbers of datasets and finally determine the image using feature extraction. As discussed above, in this proposed EEPS model, Haar cascaded classifier is used to detect the facial features such as eyes, mouth and nose. Dataset containing images of those will be referred. For Instance, if the classifier detects the features, the corresponding binary value will be denoted as 1 or 0 otherwise. The landmark feature uses the 3D coordinates of the 68 facial coordinates as in Fig. 3 is detected with convolutional network. The facial network was surrounded with bounding box which was detected using Dlib library. Nose length will be estimated by the distance between



Fig. 3 Facial coordinates

nose tip and top of nose. Extracted features will be trained using five classifiers such as Radial Basis Function, Support Vector Machine, Logistic regression, and Random Forest [18]. The input data will be divided into training and testing phase where videos will be recorded. All the recorded videos in one phase will be assigned to training stage and videos from other phase will be assigned to testing phase. Training phase will be further dived into training and validation process and finally the position will be determined.

3.5 Estimation of Eye Aspect Ratio

The status of the employee’s eyes is used to monitor tiredness. Getting the corner and form of the eyes is key. CPR (Cascaded Pose Regression) is a regression approach for estimating an object’s pose [8]. Feature points or landmarks, in particular, can be used to represent a face position. As a result, the CPR algorithm is employed to obtain face land-marks in order to assess the employee’s stance. The EAR is calculated using the following formula:

$$EAR = (\|P2 - P6\| + \|P3 - P5\|)/2\|P1 - P4\| \tag{4}$$

where P1, P2, P3, P4, P5, P6 are the coordinates of eye landmarks. The EAR is well above 0.2 when the employee’s eyes are open.

3.6 Estimation of Mouth Aspect Ratio

Yawning may be a symptom of drowsiness which is indicated visually when the mouth is open. Research on yawn detection focuses only on the size of the mouth when its opening [11]. The mouth opening is identified as yawn when the ratio is beyond a certain threshold. The lips color is one in every of a type for everybody, and manner to various lights conditions, it's critical to affirm that the threshold value is customized to the changes. The width of the mouth is measured in consecutive frames, and yawn is detected when the mouth is continuously commencing widely over a wide variety of times and is detected through analyzing the movement in the mouth area additionally detected supported a horizontal pro-record projection. When representing the face with 68-(x, y) coordinates, the mouth is represented through a set of 20-(x, y) coordinates. Similar to EAR calculation, we have used coordinates 62, 64, 66, and 68 to calculate the space among the lips to calculate the mouth aspect ratio (MAR) using the Eq. (5). The horizontal and vertical distance of the mouth is calculated as:

$$MAR = (50 - 60) + (51 - 59) + (53 - 57) + (54 - 56) / (2 * (49 - 55)) \quad (5)$$

where coordinates, 49–68 represents the locations of 2D facial landmark. With the help of these numbers and the result of mouth state, we can calculate the count of yawn over a period of time, which in turn is a very important feature in assessing the employee drowsiness in our EEPs model.

3.7 Drowsiness Detection Assessment

The drowsiness is detected and assessed real-time while the employee is front of his/her work machine. The live video will be captured by a camera in front of the employee when the system is turned on as shown in Fig. 1. Every frame of every video will be scrutinized [8]. To begin in the proposed system, the CCNN is used to detect faces, and the HCCNN is utilized to recognize employee faces. Furthermore, the Dlib toolkit is used to create facial landmarks. The EAR will be calculated later. According to the eye's landmarks and MAR will be calculated for determining the mouth coordinates which determines yawning state. An SVM classifier is trained in the offline mode such that when the live image is fed into the classifier it is able to detect whether the eyes are open or closed through EAR input, and yawning is also assessed for tiredness using MAR input.

An incremental rating score together with an alarm which will be triggered based totally on the scenarios of open and closed eyes. When the eyes are open, and there may be no yawn the system detects that the employee is active. When the eyes are closed and yawn is detected, the machine triggers the rating score and when the score

crosses certain limit, an alarm is triggered and vibrates continuously until the score falls lower back beneath the certain limit.

3.8 Stress Detection in Employees

The employee face can be broadly categorized as Stress and No Stress. The emotional states given below in Table 1. can be either categorized as Stress or No Stress. In the EEPS model of Stress detection in IT employees, initially the data cleaning process is done and a number of common approaches for ensuring data consistency and validity of the response of each individual. The hybridization of Random Forest and Radial Basis classification algorithm is involved. Radial Basis Function have good learning ability from dataset and ignore all the background noise from the video. It is more efficient as it uses gaussian kernel for separation patterns. Random Forest on the other side is used for both classification and regression where it can handle a huge number of datasets and produce high level accuracy the system [18]. Hence on the hybridization of both Radial Basis and Random Forest the EEPS model eliminates all the error causing inputs and used to deal with detection problem by learning from training dataset and produce the overall best results. Finally, the process classifies the stress and no stress condition by categorizing the employee emotions. This study also proposes to detect the stress amongst the employees by capturing different emotions through facial landmarks. The emotional states given in Table 1. can be either categorized as Stress or No Stress.

When the machine first starts up, a camera in front of the employee captures live video, which is then processed frame by frame [8]. First, the Pre-processing process is done wherein all the negative and poor pix images, irrelevant noise and dimensions related to images were eliminated and the CCNN is used to detect face of employee where HCCNN also used to recognize the face of the employee. Then, facial landmarks are determined using Dlib toolkit. Then, the emotions are classified. Each frame analyzed by this system is saved as a separate image in the root directory folder wherein the code is running. Later on, this function creates a replica of the

Table 1 Different types of emotion and stress detection

Stress	Category
Neutral	No Stress
Joy	No Stress
Surprise	No Stress
Anger	Stress
Sadness	Stress
Fear	Stress
Disgust	Stress

original video by producing a rectangular box around the face and showing live emotions within this video. Further, create a Pandas Data Frame from these values that are obtained and plot a graph with the data frame using matplotlib against instances. We can further examine this data frame by taking every individual emotion value that was analyzed by the model and finding which sentiment was dominant across the whole video. Finally, the Stress is estimated based on the type of emotion detected by the system and the classification process is done with the hybridization algorithm proposed in this system.

A threshold limit (say 10) for Stress during a particular duration (say per day) can be found and upon breaching the threshold limit of the Stress, the HR and the Reporting Manager of the employee can be notified to safeguard the mental health of the employee. This module will be further expanded after the completion of Drowsiness Detection module.

3.9 Notification to Employer and DCOUNT Functionality

A counter is set to determine how many times the employee falls drowsy based on the trigger of score above 15. A threshold limit can be set by the employer and once the counter exceeds the threshold limit, a notification as will be sent to the employer to report the same at the end of the day. This will be a Key Performance Indicator (KPI) for the employer to decide on the employee performance over a time period viz., day, week, month, year. A functionality 'D Count' has also been added in the system as shown in Fig. 4. to let the employee know his/her count of drowsiness which enables the employee to keep track of his/her drowsiness and take necessary actions in order to not exceed the threshold limit set by the employer. This will also act as an indicator to the employee's productivity.

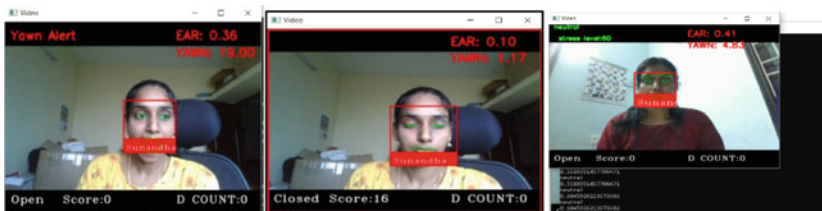


Fig. 4 Drowsiness alert and stress value estimation

Table 2 Employee efficiency illustration

S No.	Employee name	Employee score (%)	Category
1	Sunandha	90–100	Exceeds expectation
2	Harish	80–89	Good
3	Siva	70–79	Average
4	Shri	60–69	Needs improvement
5	Kumaran	<60	Bad

3.10 Estimation of Time Loss and Efficiency Calculation

The system aims to record the loss in time whenever the falls drowsy based on the alarm trigger. The duration of the alarm for each instance may be recorded and can be estimated as time loss which will be very useful in the proposed efficiency calculation. Efficiency can be determined based on the expected working hours set by the employer without any distractions. The actual working hours can be calculated by deducting the time loss that occurred due to drowsiness of the employee and the duration he/she is away from the workstation even though the workstation is logged on. The standard efficiency percentage is given by the formula,

$$\text{Efficiency}(\%) = [(\text{Actual Input})/(\text{Expected Output})] * 100 \quad (6)$$

where,

$$\text{Actual Input} = \text{Expected Output} - \text{Time Loss} \quad (7)$$

Time Loss can be estimated based on the time lost whenever the employee is drowsy. Expected Output is the output time limit of work expected by the employer without any distractions. Based on the efficiency scores the employees are categorized as Table 2:

4 Results and Discussion

The experimental set up includes three major modules such as 1) Drowsiness Detection 2) Stress Detection 3) Overall efficiency calculation. In the initial step the author will fix a camera in each working system and the employee will action the symptoms of drowsiness by means of closing of eyes and yawning and will provide a symptom of stress by facial expression. When the system detects drowsiness and stress the alert will be provided for the employee as shown in Fig. 4.

The recorded training and validation accuracy for each instance of epoch are plotted for the proposed system. The validation accuracies obtained in each instance of epoch for a particular run is averaged to record one iteration of the derived

value. Likewise, ten iterations are performed against each of the baseline value. After performing ten iterations the average derived value is calculated.

$$Derived\ Value = \bar{x} \sum_{i=1}^n Average\ Validation\ Accuracy \tag{8}$$

where n is the number of iterations.

The derived value obtained for the proposed model stood at 97.37% whereas that of the baseline model is at 94.80%.

Figures 5 and 6 shows the results evaluated and are inferred in Table 3.



Fig. 5 Training and validation accuracy

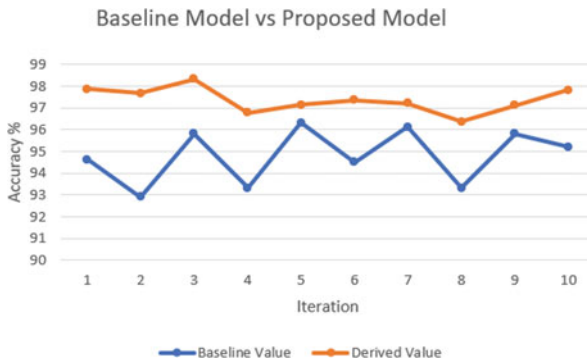


Fig. 6 Baseline vs proposed model comparison accuracy

Table 3 Baseline vs proposed model

Iteration	Baseline value	Derived value
1	94.63	97.87
2	92.90	97.68
3	95.81	98.32
4	93.33	96.79
5	96.33	97.13
6	94.50	97.37
7	96.13	97.22
8	93.33	96.37
9	95.82	97.11
10	95.21	97.83
Average	94.80	97.37

5 Conclusion

This study delves into the aspects of the that impacts employee efficiency due to the changing work culture and habits especially after the onslaught of the COVID-19 pandemic with the introduction of hybrid environment. In this study, we propose a methodology to track the drowsiness and stress levels of employees during work-time by three experiments namely drowsiness, stress and overall efficiency detection which in turn can impact the quality of work. We have developed and evaluated a Cascaded Convolution and Haar Cascade Algorithm which is used to classify employees' states using facial landmarks, with an emphasis on the selection of appropriate characteristics. Drowsiness detection from Kaggle dataset was recorded and reviewed for this purpose. The recorded eye closure signal was used to infer a large number of head movement blink features and expressions, which served as the foundation for model development. The selection of relevant features was a crucial part of the SVM Hinge-based classification.

The limitation of this study is that, this captures only the drowsiness, stress and emotional levels of the employees that can impact employee efficiency but doesn't track the actual quality of work done. The hybrid model at times makes the overall system down as the drowsiness module focuses on the SVM classifier [13]. Hence the hybrid model can be implemented for both the modules to make more accurate performance and faster results.

References

1. Tamanani R et al (2021) Estimation of driver vigilance status using real-time facial expression and deep learning. *IEEE Sens Lett* 5(5):3–6. <https://doi.org/10.1109/LESENS.2021.3070419>

2. Chen G et al (2020) EDDD: event-based drowsiness driving detection through facial motion analysis with neuromorphic vision sensor. *IEEE Sens J* 20(11):6170–6181. <https://doi.org/10.1109/JSEN.2020.2973049>
3. Pandey NN, Muppalaneni NB (2021) Real-time drowsiness identification based on eye state analysis. In: *Proceedings of the International Conference on Artificial Intelligence and Smart Systems ICAIS 2021*, pp 1182–1187. <https://doi.org/10.1109/ICAIS50930.2021.9395975>
4. Vijay M et al (2020) Real-time driver drowsiness detection using facial action units. In: *Proceedings of the International Conference on Pattern Recognition*, pp 10113–10119. <https://doi.org/10.1109/ICPR48806.2021.9412288>.
5. Girish I et al (2020) Driver fatigue detection. In: *2020 IEEE 17th India Council International Conference INDICON 2020*, pp 9–14. <https://doi.org/10.1109/INDICON49873.2020.9342456>
6. Singh J (2020) Learning based driver drowsiness detection model. In: *Proceedings of the 3rd International Conference on Intelligent Sustainable Systems ICISS 2020*, pp 698–701. <https://doi.org/10.1109/ICISS49785.2020.9316131>
7. Dreisig M et al (2020) Driver drowsiness classification based on eye blink and head movement features using the k-NN algorithm. In: *2020 IEEE Symposium Series on Computational Intelligence SSCI 2020*, pp 889–896. <https://doi.org/10.1109/SSCI47803.2020.9308133>
8. You F et al (2019) A real-time driving drowsiness detection algorithm with individual differences consideration. *IEEE Access* 7:179396–179408. <https://doi.org/10.1109/ACCESS.2019.2958667>
9. Tipprasert W et al (2019) A method of driver's eyes closure and yawning detection for drowsiness analysis by infrared camera. *2019 1st International Symposium on Instrumentation, Control, Artificial Intelligence, and Robotics ICA-SYMP 2019*, pp 61–64. <https://doi.org/10.1109/ICA-SYMP.2019.8646001>
10. Kailasam S et al (2019) Accident alert system for driver using face recognition. In: *IEEE International Conference on Intelligent Techniques in Control, Optimization and Signal Processing INCOS 2019*, pp 1–4. <https://doi.org/10.1109/INCOS45849.2019.8951320>
11. Koldijk S et al (2018) Detecting work stress in offices by combining unobtrusive sensors. *IEEE Trans Affect Comput* 9(2):227–239. <https://doi.org/10.1109/TAFFC.2016.2610975>
12. Saidi A et al (2020) Hybrid CNN-SVM classifier for efficient depression detection system. In: *Proceedings of the International Conference on Emerging Technologies and Intelligent Systems IC_ASET 2020*, pp 229–234. https://doi.org/10.1109/IC_ASET49463.2020.9318302
13. Kumar P et al (2020) Assessment of anxiety, depression and stress using machine learning models. *Procedia Comput Sci* 171(2019):1989–1998. <https://doi.org/10.1016/j.procs.2020.04.213>
14. Wongkoblap A et al (2019) Modeling depression symptoms from social network data through multiple instance learning. In: *AMIA Jt. Summits on Translational Science Proceedings. AMIA Jt. Summits Transl. Sci.* 2019, pp 44–53
15. Shanmugasundaram G et al (2019) A comprehensive review on stress detection techniques. In: *2019 IEEE International Conference on Computing, Communication and Automation Networking, ICSCAN 2019*, pp 1–6. <https://doi.org/10.1109/ICSCAN.2019.8878795>
16. Parmar A, Katariya R, Patel V (2019) A review on random forest: an ensemble classifier. In: Hemanth J, Fernando X, Lafata P, Baig Z (eds) *International Conference on Intelligent Data Communication Technologies and Internet of Things (ICICI) 2018*. ICICI 2018. *Lecture Notes on Data Engineering and Communications Technologies*, vol 26, pp 758–763. Springer, Cham. https://doi.org/10.1007/978-3-030-03146-6_86
17. Akbarian S et al (2019) Automated non-contact detection of head and body positions during sleep. *IEEE Access* 7:72826–72834. <https://doi.org/10.1109/ACCESS.2019.2920025>
18. Reddy US et al (2018) Machine learning techniques for stress prediction in working employees. In: *2018 IEEE International Conference on Computational Intelligence and Computing Research ICCIC 2018*, pp 1–4. <https://doi.org/10.1109/ICCIC.2018.8782395>

UAV-Enabled Supply Chain Architecture for Flood Recovery in Smart Cities



Theodoros Anagnostopoulos, Faidon Komisopoulos, Ioannis Salmon, and Klimis Ntalianis

Abstract An Unmanned Aerial Vehicle (UAV) supply chain architecture is used to handle flood water leak incidents occurred by physical disasters in Smart Cities (SCs). Floods produce serious problems and inefficiencies in problematic sectors of water grid. Such incidents are treated as a supply chain problem, where each incident is assigned to a certain priority. On a trigger occurrence the nearest UAV reaches a certain flood water leak incident where the assigned plumber at the SC control system assesses the problem. SC personnel evaluate if it is a high significance incident to fix the problem in real time. In this paper, there are presented certain use cases, which are evaluated with proposed metrics incorporated by control system supply chain architecture to infer the optimum use case for SC recovery. Such a use case is proposed to be adopted by SC control system architecture to handle upcoming scenarios of flood physical disasters.

Keywords Flood water leak · Physical disaster recovery · Autonomous UAV · Control system · Supply chain · Architecture · Smart city

1 Introduction

Changes in global climate warming effect result in diverse physical disasters which alter the environment of nowadays living. Such disasters have high impact to citizens'

T. Anagnostopoulos (✉) · F. Komisopoulos · I. Salmon · K. Ntalianis
DigiT.DSS.Lab, Department of Business Administration, University of West Attica, Aigaleo, Greece

e-mail: Theodoros.Anagnostopoulos@uniwa.gr

F. Komisopoulos

e-mail: fedonk@uniwa.gr

I. Salmon

e-mail: isalmon@uniwa.gr

K. Ntalianis

e-mail: kntal@uniwa.gr

wellbeing in Smart Cities (SCs) green and sustainable ecosystems. Human kind witness several serious physical disasters the last decades like floods, earthquakes and tsunamis waves. Such disasters affect citizens' daily life since they cause severe injuries to earth's population. It is proposed a control system supply chain architecture to treat the effects of physical disasters to human population living in SCs by assessing autonomous Unmanned Aerial Vehicle (UAV) technology potentiality. Certain fleet of autonomous UAVs is located in several areas within the SC. Such a fleet is available to provide remote evidence of a critical flood water leak in Smart City's public water grid network. Specifically, assigned personnel is able to decide in real time a high significant flood water leak incident and send a vehicle with well trained plumbers to fix the problematic area.

In this paper it is proposed a control system architecture to face flood water leak significant incidents in SC spatial area. Flood water leak recovery is considered as a supply chain problem where a limited amount of personnel, dedicated municipality vehicles and UAVs are invoked to serve certain flood water leak problems in a SC green and sustainable ecosystem. A number of use cases are studied to infer which is optimal for certain SC adoption. Specifically, Use Case I (UC-I) is a centralized approach where UAVs are assumed to located in the centre of the SC, while an alternative Use Case II (UC-II) assumes a decentralized distributed infrastructure to assess the proposed control system architecture efficiency. Concretely, current study adopts certain evaluation metrics, which are incorporated to infer the optimal use case for actual adoption by SCs.

Structure of the paper is presented as follows. Section 2 examines prior work in the research area. In Sect. 3 is presented the proposed control system supply chain architecture. Section 4 analyses the proposed evaluation system parameters, while Sect. 5 describes certain metrics used to assess the adopted system. Section 6 defines the examined use cases. In Sect. 7 are performed the experiments. Section 8 discusses the observed results, while Sect. 9 concludes the paper and proposes future work.

2 Prior Work

Research in flood water leak include interested related works that have been presented in the past. A satellite and UAV-based remote sensing model for assessing the risk of flood water leak in village relocation management is proposed by [1]. A systematic survey on addressing physical disasters in smart cities by incorporating UAV path planning in 5G wireless network technology is analysed in [2]. A multi-criteria optimization method for efficient deployment of UAV-enabled flood disaster management proposed by [3]. Scheduling of complex emergency tasks for assisting multiservice UAVs in cases of post-disaster scenarios is presented in [4]. A system incorporating planning UAV technology activities for achieving efficient citizen's coverage in flood physical disaster areas is supported in [5]. Incorporation of UAV is social logistics to face natural flood water leak disaster's response for providing humanitarian relief aid is proposed by [6]. Deployment of crowdsourcing technology, which faces disaster damages due to floods by incorporating distributed UAV monitoring is proposed in [7].

An integrated multimodal approach incorporating ground and UAV monitoring in flood physical disaster for enhancing detailed investigations is provided by [8]. A system based on UAV technology, which is able to monitor online and in real time flash flood disasters using Lagrangian Internet of Things (IoT) microsensors is proposed in [9]. A system design, which incorporates distributed UAV architecture for treating flood water leak area in smart cities for further data analytics and inference is provided in [10]. Flood water leak mapping is used for detection of areas prone to flood incorporating Synthetic Aperture Radar (SAR) is introduced in [11]. A system, which is able to perform human detection and rescue after physical flood disasters by incorporating UAV technology is adopted in [12]. A flood water leak area detection model, which is used on a UAV monitoring system in case of physical disasters is proposed by [13]. A flood evaluation system for smart city recovery in critical areas by incorporating UAV surveillance system methodology is analysed in [14].

An integrated methodology incorporating deep learning algorithms enhanced with region growing advancements able to exploit UAV optical data sources potentiality for flood extent mapping is adopted by [15]. Densely connected Recurrent Neural Network (RNN) architecture is able to detect flood disaster in smart cities based on UAV high resolution images as proposed in [16]. Flooding physical disaster is able to be faced by combining Unmanned Surface Vehicles (USV) and UAV infrastructure towards an integrated rescue system as analysed in [17]. Flood water leak level detection observed in physical disasters based on UAV technology for assuring smart cities sustainable environment is adopted by [18]. Disaster management enhances UAV and LoRa networks to exploit flood water leak performance evaluation in smart cities proposed by [19]. Reliable physical flood disaster response is feasible by incorporating an integrated social media platform along with a UAV sensing system as analysed in [20]. A UAV surveillance system is implemented to search of survivors during flood physical disasters in smart cities as proposed by [21].

A multicriteria supply chain model for decision support is proposed in [22], which provides decision support utilities to policy makers in selecting certain public programs. Such models are able to promote an efficient use of scarce resources to audit and control supply chain situations. Decision support systems, which assist supply chain frameworks are described in [23]. Such support systems are based on cost benefit and cost efficiency analysis to set limits on certain dimensions incorporating models are able to treat. Multi criteria techniques are able to consider value tradeoffs from the decision makers aspect point of view to produce a synthetic measure, which summarize the supply chain system performance of investment options.

A supply chain system is proposed in [24], which is based on thermal imaging as a contemporary technique to monitor agricultural crop water management. Such system is able to estimate effectively canopy surface temperature as well as the ability to predict on time crop water levels. A sustainable low power wireless sensor network based on a contemporary supply chain architecture is described, in [25], to manage a variety of physical disasters. The distributed sensor network frames the coverage area by incorporating tiny energy efficient sensors, which are able to monitor and convey the data generated by the embedded sensors into the sustainable infrastructure. Such system is used in metro cities, which need to be monitored to provide their citizens well-being in case of physical disasters.

Proposed research efforts are extensive in facing flood physical disaster in SCs by incorporating certain architectures and assisting technologies. However, the examined studies do not face in great detail the upcoming problem of plumber vehicles' fleet management based on remote assessment of the flood water grid incidents. Facing this supply chain problem, in this paper we incorporate an autonomous UAV control system supply chain architecture, which is proposed to handle flood water leak incidents. Observed incidents are prioritized as low, medium or high significance problematic situations based on the assigned personnel decision. After evaluating a high significance incident, a vehicle with plumber personnel is invoked to treat the flood water leak problem of the certain water grid place. Adopted use cases are validated and evaluated to prove the efficiency of the proposed scheme solution.

3 Control System Supply Chain Architecture

Proposed control system supply chain architecture is based on certain detection and service models used to treat an upcoming flood incident. Such models are able to mitigate the emerged disaster risk and provide recovery in real time. Control system supply chain architecture overview is presented in Fig. 1. The adopted detection model is fed with input the current status of the system, i.e., if there is a flood water leak incident to serve, as well as the location of the incident in the SC's water grid. Model's results are the verification that the incident has been served by certain plumber personnel and vehicle supply chain availability. In case a flood incident occurs, proposed model calls the service model providing the location of the incident in the SC area. The detection model is presented in Table 1.



Fig. 1 Control system architecture overview: **a** SC municipality control centre, **b** flood water leak, **c** municipality plumb vehicle, **d** personnel plumber, and **e** UAV

Table 1 Detection model

#	Detection model
1	Input: <i>status, l_i</i> //Status, incidence location
2	Output: <i>status</i> //Trigger status
3	Begin
4	<i>status = 0</i> //Originally there is no trigger to handle
5	While (True) Do
6	If <i>status = 1</i> Then //If a phone call trigger happens
9	<i>status</i> ← <i>service(status, l_i)</i>
10	//Invoke service algorithm to handle the event
11	End If
12	<i>return(status)</i>
13	End While
14	End

When the service model is called, system architecture engages the closest autonomous UAV to visit the flood water leak incident location in the SCs public water grid. Prior to the round trip it is checked if the battery lifetime is greater than the estimated time to serve the incident, thus to verify that the selected UAV has enough energy to perform the task. In case battery is drained, system engages subsequently the next closest autonomous UAV to serve the emerged flood incident location. The process is repeated until an available UAV has adequate battery level. This UAV is selected and is navigated autonomously to the incident location by incorporating shortest path algorithm. When the UAV arrives to the incident location the embedded camera and microphone are activated accordingly to provide the assigned personnel in the SC municipality control centre a view of the flood water leak damage significance. In case the incident significance is considered high by the assigned personnel the system calls a plumber vehicle to fix the flood incident and recover the damage. Subsequently, when the incident has been served the status of the system model is updated to be ready to serve a consecutive flood incident. Service model is presented in Table 2.

Table 2 Service model

#	Service algorithm
1	Input: $status, l_i$ //Status, incidence location
2	Output: $status$ //Trigger status
3	Begin
4	<i>engage UAV nearest to l_i</i>
5	While (<i>True</i>) Do
6	If (<i>battery lifetime > estimated time</i>) Then
7	//If UAV has enough energy to perform task
8	<i>navigate autonomous UAV to l_i</i>
9	//Incorporate shortest path to reach l_i
10	<i>enable remote diagnosis</i>
11	//Doctor provide remote diagnosis
12	If (<i>incidence significance = high</i>)
13	Then
14	<i>call an ambulance to transfer patient</i>
15	//Transfer patient from l_i to hospital
16	End If
17	<i>status</i> $\leftarrow 0$ //Indicate that incident has been served
18	<i>return(status)</i> //Exit service algorithm
19	Else
20	<i>engage another UAV nearest to l_i</i>
21	End If
22	End While
23	End

4 Evaluation Parameters

Certain evaluation parameters are adopted to assess the effectiveness of the proposed flood water leak incident use cases. Concretely, the expected number of incidents occurred in such a physical disaster are decomposed to average number of incidents i , which are characterized by the assigned SC control centre personnel either as low, l , medium, m or high, h , significance flood incidents. In addition, the average number of autonomous UAVs engaged, u , to reach the problematic areas during average number of incidents is also considered by the system control architecture. UAVs use battery which need recharge when drain in periodical time occurrence. Specifically, an indicator of the average number of battery recharges is also proposed for examination, b , by the system. Concretely the average amount of time required,

Table 3 Evaluation parameters

Parameter	Description
<i>i</i>	Average total number of incidents
<i>l</i>	Average number of low significance incidents
<i>m</i>	Average number of medium significance incidents
<i>h</i>	Average number of high significance incidents
<i>u</i>	Average number of autonomous UAVs
<i>b</i>	Average number of battery recharges
<i>t</i>	Average amount of time required
<i>d</i>	Average distance covered

t, as well as average distance to be covered in each invocation, *d*, to serve a flood water leak incident is also important. System’s adopted evaluation parameters are presented in Table 3.

5 Evaluation Metrics

Assessing the performance of the proposed system per each use case can be performed by defining certain evaluation metrics, which assess the adopted values of a given evaluation parameter. Assuming, that $r_1 = \frac{u}{i} \in [0, 1]$ is the average number of UAVs per average incidence ratio it is able to measure the average number of UAVs engaged per average number of incidents. Specifically, low value of r_1 indicates an optimal use case since less average number of UAVs are engaged. Concretely, let us assume, $r_2 = \frac{h}{l+m+h} \in [0, 1]$ to be the average number of high significance incidences per average number of low, medium or high incidents occurred. It holds that low value of r_2 indicates an efficient use case since less average number of high significant incidents occurred and need treatment. Let us, also, consider $r_3 = \frac{b}{t} \in [0, 1]$ to be the ratio of average number of battery recharges per average amount of time required to serve an incident. In this case a low value of r_3 indicates an effective way to serve average amount of incidents. Subsequently, assume, $r_4 = \frac{d}{t} \in [0, 1]$ to be the ratio of average velocity (i.e., ratio of average distance covered per average time required) observed during average UAVs certain invocation. A high value of r_4 indicates optimal service average velocity of the proposed system. Proposed evaluation metrics are presented in Table 4.

Table 4 Evaluation metrics

Metric	Measures
$r_1 = \frac{u}{i} \in [0, 1]$	Average number of UAVs engaged per average total number of incidents
$r_2 = \frac{h}{l+m+h} \in [0, 1]$	Average number of high significance incidences per average number of low, medium or high incidents occurred
$r_3 = \frac{b}{t} \in [0, 1]$	Average number of battery recharges per average amount of time required to serve an incident
$r_4 = \frac{d}{t} \in [0, 1]$	Average velocity (i.e., ratio of average distance covered per average time required) observed during average UAVs invocation

6 Use Cases

Proposed control system supply chain architecture is evaluated on certain categories of SC sustainable infrastructure. Let us define, UC-I, which describes a centralized system infrastructure where certain number of autonomous UAVs is located in a central base station at the centre of the SC spatial area. In addition, assume that is incorporated a number n of UAVs to serve a possible flood water leak incident in the SC area. In case a phone call triggers the control system, the closest UAV is invoked to serve the emerged flood incident. This is a continuing process until all the available UAVs are invoked. Please note that in this use case is assumed that the SC terrain is not divided to certain area sectors, which in turns means that on a next control system trigger a UAV is possible to engaged at any location within the whole SC spatial area. Intuitively it can be inferred that in such a case the system resources are not used effectively and may not be sufficient to serve the upcoming citizens' needs, on real time, to provide adequate recovery services. UC-I is presented in Fig. 2.

Assume a UC-II, which describes a decentralized distributed infrastructure where the SC spatial area is divided to certain number of equal sized p sectors according to the number n of the available autonomous UAVs by the SC control system supply



Fig. 2 UC-I

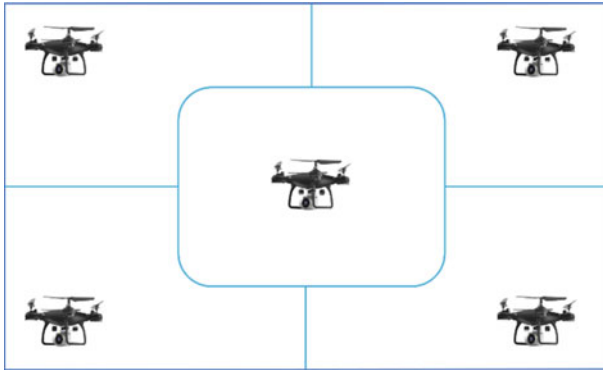


Fig. 3 UC-II

chain architecture. In this case, each sector has assigned a unique UAV to serve its flood water leak recovery needs. It is assumed that there is a certain n number of base stations at the centre of each sector. When a control system trigger occurs, it is invoked the closest UAV, which is assigned to the nearest sector. In this case, it also holds that the process is continued until all the available and charged UAVs are invoked. However, the main difference is that when a new trigger is occurred it is served by the closest UAV, which is actually located in the nearest neighbour sector of the upcoming flood incident. UC-II is presented in Fig. 3.

7 Experiments

Specific experiments are performed in this research effort to assess the effectiveness of each use case and define which is more efficient in case of flood physical disasters in SCs. It is assumed that the control system architecture of the SC municipality has $n = 10$ autonomous UAVs in the whole SC spatial coverage. In case of UC-I the autonomous UAVs are all of them located at the city centre, while in case of UC-II UAVs are equally distributed to the available $p = 10$ sectors of the SC area. Specifically, for UC-II it holds that for each sector is assigned a unique UAV. We run the experiments for a number of $it = 1000$ iterations. For certain iteration, the control system supply chain architecture is invoked several times according to the incidents, ic , occurred, which are following a random distribution between the interval $ic \in (0, 100]$ incidents. We used a random distribution of flood water leak incidents in several places of the SC to eliminate bias of the system's observed results. Experimental parameters are presented in Table 5.

Table 5 Experimental parameters

Experimental parameter	Value
n	10
p	10
it	1000
ic	(0,100]

8 Results and Discussion

The adopted control system supply chain architecture is fed with certain experimental parameters as input, which output the values as a result of the evaluation parameters. Specifically, proposed approach is based on the values of the evaluation parameters, which leads the system to result on certain metrics. Such metrics, are able to assess the effectiveness of each use case. Concretely, in Fig. 4 are presented the results of r_1 metric for UC-I and UC-II, respectively. It can be observed that r_1 values for UC-II are less than that of UC-I. This result indicates that UC-II is an optimal use case since less average number of UAVs are engaged by the control system supply chain architecture.

Subsequently, results for r_2 metric are presented in Fig. 5. It can be observed that values of r_2 metric for UC-I are greater than that of UC-II. This outcome indicates that UC-II is an effective use case since less average number of high significant incidents occurred and need treatment from the available supply chain infrastructure.

Subsequently, in Fig. 6 are presented the r_3 metric results. Please note that r_3 values for UC-II are less than that of UC-I, which is an indication that UC-II is a use case with an efficient feature to serve average amount of flood water leak incidents in the SC coverage area.

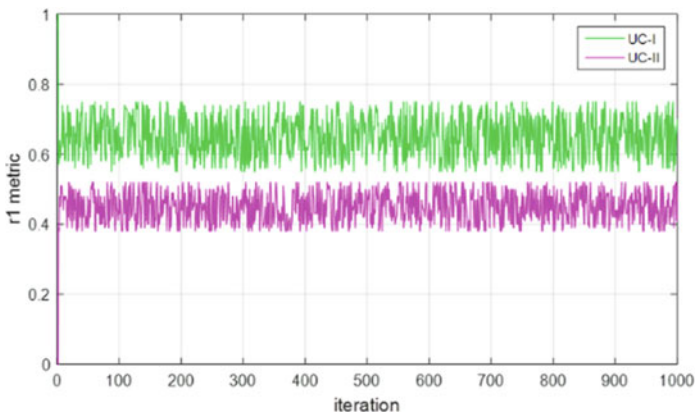


Fig. 4 r_1 metric for UC-I and UC-II

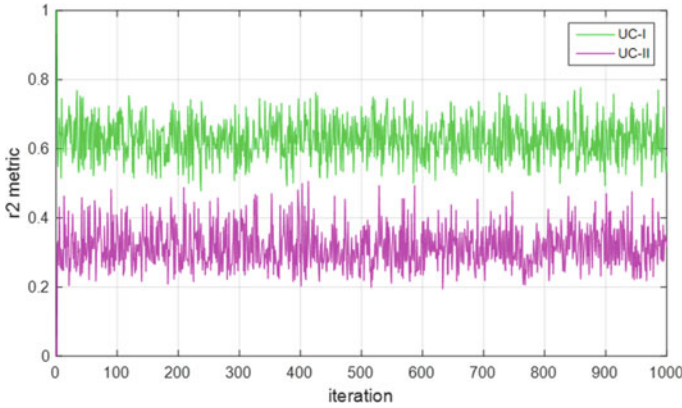


Fig. 5 r_2 metric for UC-I and UC-II

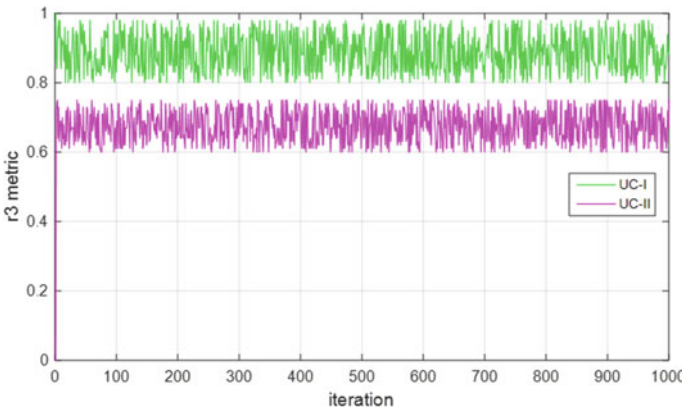


Fig. 6 r_3 metric for UC-I and UC-II

In addition, r_4 metric results are presented in Fig. 7. It can be observed that values of r_4 metric for UC-I are less numerically than that of UC-II, which is an indication that UC-II is a use case with optimal service of average velocity for the adopted flood physical disaster recovery system architecture.

Concretely, a comparison of the observed results from the performed experiments in the current research effort leads to the inference that UC-II use case is optimal compared with the UC-I, since it has less r_1 , r_2 , and r_3 values than UC-I. Subsequently, UC-II has higher r_4 metric values than UC-I. The inferred optimality is based on the adopted evaluation metrics used to assess the efficiency of both use cases as presented in Sect. 5, where is analysed the evaluation methodology which leads to certain outcome of the current research effort.

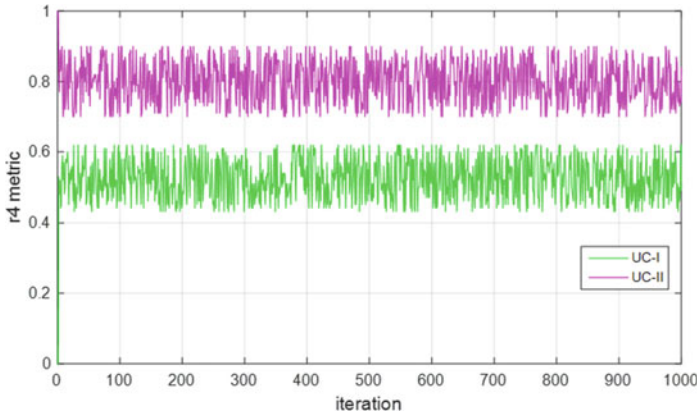


Fig. 7 r_4 metric for UC-I and UC-II

9 Conclusions and Future Work

Due to the global climate warming effect there is an increase of physical disasters like floods, earthquakes and tsunami waves. Such disasters affect citizens' wellbeing since they cause severe injuries to the SCs' population. In this paper it is proposed a control system supply chain architecture to face the effects of flood physical disasters to citizens living in green and sustainable SCs. It is incorporated a fleet of autonomous UAVs, to serve flood water leak incidents of certain significance, such as low, medium and high emerged incidents. Control system architecture is based on detection and service models to serve an emergency flood incident after a physical disaster in SCs. A fleet of autonomous UAVs is located in certain areas within the SC spatial coverage area according to certain adopted use cases. Such use cases are a centralized UC-I and a decentralized distributed UC-II, which are evaluated to assess to the efficiency of the proposed control system supply chain architecture.

Certain evaluation metrics and parameters are proposed to infer which use case is optimal for adoption by SCs in cases of flood physical disasters in the city's water grid network. Such a service is considered as a supply chain problem where a limited number of UAVs, municipality vehicles and plumber dedicated personnel is assigned to treat an emerging flood water leak incident. In current research effort it is performed certain experiments where it is proved that UC-II is optimal compared with UC-I according to the proposed evaluation metrics. Future research directions are to provide further exploitation of the adopted detection and service models by incorporating Artificial Intelligence (AI) and Multi-Agent System (MAS) design. Such intelligent systems are able to serve the upcoming incidents based on stochastic knowledge of prior incurred flood water leak incidents observed at the SC coverage area.

References

1. Cheng J et al (2022) Satellite and UAV-based remote sensing for assessing the flooding risk from Tibetan lake expansion and optimizing the village relocation site. *Sci Total Environ* 802:1–14
2. Qadir Z, Ullah F, Munawar HS, Turjman FA (2021) Addressing disasters in smart cities through UAV's path planning and 5G communications: a systematic review. *Comput Commun* 168:114–135
3. Masroor R, Naeem M, Ejaz W (2021) Efficient deployment of UAVs for disaster management: a multi-criterion optimization approach. *Comput Commun* 177:185–194
4. Rottondi C, Malandrino F, Bianco A, Chiasserini CF, Stavrakakis I (2021) Scheduling of emergency tasks for multiservice UAVs in post-disaster scenarios. *Comput Netw* 184:1–13
5. Malandrino F, Chiasserini CF, Casetti C, Chiaraviglio L, Senacheribbe A (2019) Planning UAV activities for efficient user coverage in disaster areas. *Ad Hoc Netw* 89:177–185
6. Estrada MAR, Ndoma A (2019) The uses of unmanned aerial vehicles—UAV's—(or drones) in social logistic: natural disasters response and humanitarian relief aid. *Procedia Comput Sci* 149:375–383
7. Guntha R, Rao SN, Shivdas A (2020) Lessons learned from deploying crowdsourced technology for disaster relief during Kerala floods. *Procedia Comput Sci* 171:2410–2419
8. Diakakis M et al (2019) An integrated approach of ground and aerial observations in flash flood disaster investigations. The case of the 2017 Mandra flash flood in Greece. *Int J Disaster Risk Reduct* 33:290–309
9. Abdelkader M, Shaqura M, Claudel CG, Gueaieb W (2013) A UAV based system for real time flash flood monitoring in desert environments using Lagrangian microsensors. In: *International Conference on Unmanned Aircraft Systems (ICUAS)*, Atlanta, GA, USA, pp 25–34, May 2013
10. Mohanty MD, Mohanty MN (2018) Design of a quadcopter UAV for flood area data analysis. In: *International Conference on Data Science and Business Analytics (ICDSBA)*, Changsha, China, Sep., pp 3–7
11. Jardosh P, Kanvinde A, Dixit A, Dholay S (2020) Detection of flood prone areas by flood mapping of SAR imagery. In: *International Conference on Smart Systems and Inventive Technology (ICSSIT)*, Tirunelveli, India, Aug., pp 814–819
12. Tariq R, Rahim M, Aslam N, Bawany N, Faseeha U (2018) DronAID: a smart human detection drone for rescue. In: *International Conference on Smart Cities: Improving Quality of Life Using ICT & IoT (HONET-ICT)*, Islamabad, Pakistan, Nov., pp 33–37
13. Popescu D, Ichim L, Caramihale, T (2015) Flood areas detection based on UAV surveillance system. In: *International Conference on System Theory, Control and Computing (ICSTCC)*, Cheile Gradistei, Romania, Oct., pp 753–758
14. Sumalan AL, Popescu D, Ichim L (2016) Flood evaluation in critical areas by UAV surveillance. In: *International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, Ploiesti, Romania, Jun., pp 57–62
15. Beni LH, Gebrehiwot AA (2021) Flood extent mapping: an integrated method using deep learning and region growing using UAV optical data. *IEEE J Sel Topics Appl Earth Obs Remote Sens* 14:2127–2135
16. Rahnemoonfar M, Murphy R, Miquel MV, Dobbs D, Adams A (2018) Flooded area detection from UAV images based on densely connected recurrent neural networks. In: *International Geoscience and Remote Sensing Symposium*, Valencia, Spain, Jul., pp 1788–1791
17. Zhang J, Xiong J, Zhang G, Gu F, He Y (2016) Flooding disaster oriented USV & UAV system development & demonstration. In: *IEEE OCEANS*, Shanghai, China, Apr., pp 1–4
18. Rizk H, Nishimur Y, Yamaguchi H, Higashiro T (2022) Drone-based water level detection in flood disasters. *Environ Res Public Health* 19(237):1–15
19. Saraereh OA, Alsaraira A, Khan I, Uthansakul P (2020) Performance evaluation of UAV-enabled LoRa networks for disaster management applications. *Sensors* 20(2396):1–18

20. Rashid MT, Zhang DY, Wang D (2020) SocialDrone: an integrated social media and drone sensing system for reliable disaster response. In: International Conference on Computer Communications (INFOCOM), Toronto, ON, Canada, Jul., pp 218–227
21. Ravichandran R, Ghose D, Das K (2019) UAV based survivor search during flood. In: International Conference on Unmanned Aircraft Systems (ICUAS), Atlanta, GA, USA, Jun., pp 1407–1415
22. Vivas V, Duarte Oliveira M (2017) Structuring multicriteria resource allocation models—a framework to assist auditing organizations. In: International Conference on Operations Research and Enterprise Systems (ICORES), Porto, Portugal, Feb., pp 321–328
23. Phelps C, Madhavan G (2018) Resource allocation in decision support frameworks. *Prior Setting Glob Health* 16:85–91
24. Shakya S (2021) Unmanned aerial vehicle with thermal imaging for automating water status in vineyard. *J Electr Eng Autom* 3(2):79–91
25. Kamel K, Smys S (2019) Sustainable low power sensor networks for disaster management. *IRO J Sustain Wirel Syst* 4:247–255

Outlier-Based Sybil Attack Detection in WSN



A. Jeyasekar, S. Antony Sheela, and J. Ansulin Jerusha

Abstract Sybil attack and its false-data injection are the most harmful attack in the Wireless Sensor Network. The sybil node impersonates like an authenticated user by bogusly generating node identity and sends the false/biased data to wireless sensor nodes and gateway nodes. Therefore, in this paper, we propose a security mechanism that uses outlier detection and trust estimation to detect the false/biased data generated by the sybil attack. Outlier detection is performed over the estimated trust and sensed data to detect the node outliers and data outliers which in turn used to detect the sybil attack and the false data injected by it. Further to it, triangle inequality between the nodes based on the geographical position of the nodes is used to detect the sybil attacks in the Wireless Sensor Network. Based on the node outlier detection and data outlier, sybil attack and false data injected by it in the network is identified and removed. The proposed security mechanism is experimented in a Wireless Sensor Network using ZigBee nodules and Raspberry Pi. Based on the empirical analysis, the median based outlier detection performs well than the hampel identifier and z-score. So, in the proposed system, the median based outlier detection method was used to detect the node outliers and data outliers. The experimental results show that the proposed method detects the sybil attack with more than 90% of detection rate and prevents the attack consequences from the decision making in the gateway node.

Keywords Outlier detection · Trust estimation · Sybil attack · False-data injection

A. Jeyasekar (✉)

Department of Computer Science and Engineering, Faculty of Engineering and Technology, SRM Institute of Science and Technology, Kattankulathur, Chennai, Tamilnadu, India
e-mail: ajeyasekar@yahoo.com

S. A. Sheela

University College of Engineering, Nagarcoil, Tamilnadu, India

J. A. Jerusha

Department of Electronics and Communication, SRM Valliammai Engineering College, Kattankulathur, Chennai, Tamilnadu, India

1 Introduction

Wireless Sensors are being used widely in the network application for detecting, monitoring and measuring the event in the geographic space. These applications measure the parameters of interest and send them to the centralized control unit for interpretation of measured data to know the criticality of events happened remotely. While sending and processing the data in the centralized unit, the system is prone to vulnerability and possible malicious attacks such as false data injection, impersonation attack, identity theft and fake identity creation etc. Among these attacks, fake identity creation and false data injection are most popular attack in the applications used for monitoring the water quality, pollution. In these systems, the attacker generates the fake identities and connects with centralized unit in order to compromise the effectiveness of the system. They send the false data to the centralized unit instead of actually measured data [19, 26]. Attacking the network with manipulated fake identity is called Sybil attack and injecting the false data instead of measured data is called false data injection attack. In the presence of these attacks, the monitoring system generates wrong reports. Not only the environment monitoring system, the impact of Sybil attack is more in the social network where the attackers spread spam and advertisements using other's identity or fake identity [17]. In addition, in the vehicular ad hoc network, these attackers send the biased traffic data that is used for data aggregation at the centralized unit which may collapse the traffic regulation system. Hence it is paramount importance to detect the Sybil attack and false data injection in the environment monitoring system because these attackers behave like a legitimate user.

Figure 1 depicts an example scenario which explains the direct and indirect sybil attack using two clusters. In direct attack, the malicious node in the cluster A creates three fake identities (black colored node) and connects them directly with gateway node to inject faulty data to the system. In case of indirect attack, the Sybil attacker uses the legitimate identity, pretends like a legitimate node and communicates to gateway node as shown in the cluster B. It pollutes the network with bogus information which is interpreted as genuine information by the honest node. Recent research efforts have been focused on detection and prevention of Sybil attack in the different network like social network, mobile ad hoc network, wireless sensor network etc. [12, 19–22, 24, 28–30, 35]. These methods have their own pitfalls also. For example, in the trust-based method [30], there is a chance that the dishonest node may send the biased recommendations about its neighbor node. Suppose these recommendations are taken into account to estimate the trust value, the honest node may be marked as sybil attack or sybil attacker may be marked as honest node. In order to avoid it, in this paper we propose a system which uses the outlier detection method. It estimates the biased recommendation given by the dishonest nodes. The same outlier detection method is also used to detect the bogus data or false data injected by the dishonest nodes.

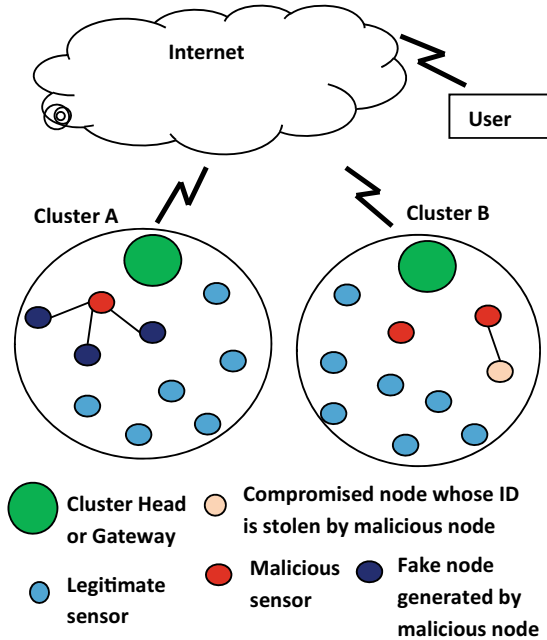


Fig. 1 Sybil attack scenario

In the proposed system, every node in a cluster evaluates the neighbor nodes and sends the trustworthy of its neighbors. In Fig. 1, the honest node (blue color) evaluates the malicious node (black color) based on its transaction experience and sends its evaluation reports to the gateway node. All kind of messages exchanges between the nodes are called transaction experiences. Similarly, all the nodes send its evaluation report about the neighbor nodes to gateway node. Normally these evaluation reports were used to estimate the trust value. But in the proposed system, we used the outlier detection method to detect and remove the biased evaluation reports and corresponding dishonest nodes. Further to it, the outlier detection is applied over the environmental data sent by the sensor nodes. Thus, the proposed mechanism detects the Sybil attack and the false data injection. Research contributions in this paper are given below.

1. We propose a registration protocol which creates the unique identity to the sensor nodes. It avoids the attackers to get into the WSN
2. We propose a outlier detection method which detects the sybil attackers based on the sensor node location and sensor node transactions with gateway node.
3. We propose another outlier detection method which removes the bogus or false data injected by the sybil attackers using histogram

2 Related Works

More attention has been given to avoid Sybil attack in recent days especially in social network, peer-to-peer network etc. Therefore, in this paper, we propose a mechanism which detects the Sybil attack and their false data injection using outlier detection method and trust estimation method.

2.1 Sybil Attack Detection

There are various methods used to detect the Sybil attack in the wireless sensor network such as Cryptography method, Random key pre-distribution method, Radio resource-based testing method, Neighbourhood data-based method [6], Node position and mobility-based method, Received signal strength based method, Trust based method and Energy consumption based method [1, 2, 5, 7]. There are various approaches such as Statistical En-route Filtering (SEF) [9], Dynamic En-route Filtering (DEF) [33], SEF using neighboring nodes [13]. These techniques are suitable for the wireless sensor network with large number of sensor nodes and but not suitable for detecting the impersonation on legitimate nodes [18]. Therefore, in this paper, we propose the mechanism that detects the sybil attack and the false data injected by them using outlier detection and trust estimation.

2.2 Outlier Detection

Generally, the data measured and collected by the sensor nodes are affected by error, noise and missing values which is called outliers. Sometimes these outliers are artificially generated by the malicious nodes. In order to remove the outliers from the collected data, statistical based outlier detection method, nearest neighbor-based outlier detection, clustering based outlier detection, classification-based outlier detection and spectral decomposition-based outlier detection are used [15, 16, 32]. The statistical based outlier detection is very simple, light weight method so that it is suitable for WSN. Hampel Identifier, z-score are such outlier detection methods [30, 32].

Hampel Identifier [15, 34] is a popular outlier detection method that uses median and median absolute deviation as a robust estimate of the spread of outliers. Hampel identifier has two configurable parameters that are size of the sliding window and the number of standard deviations which identify the outliers correctly [15]. z-score [3] estimates how far the data points are from mean. Z-score is a signed number of standard deviations by which the data point is above the mean of data points. If the z-score value is greater or less than the threshold value, the data point is identified as outliers. Therefore, in this paper, we used a median based outlier detection method for

identifying the outliers in the transaction success rate of neighbor nodes. It enables the proposed security mechanism to detect the attackers like sybil, grey hole and selective forwarding attacks because malicious nodes drop the packets received from the neighbor instead of forwarding it to other nodes.

2.3 Trust Estimation

A trust management system deals with monitoring neighbouring nodes, detecting misbehaviour, estimating trust values based on the recommendation, alerting the neighbour nodes about the malicious node and isolating the malicious nodes from the network. Estimating the trust value is an important process in the trust management system which uses various techniques such as probability-based trust estimation, weight-based trust estimation and fuzzy logic-based trust estimation [4, 5, 10, 23, 25, 30]. In the probability-based trust estimation method, the trust value is estimated as expectation value of probability distribution of node's future behavior. Each sensor node estimates a reputation value for other nodes by monitoring the transactions and updates it using either bayes theorem or dirichlet process [10, 11]. Since this type of approaches have high computational complex, it is not suitable for wireless sensor applications. In [25], Gaussian distribution and Bayesian theory are used to estimate the reputation and trust of sensor nodes in wireless sensor network. This scheme also has high computational complexity [10] and gives higher overheads to wireless sensor network. In the weighting-based trust estimation method, trust and reputation are estimated by weighting the behavior/performance of the sensor nodes over time [10, 31]. This method makes it simple to estimate trust, and its implementation is also very simple.

In [31], the trust is calculated at node level, cluster head level and base station level. A sliding window scheme is used to update trust. The cluster head integrates trust values obtained from nodes using standard normal distribution and detects the trustworthy nodes, uncertain nodes, and untrustworthy nodes. The nodes with equal number of successful and unsuccessful operations are not considered as untrustworthy nodes because it degrades the system performance. The probability-based trust estimation has high computational complexity which is not suitable for wireless sensor network. The weighting-based trust estimation is lightweight, computation less complex and consumes less energy which meets the requirement of wireless sensor network. Therefore, in this paper, we propose the weighting trust estimation along with outlier detection and triangle inequality to improve the performance of detection of sybil attack.

3 Overview of Proposed System

The major components of the proposed system shown in Fig. 2 are registration and authentication, node outlier detection and data outlier detection. In registration and authentication, the sensor nodes that are inside the transmission area of gateway node, register and get a unique ID from gateway node. The attacks are identified and removed using node outlier detection and data outlier detection method. The node outlier detection takes the trustworthiness and triangle inequality property of geographical position of sensor nodes. The trustworthiness of a sensor node is estimated based on direct trust estimation and indirect trust estimation. Direct trust is estimated at the sensor nodes and the gateway node. In order to estimate the direct trust, the gateway node considers the transaction experiences with other sensor nodes. Similarly, every sensor node considers the transaction experiences with its neighbour sensor nodes to estimate the direct trust.

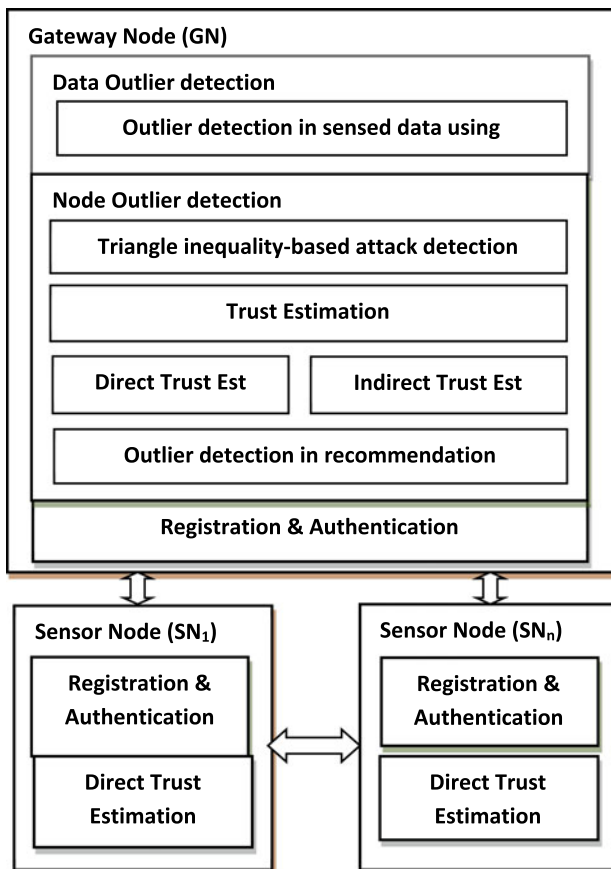


Fig. 2 Overview of proposed security mechanism

The indirect trust is estimated at the gateway node based on the recommendations given by a sensor node against its neighbour sensor nodes. The node outlier detection method identifies and removes the outliers in the recommendations i.e. the biased recommendation, bad recommendation etc. and estimates the indirect trust. A Sybil node generates multiple node identities bogusly and sends the environmental data to gateway node. But the geographical position of these multiple node identities is same. Therefore, triangle inequality property is applied over the geographical position of the multiple node identities to detect the sybil node. By these ways, the dishonest nodes and attack nodes are detected and removed. The data outlier detection method is applied over the environmental data collected from the honest sensor nodes. The outliers in the measured data is detected by using histogram and removed. Eventually, only the environmental data from the honest and legitimate node is sent to the web application.

4 Proposed Security Mechanism

The proposed approach is explained in threefold: Registration phase, Node outlier detection phase and Data outlier detection phase. The registration process of the sensor nodes with gateway node is carried out as a preliminary authentication mechanism and in second phase, node outlier detection identifies and removes the suspicious nodes based on outliers in transaction experiences, degree of trust of sensor nodes and using triangle inequality property. The third phase deals with detection of outliers in the environmental data and removes the dissimilar data using histogram. Eventually the gateway node alerts all the sensor nodes not to send the data to malicious node.

4.1 Registration and Authentication

In this registration phase, Gateway Node (GN) generates private/public key (KR_{GN} , KU_{GN}) and broadcasts its public key (KU_{GN}) and its location to all the sensor nodes in its radio coverage area periodically.

$$GN \rightarrow All : [KU_{GN} \| P(GN)] \quad (1)$$

The sensor node may be neighbour node (1 hop distance) or non-neighbour node (multi hop distance) to GN. We assume that each sensor node has an algorithm to detect its neighbour nodes. A sensor node (SN) who wants to join in this cluster generates private/public key (KR_{SNi} , KU_{SNi}) and sends a request message to GN.

$$SN_i \rightarrow GN : E_{KU_{GN}}[IP(SN_i \| C_{max} \| \|L_i \| KU_{SNi})] \quad (2)$$

where IP denotes the IP address of the node and L_j means the list of parameters that the sensor node is capable of measuring the environmental parameters. C_{max} is maximum flow path that a sensor node permits to flow through it. C_{max} restricts the malicious node to create more flow through the sensor node. Even though KU_{GN} is known to all the sensor nodes, this transaction cannot be decrypted by the malicious nodes because the private key KR_{GN} used for decryption is known only to GN. Therefore the malicious node cannot disclose any value like IP, C, L, K. GN decrypt the encrypted message using its private key (KR_{GN}) and performs the Hashing of IP address, C_{max} and L_i to create a unique node identifier $ID(SN_i)$

$$GN : D_{KR_{GN}}[E_{KU_{GN}}[IP(SN_i) \| C_{max} \| \| L_i \| KU_{SN_i}]] \quad (3)$$

$$GN : ID(SN_i) = H[IP(SN_i) \| L_i \| C_{max}] \quad (4)$$

GN maintains a table in which the $IP(SN_i)$, $ID(SN_i)$, C_{max} , and L_i are stored. The node identifier is sent to the sensor nodes. Therefore, each sensor nodes in the cluster have a unique node identifier issued by the gateway node. Creating the unique node identity does not solve the Sybil attack problem because the malicious node may have many such node identifiers. But C_{max} is used to control the number of attackers sending the data to GN through a sensor node.

$$GN \rightarrow SN_i : E_{KU_{SN_i}}[E_{KR_{GN}}[ID(SN_i), L_i]] \quad (5)$$

where L_j is list of parameters that GN requests the SN to measure and send. Since the KR_{SN_i} is a private key that is known only by SN_i , this transaction cannot be disclosed by the malicious node. The sensor node decrypts it using KR_{SN_i} and KU_{GN} .

$$SN : D_{KR_{SN}}[D_{KU_{GN}}[E_{KU_{SN_i}}[E_{KR_{GN}}[ID(SN_i), L_i]]]] \quad (6)$$

The SN_i is now registered with GN so that GN can get the intended environmental parameters from the sensor nodes. Equations (1)–(6) shows the message transaction between the gateway node and sensor node during the registration phase.

Algorithm 1 - Registration and Authentication

1. Gateway node broadcast its position and its public key 4
 2. Every node in the coverage of gateway node encrypts the IP address, List of sensor parameters etc using public key of gateway node
 3. All nodes unicast the encrypted message to gateway node
 4. Gateway node decrypts the message using its private key
 5. Gateway node generates the node ID for all responded nodes and encrypts the ID using the public key of corresponding nodes
 6. Gateway node sends the encrypted ID to corresponding nodes
 7. Sensor nodes decrypts the message using their private key
-

4.2 Node Outlier Detection

A malicious node generates the large number of identities bogusly or steals them from honest nodes and creates Sybil identities to attack the honest nodes in the network. These nodes are known as node outliers, which are detected by estimating the trustiness of nodes and geographical distance from the gateway node [16, 28].

Trust Estimation based Node Outlier Detection: The trustworthiness of a sensor node is estimated based on transaction experiences which mean all kind of message exchanges between the sensor nodes and GN [5, 10]. The successful and unsuccessful message exchanges are considered as transaction experiences. The trust value of a sensor node j is estimated by GN as given below

$$T_{GN,j}(t) = \alpha * D_{GN,j}(t) + (1 - \alpha) * R_{GN,j}(t) \quad (7)$$

where $T_{GN,j}(t)$ is the trust value of node j estimated by node GN at time t . $D_{GN,j}(t)$ is the direct trust estimated by GN about node j at time t based on the transaction experiences between node GN and j . $R_{GN,j}(t)$ is the indirect trust by node GN for the node j based on the transaction experiences between node k and j . α is a constant that lies between 0 and 1 which determines the weight to be given for direct trust and indirect trust. The experimental results show that $\alpha = 0.5$ provides better performance. If the estimated trust value ($T_{GN,j}$) of node j is greater than the threshold φ , then the node j is honest node and can be used for environmental data collection. Similarly, GN evaluates all the neighbour nodes and finds whether the neighbours are honest or not. The direct trust is estimated as given below

$$D_{GN,j}(t) = \beta * D_{GN,j}(t - 1) + (1 - \beta) * \left[\frac{S_{GN,j}(t)}{S_{GN,j}(t) + F_{GN,j}(t)} \right] \quad (8)$$

where $S_{GN,j}(t)$ is the number of successful transaction experiences between the nodes GN and j . $F_{GN,j}(t)$ is the number of failure transaction experiences between the nodes GN and j . β is a constant that lies between 0 and 1. The experimental results show that $\beta = 0.2$ gives better performance.

There are many neighbours around the gateway node GN that send/receive the messages from each other. Based on the transaction experiences, every node in the cluster estimates the direct trust of its neighbour nodes and sends them to GN periodically. The GN considers them as indirect trust ($r_{k,j}(t)$). The recommendations from neighbours are used to determine the trustworthiness of a node. Some smart attacker provides bad recommendations or biased recommendation against the honesty node which is called bad-mouthing attack. In the presence of dishonest recommendations, the trust estimation becomes biased. Therefore, it is necessary to remove the dishonest recommendations before the estimation of indirect trust. The dishonest recommendations are the outliers in the set of recommendation

$$R = \{r_{1,j}(t), r_{2,j}(t), r_{3,j}(t), r_{4,j}(t) \dots \dots r_{m,j}(t)\} \quad (9)$$

where m indicates the number of nodes sent the recommendation about node j . The dishonest recommendations are the subset of recommendation set R that is considered as outliers of the set R or dissimilar recommendation set. In order to find out the set of dishonest recommendation from the set R , median and standard deviation are used. Let μ be the mean, M be the median, s be the standardizing level varying between 0 to 3 and σ be the standard deviation of set R . The lower control limit (LCL) and upper control limit (UCL) are given below respectively [3, 15]

$$\text{LCL} = \mu - Ms\sigma, \quad \text{UCL} = \mu + Ms\sigma \quad (10)$$

The trust recommendation between these limits is assumed as honest node. Otherwise, the node j is assumed as dishonest node and it is marked as suspicious node in the table maintained by GN. If j is honest node, the recommendations lying outside the boundary are removed. The resulting set is denoted as

$$R' = \{r'_{1,j}(t), r'_{2,j}(t), r'_{3,j}(t), r'_{4,j}(t) \dots r'_{n,j}(t)\} \quad (11)$$

where n indicates the number of recommendations from honest nodes and R' is the subset of R . The indirect trust of node j is estimated by doing simple averaging of R' as given below

$$R_{GN,j}(t) = \frac{\sum_{k=1}^n r'_{k,j}(t)}{n} \quad (12)$$

Triangle Inequality-based Node Outlier Detection: Since the GN knows the position of all the honest nodes, it calculates the distance between the nodes and applies the triangular inequality to detect the Sybil attack. The distance between the two nodes (SN_i, SN_j) is denoted as a distance function $d(SN_i, SN_j)$. GN calculates the distance $d1 = d(GN, SN_j)$, $d2 = d(GN, SN_i)$ and $d3 = d(SN_i, SN_j)$. If the distance satisfies the triangle inequality ie $|d2 - d3| < d1 < |d2 + d3|$, then SN_i and SN_j are neighbor. Since the malicious node (MN) creates n number of fake identities with same geographical location, $d1$ is much higher than or equal to $|d2 + d3|$ and $d1$ is much lower than or equal to $|d2 - d3|$. Hence the triangle inequality does not hold good and therefore the MN is assumed as Sybil attacker.

Algorithm 2 - Data collection and trustiness estimation

1. Gateway node collects the encrypted message of Node ID, Location information, sensor data, Recommendation about the neighbor nodes whenever required.
 2. Gateway node decrypts the encrypted message received from all sensors using sensor's public key and estimates the indirect trust based on the recommendation given by the sensor nodes.
 3. Gateway node estimates the direct trust for all sensor nodes based on the transactions it made with the sensor nodes
 4. Gateway node estimates the trust worthiness of sensor nodes using direct and indirect trust
 5. Gateway node estimates the distance to sensor nodes
-

4.3 Data Outlier Detection

Sybil attacker impersonates with fake identities or stolen identities and inject the false data/biased data to the gateway node. Sometimes, the attacker like bad mouthing attacker, collusion attacker and false data injection attacker does the successful transactions with neighbours with the intension of injecting the false environmental data. So GN may consider it as honest node because the trust estimation takes the transaction experiences for estimation. Therefore, there is possibility of having the false data in the collected environmental data. The smart attacker may give biased data to go undetected. Hence, we propose an approach in which the collected data is divided into 5 to 10 bins based on the range of possible values of environmental data. (DR1, DR2, ... DR10). By doing so, we can find which data are similar to each other. After grouping the collected data in their respective bins, the bin with higher count of data falling in, is considered as correct data.

$$H(R') = \{(DR1, c1), (DR2, c2), (DR3, c3), \dots (DR10, c10)\} \quad (13)$$

where c_i ($i = 1,2,\dots,10$) denotes the total number of time that the measured data falls in the range DRi. From the histogram $H(R')$, the data range (DR) with high frequency is selected.

Algorithm 3 - Node/data outlier detection

1. If the trustworthiness of sensor node is less than the threshold value, the recommendation given by the particular sensor node is removed and the node is marked as suspicious node
 2. If the triangle inequality is not satisfied with respect to the estimated distances, the recommendation and sensed data are removed and the node is marked as suspicious node
 3. Apply the histogram on the sensed data and select the data with high frequency and consider them as trustworthy sensed data from legitimate node.
 4. The ID of nodes which sends the data with low frequency are compared with suspicious nodes. If it matches, the nodes are considered as sybil attackers.
-

5 Experimental Setup

In this paper we consider a test bed that has three layers. They are sensor layer, gateway layer and web layer. At the sensor layer, a variety of sensors are connected with a sensing node which has two components i.e. zigbee transceiver and raspberry pi. The zigbee transceiver in sensor layer is configured as end device. The sensor nodes connected with a GN is known as a cluster. At gateway layer, Raspberry Pi is connected with zigbee transceiver which is configured as coordinator. GN collects the measured data from the sensor nodes and sends the data to digi cloud manager. The data is collected from the sensor every one hour. The gateway application provides service such as registration and authentication service, data collection service, attack detection service and file transfer service to web layer. The web layer consists of a web database and web application. The web database collects the environmental data from the cloud and web application visualizes the sensed data. In order to simplify the analysis, we assume only one gateway in the test bed. The application in web server provides services to web client interface. In this paper, a cluster is formed with sensor nodes and number of sensor nodes in a cluster is 40 including the attackers. We introduced Sybil attack and their false data injection in the network and analyzed the performance of proposed system. The data is collected from the sensor every one hour. The number of attackers in a cluster varies from 10 to 50% of total nodes. We assume 50% Sybil attackers and 50% false data injection in the experiments.

6 Performance Analysis

The performance of proposed mechanism is analysed based on registration phase, outlier detection and trust estimation. As compared with sybil attack detection methods given in the Table 1, the accuracy of proposed detection method is better

even though it has high communication cost and high memory cost. As compared with trust-based and outlier-based detection systems given in the Table 2, the proposed system has all features like registration and authentication modules. Trust estimation and outlier removal, location estimation and outlier removal, false-data removal etc.

Table 1 Comparison of sybil attack detection methods

Detection methods	Overhead	Memory cost	Accuracy
Cryptography method [2003] [21]	High communication overhead	High cost for key generation and key distribution. High memory cost	Depends on complexity of cryptographic algorithm and mechanism to store the secret key in the centralized system
Random key pre-distribution method [2004] [28]	High communication overhead	High implementation cost and memory cost because of pairwise keys, hash values and voting keys	Depends on number of nodes in the network
Radio resource-based testing method [2015] [29]	High energy consumption	High implementation cost	Depends on number of channels used by the attackers
Neighborhood data-based method (2009) [22]	Less energy consumption	High memory cost	Depends on number of honest nodes having same set of neighbor nodes in the network and the communication range
Node position and mobility-based method (2006) [8]	High energy consumption	High memory cost	Depends on environmental interference and similar movement of legitimate nodes
Received signal strength-based method (2022) [37]	Less communication overhead	Less memory cost	Depends on environmental interferences and antenna gain
Trust based method (2015) [10]	High communication overhead	high memory cost	Depends on number of sybil nodes in the network
Energy based method (2017) [27]	High communication overhead	Less memory cost	Depends on network mobility and energy used to send the legitimate nodes
Proposed method (outlier-based method)	High communication cost,	High memory cost	Depends on the threshold used to detect the node outliers

Table 2 Comparison of trust-based and outlier-based detection system

Detection method	Registration and Authentication	Trust estimation	Outlier removal using trust	Location estimation	Outlier detection using location	False data detection	Scalability support	Mobility support
Robust trust establishment scheme (2015) [10]	No	Yes	No	No	No	No	Yes	No
Trust-Aware Routing Framework (2012) [12]	Yes	Yes	No	No	No	No	Yes	No
Power-aware sensor selection based on trust (2010) [14]	No	Yes	No	No	No	No	Yes	Yes
Temporal and spatial outlier detection (2019) [15]	No	No	No	Yes	Yes	Yes	Yes	Yes
System based on energy trust system (2017) [27]	No	Yes	No	Yes	No	Yes	Yes	No
Group based trust management scheme (2009) (2017) [13, 31]	No	Yes	No	No	No	No	Yes	Yes

(continued)

6.1 Analysis of Registration Phase

By knowing the KU_{GN} , the malicious peer can send many requests with different IP and register as a legitimate peer. A GN is limited with max-flow paths from itself to web server and from the sensor nodes to GN. Hence the number of sensor nodes that can be registered with GN is limited based on the number of environmental parameters required. It avoids the malicious node that creates many IDs with GN. The GN first assigns max-flow-paths to C_{min} initially and increases linearly until reaching the C_{max} based on the number of successful transaction and number of honest nodes in the cluster. Value of C_{min} is initially set to 5 and value of C_{max} is set based on number of honest nodes in the cluster to get the intended environmental parameters.

During the registration process, the GN first checks for existence of IP address of the requested sensor node in the malicious node table. If the IP address of sensor node is already in the table, the concerned sensor node is considered as malicious node. If IP address and position of sensor node is same as in the GN's table, the concerned sensor node is considered as malicious node. Since the sensor node knows the position of GN, it will not register again and again with GN. Suppose a malicious node generates the private/public key, broadcasts the public key to all the sensor nodes, acts as a gateway node and starts a cluster, the malicious gateway node is not able to send the bogus data to the web server since there is a registration process for the gateway with web server.

6.2 Analysis of Outlier Detection

Every sensor node keeps track of transactions done with neighbour nodes and calculates the transaction success rate (TSR) as given below.

$$TSR = \left[\frac{S_{i,j}(t)}{S_{i,j}(t) + F_{i,j}(t)} \right] \quad (14)$$

where $S_{i,j}(t)$ is the number of successful transaction experiences between the nodes i and j . $F_{i,j}(t)$ is the number of failure transaction experiences between the nodes i and j . Outliers in TSR are detected using hampel identifier, z-score and median based outlier detection method. The performance of these outlier detection methods is analysed by varying the sigma s value in Eq. (3, 4 and 14) from 0 to 3. Figure 3(a), (b) and (c) shows the performance of these outlier filters for varying sigma value and varying number of attackers.

As the sigma value changes from 0 to 3, the detection rate is decreased. But these outlier detection methods are robust with respect to the increasing number of attackers. When the sigma is set to 0, the hampel identifier becomes as the standard median filter which in turn reduces the outlier detection. So the sybil attack detection

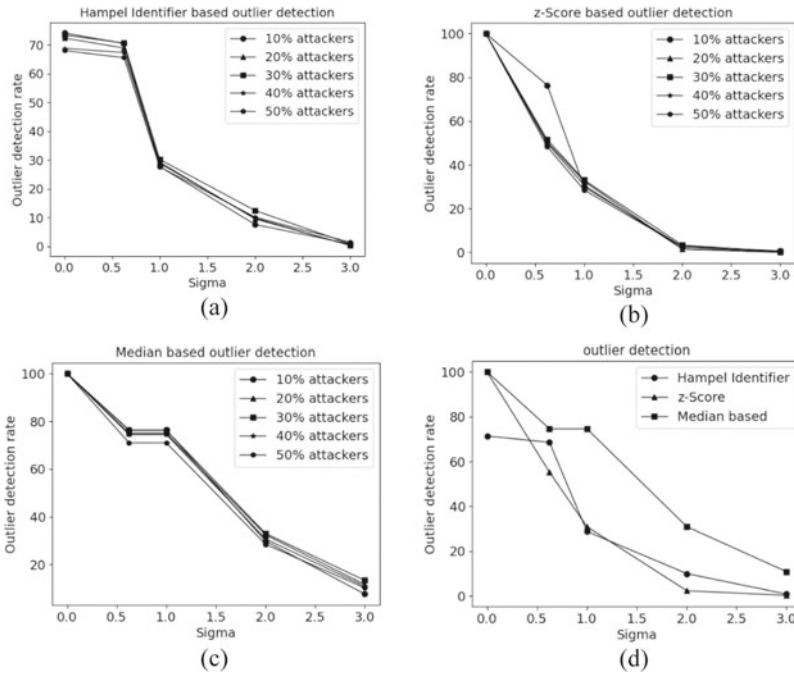


Fig. 3 Performance analysis of outlier detection methods **a** Hampel Identifier **b** z-Score **c** Median based outlier detection **d** comparison of three outlier detection methods

rate is decreased. As the sigma value increases, the outlier detection is more forgiving which allows more outliers. It increases the false positive which in turn degrade the detection rate of the sybil attacks. Therefore, the value of sigma plays vital role in deciding the attack detection rate. Among three outlier detection methods, the median based outlier detection performs well than the hampel identifier and z-score as shown in Fig. 3(d). Therefore, in the proposed system, the median based outlier detection method is used to detect the outliers in the transaction success rate. The gateway node and all the sensor nodes estimate the direct trust based on transaction experienced with other nodes directly using Eq. (12). The experiments are conducted by setting the value of beta in this equation varying from 0.2 to 0.8 and the direct trust threshold varying from 0.5 to 0.65. Based on the empirical analysis, $\beta = 0.2$ and direct threshold = 0.65 gives the better performance with respect to attacker detection rate. It means that more weight should be given to current transaction success rate (TSR) in Eq. (12). If the direct trust of a sensor node based on the transaction done at time t is greater than the threshold, then the concerned node is considered as attacker.

Figure (a) shows the detection rate of attackers using only direct trust estimation against varying beta and direct threshold. Figure 4(b) shows the detection rate of attackers based on trust estimation which include direct trust and indirect trust as given in Eq. (11). The trust estimation is evaluated by varying the trust threshold from 0.5 to 0.65, varying alpha value in the Eq. (11) from 0.2 to 0.8 and varying number of attackers from 10 to 50% of node deployed. The average of detection rate is shown in the Fig. 4(b). From the experimental results, alpha = 0.5 or 0.6 and trust threshold = 0.65 gives better performance than other values of alpha and threshold. From Fig. 4(c), it is observed that alpha = 0.6 and trust threshold = 0.65 give better performance with respect the detection rate.

Figure 4(d) shows the histogram of the row environmental data. In order to plot the histogram, the entire range of data sequences are divided into equal sized bins called classes and then for each bin, the numbers of points fallings into bin are counted. The bin with a smaller number of data points is considered as outliers and the bin with higher number of data points are considered as data from the legitimate nodes. In this paper, we develop the water quality monitoring system which measures the pH, EC value of water tank. Therefore, many of the data points lie in the bin ranging from 6.5 to 7.5 which are considered as data from legitimate node. The filtered environmental data after the removal of false data and attackers gives only the values ranging from

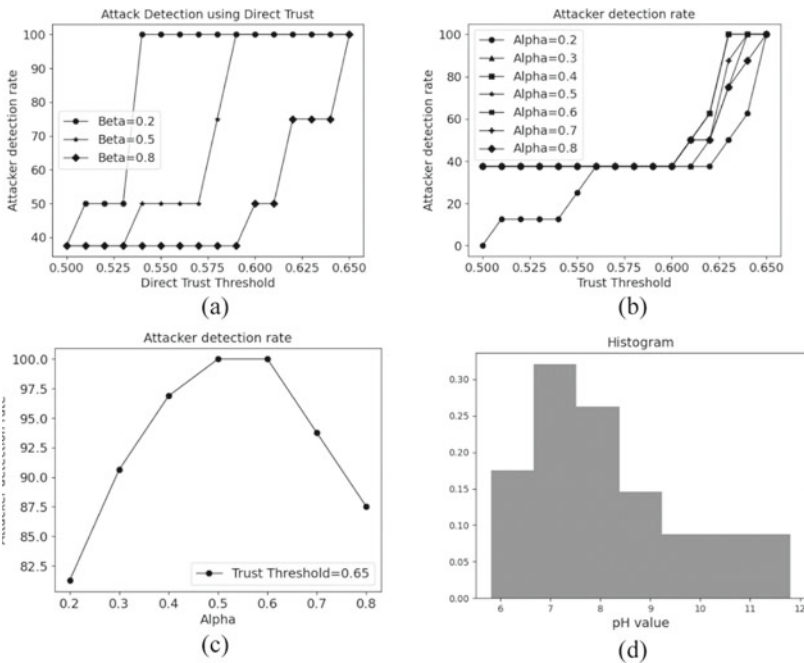


Fig. 4 Performance analysis of trust estimation **a** Direct trust estimation for varying weight (β) **b** Trust estimation for varying weight (α) **c** Attack detection rate of proposed system **d** Histogram of filtered data points

6.5 to 7.5. Eventually, the proposed system detects the Sybil attackers and removes the false data injected by sybil attacker. The computational complexity of the proposed outlier-based sybil attack detection system is higher than the trust-based sybil attack detection system explained in [10, 12, 15, 31] because the proposed system has more features like registration and authentication modules. Trust estimation and outlier removal, location estimation and outlier removal, false-data removal etc. as shown in the Table 2.

7 Conclusion

This paper presents a security mechanism to detect the Sybil attack and their false-data injection using outlier detection and trust estimation. Since the previous research work on sybil attack detection using trust estimation lacks in accurately estimating the trust value because of the biased recommendation. But in the proposed system, we used the outlier detection to remove the biased recommendations and false data injected by the sybil attackers. Based on the experiments, we found that the value of alpha, beta and sigma to 0.5, 0.2 and 1 respectively provides better performance in detecting the sybil attacker accurately.

However, the proposed system lacks in the estimation of normal and abnormal thresholds accurately for outlier detection because of the changing nature of attacker's behaviours in injecting the false data. In the proposed system, suppose the outliers are generated by the attackers not frequently, it is very difficult to assess the performance the outlier techniques effectively. Further to it, the outliers are sometimes generated by the legitimate sensor nodes because the sensors can be on or off sporadically over the period. Therefore, we plan to use the deep learning techniques for outlier detection in future which will definitely improves the performance of the proposed system [36]. While going for deep learning techniques, the extraction of features from the high volume of data becomes a big challenge because learning the complex structure of large volume of data is crucial.

References

1. Vasudeva A, Sood M (2018) Survey on Sybil attack defense mechanisms in wireless adhoc network. *J Netw Comput Appl* 120:78–118
2. Bhise AM, Kamble SD (2016) Review on detection and mitigation of Sybil attack in the network. *Procedia Comput Sci* 78:395–40
3. Kolbasi A, Unsal A (2019) A comparison of the outlier detection methods: an application on Turkish foreign trade data. *J Math Stat Sci* 5:213–234
4. Mohajer A, Bavaghar M, Saboor R, Payandeh A (2013) Secure dominating set-based routing protocol in MANET: using reputation. In: 10th International ISC Conference on Information Security and Cryptology, Yazd, Iran, 29–30 August 2013

5. Mohajer A, Hajimobini MH, Mirzaei A, Noori E (2014) Trusted-CDS based intrusion detection system in wireless sensor network (TC-IDS). *Open Access Libr J* 1:e848. <https://doi.org/10.4236/oalib.1100848>
6. Mohajer A, Somarin A, Yaghoobzadeh M, Gudakahriz S (2016) A method based on data mining for detection of intrusion in distributed databases. *J Eng Appl Sci* 11:1493–1501
7. Bavaghar M, Mohajer A, Taghavi Motlagh S (2020) Energy efficient clustering algorithm for wireless sensor networks. *J Inf Syst Telecommun* 4:238–247
8. Piro C, Shields C, Levine BN (2006) Detecting the Sybil attack in mobile ad hoc networks. In: *Proceedings of the Workshop on Secure Communication*, 2006, pp 1–11
9. Ye F, Luo H, Lu S, Zhang L (2005) Statistical en-route filtering of injected false data in sensor networks. *IEEE J Sel Areas Commun* 23:839–850
10. Ishmanov F, Kim SW, Nam SY (2015) A robust trust establishment scheme for wireless sensor networks. *Sensors* 15:7040–7061
11. Ganewaral S, Laura KB, Srivastava MB (2008) Reputation based framework for high integrity sensor networks. *ACM Trans Sens Netw* 4:1–36
12. Zhan G, Shi W, Deng J (2012) Design and implementation of TARF: a trust-aware routing framework for WSNs. *IEEE Trans Dependable Secur Comput* 9:184–197
13. Zhang H, Wang X, Jia C (2017) False data filtering strategy in wireless sensor network based on neighbor node monitoring. *Int J Online Biomed Eng* 13:174–185
14. Han G, Shu L, Ma J, Park JH, Ni J (2010) Power-aware and reliable sensor selection based on trust for wireless sensor networks. *J Commun* 5:23–30
15. Nguyen HT, Thai NH (2019) Temporal and spatial outlier detection in wireless sensor network. *Wiley ETRI J* 41:437–451
16. Wang H, Bah MJ, Hammad M (2019) Progress in outlier detection techniques: a survey. *IEEE Access* 7:107964–108000
17. Yu H, Kaminsky M, Gibbons P, Flaxman A (2008) SybilGuard: defending against Sybil attacks via social networks. *IEEE ACM Trans Netw* 16:576–589
18. Jeyasekar A, Sahil S, Singh R (2020) DnD: filtering false data injection attacks in wireless sensor network. *Int J Adv Sci Technol* 29: 4277–4284
19. Liu J, Labeau F (2018) From wired to wireless: challenges of false data injection attacks against smart grid sensor networks. In: *IEEE Canadian Conference on Electrical and Computer Engineering*, 2018, pp 1–6
20. Wang J, Liu Z, Zhang S, Zhang X (2013) Defending collaborative false data injection attacks in wireless sensor networks. *Inf Sci* 254:39–53
21. Karlof C, Wagner D (2003) Secure routing in wireless sensor networks: attacks and countermeasures. *Ad Hoc Netw* 1:293–315
22. Ssu KF, Wang WT, Chang WC (2009) Detecting Sybil attacks in wireless sensor networks using neighboring information. *Comput Netw* 53:3042–3056
23. Maarouf UB, Naseer AR (2009) Efficient monitoring approach for reputation system-based trust aware routing wireless sensor networks. *IET Commun* 3:846–858
24. Demirbas M, Song YW (2006) An RSSI-based scheme for Sybil attack detection in wireless sensor networks. In: *Proceedings of the International Symposium on a World of Wireless, Mobile, Multimedia Network*, 2006, pp 564–570
25. Mohammad M, Subhash CH (2008) A GTRSSN: Gaussian trust and reputation system for sensor networks. In: *Proceedings of the Advances in Computer and Information Sciences and Engineering*, 2008, pp 343–347
26. Yang M, Zhang Z, Li X, Dai Y (2005) An empirical study of free-riding behavior in the maze P2P file-sharing system. In: Castro M, Van Renesse R (eds) *Peer-to-Peer Systems IV. IPTPS 2005*. LNCS, vol 3640, pp 182–192. Springer, Berlin, Heidelberg. https://doi.org/10.1007/11558989_17
27. Alsaedi N, Hashim F, Sali A, Rokhani FZ (2017) Detecting Sybil attacks in clustered wireless sensor networks based on energy trust system (ETS). *Comput Commun* 110:75–82
28. Newsome J, Shi E, Song D, Perrig A (2004) The Sybil attack in sensor networks: analysis and defenses. In: *Proceedings of the International Symposium on Information Process, Sensor Networks*, 2004, pp 256–268

29. Sarigiannidis P, Karapistoli E, Economides AA (2015) Detecting Sybil attacks in wireless sensor networks using UWB ranging-based information. *Expert Syst Appl* 42:7560–7572
30. Wu R, Deng X, Lu R, Shen Z (2012) Trust-based anomaly detection in wireless sensor networks. In: *Proceedings of the IEEE International Conference on Communications in China, 2012*, pp 203–205
31. Shaikh RA, Jameel H, D’Auriol BJ, Lee H, Lee S, Song YJ (2009) Group based trust management scheme for clustered wireless sensor networks. *IEEE Trans Parallel Distrib Syst* 20:1698–1712
32. Zhang Y, Meratnia N, Havinga P (2010) Outlier detection techniques for wireless sensor networks: a survey. *IEEE Commun Surv Tutor* 12:159–170
33. Yu Z, Guan Y (2010) A dynamic en-route filtering scheme for data reporting in wireless sensor networks. *IEEE/ACM Trans Netw* 18:150–163
34. Yao Z et al (2019) Using hampel identifier to eliminate profile isolated outliers in laser vision measurement. *J Sens* 2019:1–12
35. Suma V (2021) Detection of localization error in a WSN under Sybil attack using advanced DV-Hop methodology. *IRO J Sustain Wirel Syst* 3(2):87–96
36. Mehbodniya A et al (2021) Machine learning technique to detect Sybil attack on IoT based sensor networks. *IETE J Res*. <https://doi.org/10.1080/03772063.2021.2000509>
37. Sefati SS et al (2022) Detecting Sybil attack in vehicular adhoc networks by using fitness function, signal strength index and throughput. *Wirel Pers Commun* 123:2699–2719

An Innovative Novel Method of Reducing the Impact of Traffic Jam Using the Vehicular Ad-Hoc Network



Md. Nahidul Alam, Shahrukh Hossain Rian, Maruf Haider Chowdhury,
Md. Rayhanul Islam, and Mahfuz Ullah

Abstract This paper presents a promising novel method of reducing the traffic jam issue in the country of Bangladesh using the VANET system simulated inside the NS3 software module. Real time traffic data was collected from test locations situated in Cumilla city and used to run the simulation. The street routes were designed in Anylogic software where all the possible vehicles in the street were added to replicate the real time movement of traffic on the streets. The VANET system utilises the trifecta of communication between vehicle to vehicle and roadside unit (RSU) to share valuable information to present the best possible scenario in the road ambience. This paper also proposes a novel data communication method between all the nodes in the VANET system. Overall, the simulation results of the proposed module seem to be very efficient in reducing traffic jam issues, creating a safe and time conserving day -to- day life.

Keywords VANET · NS3 · Test location · Real time · Communication · Anylogic

1 Introduction

Traffic jam has been the by-product of modernisation for decades now. The growing population has not been kind to the limited number of streets that are present in the world. This problem has been more grievous in developing nations with huge populations like Bangladesh. Bangladesh has been on the rise in multiple aspects of industrial growth in the South-East Asia region. The standard of living in this country

Md. N. Alam (✉) · S. H. Rian · M. H. Chowdhury · Md. R. Islam · M. Ullah
Bangladesh Army International University of Science and Technology, Cumilla Cantonment,
Cumilla, Bangladesh
e-mail: nahidul@baiust.edu.bd

S. H. Rian
e-mail: shahrukh.rian@baiust.edu.bd

M. Ullah
e-mail: mahfuz@baiust.edu.bd

has risen dramatically. With the growing need for freight transport across the country, roads have become congested with trucks, cars, and buses. Traffic jam has now spread from the inner city routes to the country's crucial highway connecting roads, creating a whole new degree of problems. The Dhaka-Chittagong highway is considered the busiest and most important highway in the country due to its linkage between the port city of Chittagong and the capital city of Dhaka. In the midst of these two big cities sits the city of Cumilla, the connecting junction of this mega industrial route. As a result, it experiences massive traffic jams that disrupt production in half of the country's industries. This paper focuses on this key point in the country's transportation network and tries to come up with a good way to get around the traffic jams that happen along this route.

This paper's main focus is the crucial junction point on the Dhaka-Chittagong highway called Podduar Bazar Bishwa Road, right outside Cumilla city. This road junction connects the routes of four cities – Dhaka, Chittagong, Laksam, and Cumilla. It also houses a lot of bus stands and other small transport services, all of which congest the highway. On-going and outgoing buses, trucks, and cars get stuck in this area like insects in a spider web. Traffic police alone can't coordinate this huge number of vehicles passing through this area. Another point to be noted is that this route is also very prone to road accidents and the unfortunate gruesome demise of pedestrians. Because of these things, the main goal of this research has been to find a modern solution to this problem and stop people from dying needlessly.

The proposed method utilizes the VANET system to reduce traffic jam issues and create a safe road environment. A Vehicular ad-hoc network, also known as VANET, is a network established through the interconnection of vehicles and roadside units (RSU) present in a given area. This network transmits and receives data to and from the vehicles and RSUs to a central processing unit to present the driver's important information about the desired destination. The VANET system works by sharing data such as real-time positioning, spacing, and speed of the vehicles to predict the occurrence of traffic jams and advise the drivers to choose alternate paths for the most time-efficient drive-through. VANET also maintains the traffic control system by using intelligent protocols derived from the number of vehicles present on each side of the road. There have been other system modules developed in this field. For the purpose of implementing this proposed system, traffic data was collected from the test region. In real life, there was a lot of traffic jam when the data was being collected, and the simulation's exact copy of the traffic jam showed that if the proposed module was used, there would be a big drop in traffic jams. All the previous related versions of such work did not implement any road networks. The novelty of this proposed method is that it provides a defined, visible road network system developed in Anylogic software which shows the vehicle distribution according to the algorithm. It can adapt to any real-life path network and can be easily deployed in any traffic congested area to reduce, if not eliminate, traffic jam.

2 Literature Review

In [1] the authors discussed the traffic jam problem issue using VANET and an ant colony optimization algorithm. This is very efficient for solving the traffic jam issue and also provides the simplest route. The drawback of the paper is that the author did not discuss large-scale vehicle-to-vehicle communication. The system is going to be applicable to an excessive number of cars, which has not been discussed by the authors. Raid et al. [2] discuss load balancing and intersection waiting time reduction using queuing theory. What happened to multiple intersections and busy roads? They did not provide any solutions. In [3], researchers used GPSR and AODV methods to test the traffic light and congestion of any junction, but the drawback is that when the number of vehicles is too high, SUMO and NS2 generate an excessive number of nodes. For each vehicle, IP and MAC addresses are different.

Fahad G. and colleagues [4] deliberate on a strategy if vehicle communication is disrupted by a roadside unit, then communication will be held through a smartphone, but the problem is that it did not discuss the intersection point issue. In [5] researchers discussed a feasible solution to reduce the traffic management service. They did not provide any solution to the data theft issue. Researchers proposed secure communication between vehicles in [6]. They also proposed an algorithm where they used a weighted variable concept to predict the most viable path and less traffic density. The drawback of that research is that they did not test their data in real time. In real time, the communication between vehicles each second creates a lot of nodes.

In [7], researchers discussed a promising method to overcome the traffic jam issue using an automatic green channel facility in the case of heavy traffic flow. Researchers also did not test their simulation in real time, as well as did not compare test data to real time data. In [8], the authors talked about a model for using V2V communication to cut travel time in real time. The problem with their work is that they can't figure out the latency limit and network load. In [9] authors discussed an automated traffic control in the unmanned railway gates in India. In [10] the authors tried to make a CCTV-based surveillance system to stop road violence and keep an eye on traffic. In [11] authors presented a real-time abnormal traffic data dissemination protocol. In [12] authors discussed a new method for clustering in VANET. In [13] the authors did a survey on how important RSUs are in VANET and how they can be used with the new 5G standards. In [14] authors provided an overview of a complete VANET system. In [15] authors presented a dynamic vehicular path planning solution in VANET which improves the overall spatial utility and travel cost.

3 Methodology

In this section, the step-by-step working procedure of the proposed model will be explained. The first section will go over the SUMO software. SUMO is used to create the road network. Real time locations are selected from the Google map and vehicle

numbers and types are also inserted using SUMO. The second section will go over the formation of algorithms in Anylogic software.

3.1 Replicating Practical Scenario in SUMO

The practical test region of Podduar Bazar Bishwa Road is selected using the software SUMO shown in “Fig. 1.”

This now presents the main work space of the proposed simulation. The vehicle numbers and types were collected from a real-time survey of the test region for 300 s as shown in “Fig. 2.”

The number of different types of vehicles were inserted into the SUMO software manually as shown in “Fig. 3.”

With all of this data, SUMO can now generate a simulation of vehicles moving in the test region workspace shown in “Fig. 4.”

Using the NS3 software, a node is now generated for each and every single vehicle moving inside the simulation region. These nodes all have a unique IP and MAC address. These IP and MAC addresses are used for data communication. Each node has a circular network coverage range. Each vehicle can communicate with every

Fig. 1 Selection of the test region in SUMO



Fig. 2 Real time vehicle distribution



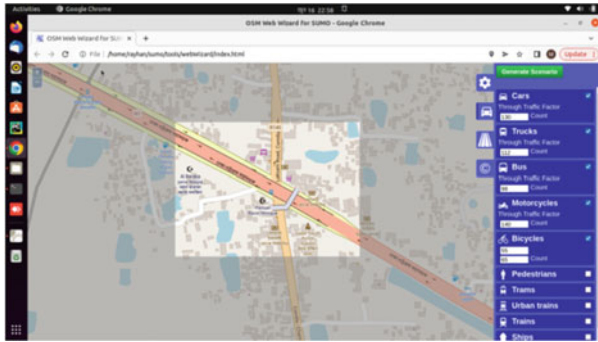


Fig. 3 Inserting vehicle data in SUMO

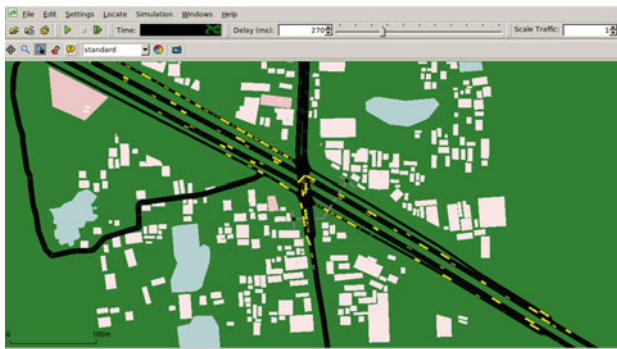


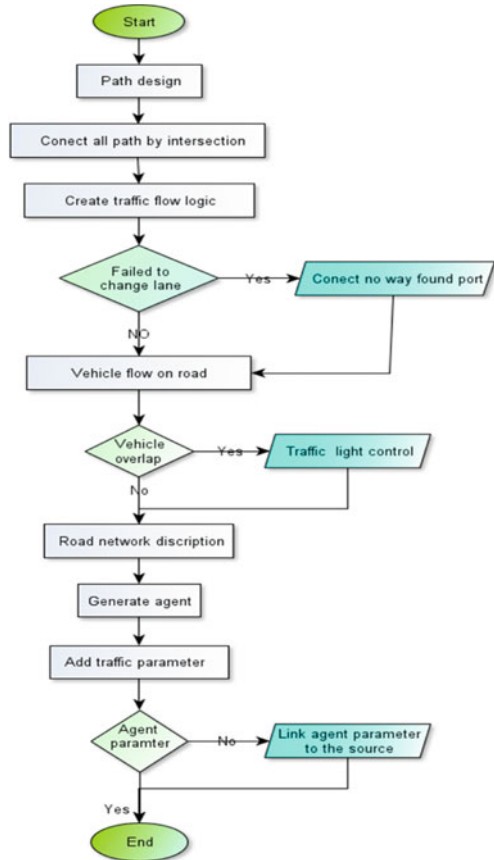
Fig. 4 Vehicle movement simulation

single other vehicle present in its corresponding network range. As the vehicles are always moving, this mesh network is always changing.

3.2 Assembling the Logic Algorithm

Using the Anylogic software, the four practical roads were digitally built inside the proposed module. Now, the traffic flow algorithm is built inside the Anylogic software shown in “Fig. 5”. There are four inter-section points in this design, which connects the four roads. The traffic flow logic block determines where the cars should enter the road system. The lane change command must be entered into the logic portion of the simulation. “Connect no way found port” is a function that decides whether or not to alter the vehicle’s lane. If it fails to change lanes and is ‘positive,’ it will link to the “no way found port,” however if it is ‘negative,’ vehicles will move on the road. Vehicles can overlap and run constantly in simulation. However, this does

Fig. 5 Traffic flow algorithm



not happen in real life. The “Road network descriptor” function is used to display vehicle numbers that are part of the current road network, as well as the speed limit for a specific sector of the traffic network.

Then agents must be created to add network traffic parameters. This network has three parameters: speed, car type, and time. When the road is selected, the agent parameters will work. If the agent parameter is ‘negative,’ it will go to the source’s link agent parameter, and if it is directly ‘positive,’ it will go to the end.

4 Proposed Path Algorithm

The path algorithm that operates the proposed module was built using the Anylogic software. As there are 4 intersectional roads considered for the test research, the logic algorithm has to be made for each of these 4 respective roads. First, the road that

connects to the capital city of Dhaka is considered. From the Dhaka road, vehicles have 3 possible routes to go to – Chittagong, Laksham, and Chadpur. This probability function is generated in the Anylogic software shown in “Fig. 6.”

Here 2 functions, - “carMoveTo” and “carDispose”, are used. The “carMoveTo” function is in charge of moving the vehicles along the simulation’s routes, while the “carDispose” function basically disposes of the vehicles at the end of the road simulation to simulate a seamless practical road scenario.

The same logic algorithm is created for each of the other 3 roads (Chittagong, Laksham and Chadpur) shows in “Fig. 7”, “Fig. 8” and “Fig. 9.” respectively.

Fig. 6 Vehicle movement algorithm for Dhaka



Fig. 7 Vehicle movement algorithm for Chittagong



Fig. 8 Vehicle movement algorithm for Laksham



Fig. 9 Vehicle movement algorithm for Chandpur

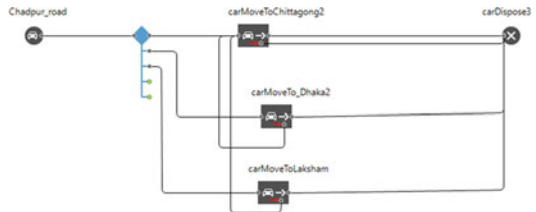
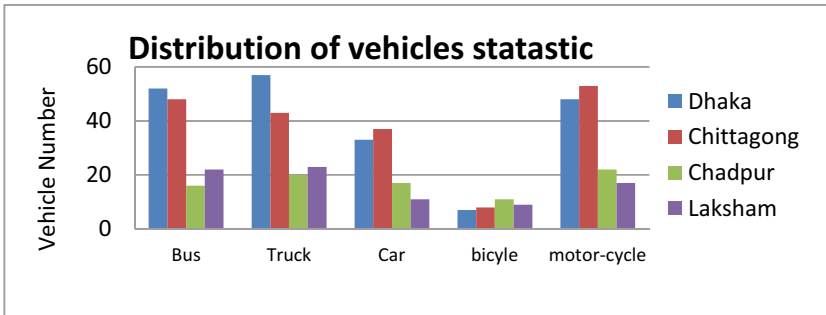


Table 1 Distribution of vehicle types and numbers among the four roads

Road name	Bus	Truck	Car	Bicycle	Motor-cycle
Dhaka	52	57	33	7	48
Chittagong	48	43	37	8	53
Chadpur	16	20	17	11	22
Laksham	22	23	11	9	17

**Fig. 10** Real time traffic flow statistics on the roads

5 Implementation and Data Sheet

5.1 Peak Time Vehicle Data and Statistics

The practical vehicle data is shown in Table 1. Here, vehicle types and distribution among the 4 roads is presented.

The distribution of vehicle statistics and the vehicle frequency in the four roads is shown in “Fig. 10” and “Fig. 11” respectively.

5.2 Proposed Module Circuit Implementation

The circuit of the proposed module is implemented in Proteus 8 software. The main components required to implement this circuit are as follows - Arduino which serves as the main CPU, 16×2 LCD display and transmitter & receiver module. The transmitter and receiver module will be inside the vehicle and once each vehicle is within the range, they will start communicating between them based on the implemented logic algorithm. The circuit diagram of the proposed module is shown in “Fig. 12.”

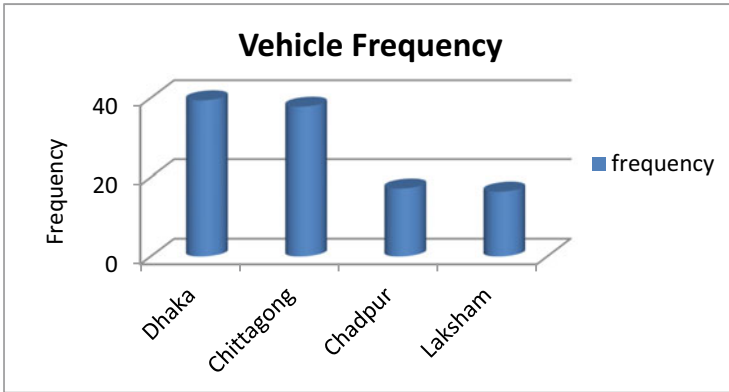


Fig. 11 Individual roads frequency at peak hour

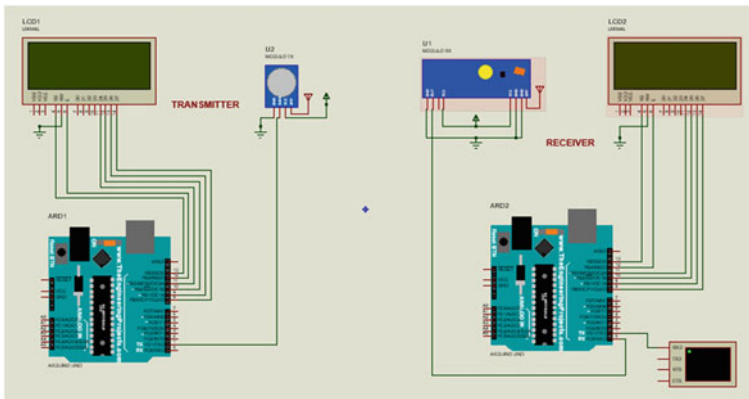


Fig. 12 Proposed prototype model circuit diagram

6 Results and Discussion

6.1 Result for Anylogic Software

The vehicle movement simulation is conducted inside the Anylogic workspace. The 4 intersectional roads are constructed inside the Anylogic software as shown in "Fig. 13."

When the proposed algorithm is deployed, the simulation shows a traffic jam free, smooth movement of vehicles going through all four roads as shown in "Fig. 14." Here, the colour of the roads indicates the density of vehicles.



Fig. 13 Design outcome of entire road network



Fig. 14 Simulated scenario for the selected area

6.2 Result for NS3 Software

Now, the communications between the vehicles are produced using the NS3 software. NS3 software formulates a single node for each of the vehicles inside the workspace. Each of these nodes consists of a unique ID and MAC address as shown in “Fig. 15.”

These IP and MAC addresses enable communication between vehicles for each node. The MAC address ensures that the actual addresses of the nodes are unique. An IP address is a node’s logical address that is used to uniquely identify a node connected to a network. These IP and MAC addresses move in packets between the vehicles and can be seen in the flow monitor of the NS3 software. The proposed module’s IP and MAC address flow is depicted in “Fig. 16.”

The flow monitor module’s goal is to provide a flexible system to measure the performance of network protocols. Network flow monitoring is the collection, analysis, and monitoring of traffic segments. It provides more information like packet data, delay time, etc.

Now, the vehicles are represented as dots and are presented in NetAnim. NetAnim is a stand-alone program that uses the custom trace files generated by the animation

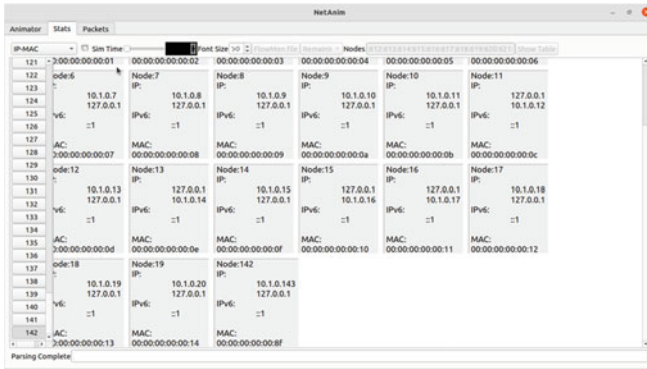


Fig. 15 Generated IP and MAC address for particular vehicles

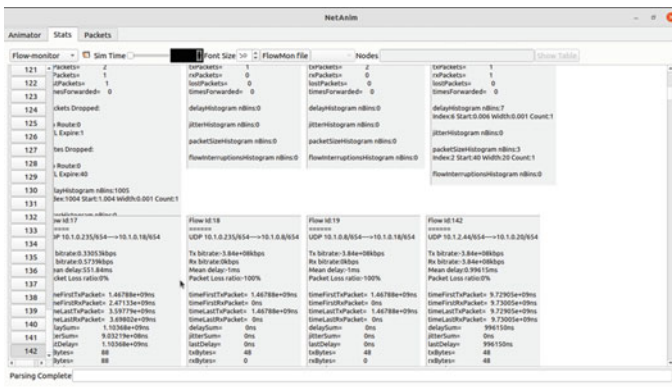


Fig. 16 Generated flow ID of each node

interface to graphically display the simulation. Here, nodes have been created for each vehicle. Since the work has been done with many vehicles, the nodes here are seen to be overlapping. These nodes are constantly on the move, so the response time for the movement of the nodes is set at a per second rate in the simulation calculation. When all the vehicles are converging at the intersection point, nodes seem to be congested at one point, as shown in “Fig. 17.”

When the vehicles move in all four direction from the intersection, the nodes seem to be scattered as shown in “Fig. 18.”

A separate mesh is generated for each node. These generated meshes are connected to each other at nodes. Each node directly communicates with another node. Here, centers are being created in different nodes from time to time. The network continuously moves and sends data to each vehicle on the road network. “Figure 19” shows the mesh network generated when cars converge at the intersection point.

“Figure 20” shows the mesh network when vehicles are scattered.

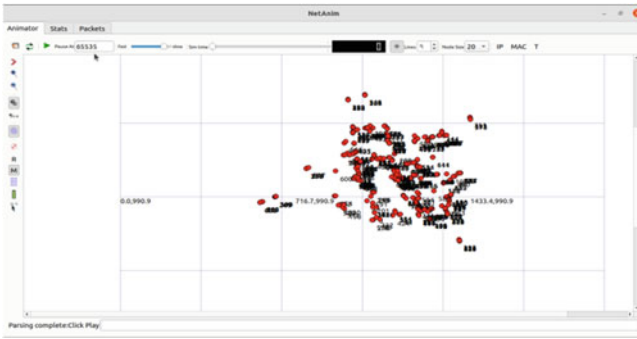


Fig. 17 Generated output NetAmin file in intersection point



Fig. 18 Generated output NetAmin file in scattered point

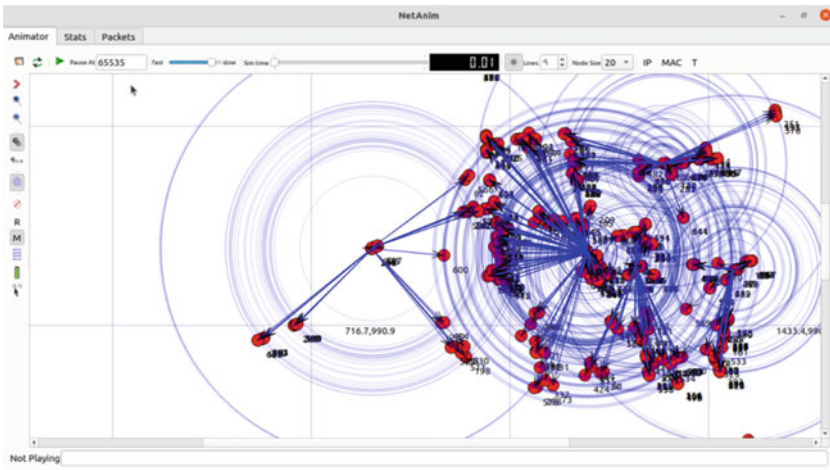


Fig. 19 Generated output mesh network in intersection point

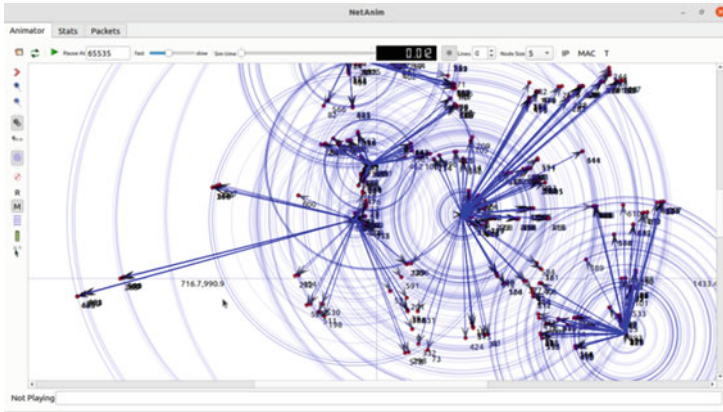
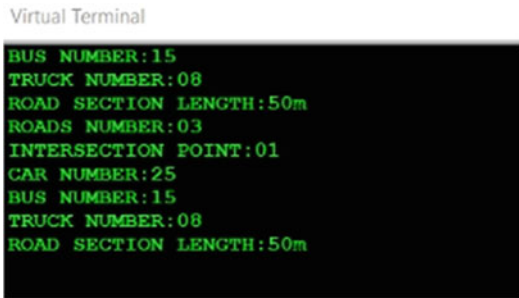


Fig. 20 Generated output mesh network in scattered point

Fig. 21 Generated output of virtual terminal



6.3 Result for Proteus 8 Software

The output of the implemented circuit combined with the logic flow diagram is shown in “Fig. 21.”

7 Conclusion

This paper presents an innovative approach to solving traffic jam issues in densely populated intersection roads in highways using the VANET. Real-time practical road ambience was recreated using SUMO and NS3 software and the logic algorithm used to run the proposed module was developed inside the Anylogic software. The proposed module showed promising results in solving the jamming issues observed in real-time scenarios. The proposed module circuit was implemented using the Proteus 8 software and progress is already on the way to fabricate it practically as a

plug and play device to be connected with any type of vehicles. A more defined and versatile central processor is also in early stages of development to ensure the best performance of the proposed device in highways as well as in urban scenarios. We believe the proposed module can add a new dimension to the traffic control system and pave the way to a safe and comfortable travel environment across the country. Most of the researchers have discussed in their research either protocol design or the communication between vehicles and RSU. The performance of their system is good but not efficient. In our proposed method, we discussed the path logic algorithm and the road network with individual vehicle IP and MAC addresses. The algorithm is user-friendly and the efficiency is also good in our proposed system.

References

1. Ferdous F, Mahmud MS (2016) Intelligent traffic monitoring system using VANET infrastructure and ant colony optimization. In: 2016 5th International Conference on Informatics, Electronics and Vision (ICIEV), May 2016, pp 356–360
2. Zaghal R, Thabatah K, Salah S (2017) Towards a smart intersection using traffic load balancing algorithm. In: 2017 Computing Conference, London, July 2017, pp 485–491
3. Gupta P, Singh LP, Khandelwal A, Pandey K (2015) Reduction of congestion and signal waiting time. In: 2015 Eighth International Conference on Contemporary Computing (IC3), Noida, India, August 2015, pp 308–313
4. Abdulkadhim FG, Yi Z, Tang C, Onaizah AN, Radie AH (2020) Optimizing roadside via unit deployment mechanism in VANET. In: 2020 3rd International Conference on Engineering Technology and its Applications (IICETA), Najaf, Iraq, September 2020, pp 144–149
5. Wang X et al (2019) Optimizing content dissemination for real-time traffic management in large-scale internet of vehicle systems. *IEEE Trans Veh Technol* 68(2):1093–1105
6. Bhargava S, Prakasha K, Sinha I (2017) Predicting traffic density and increasing fuel efficiency in vehicles using secure vehicular networks. In: 2017 International Conference on Computer Communication and Informatics (ICCCI), Coimbatore, January 2017, pp 1–3
7. Jayaraj V, Hemanth C (2015) Emergency vehicle signalling using VANETS. In: 2015 IEEE 17th International Conference on High Performance Computing and Communications, 2015 IEEE 7th International Symposium on Cyberspace Safety and Security, and 2015 IEEE 12th International Conference on Embedded Software and Systems, August 2015, pp 734–739
8. Lebre M-A, Mouel FL, Menard E (2020) Efficient vehicular crowdsourcing models in VANET for disaster management. In: 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), Antwerp, Belgium, May 2020, pp 1–5.
9. Anand JV (2020) Automatic traffic control technologies for remote monitoring of unmanned railway gates. *J Electron Inform* 2:30–37
10. Dhaya R (2020) CCTV surveillance for unprecedented violence and traffic monitoring. *J Innov Image Process* 2:25–34
11. Chen W (2010) VANETs-based real-time traffic data dissemination
12. Hassan H, Hasan Z, Taie R (2022) A simulation approach to improve the VANETs communication. *Int J Interact Mob Technol IJIM* 16:137–144
13. Karunathilake T, Förster A (2022) A survey on mobile road side units in VANETs. *Vehicles* 4:482–500
14. Mahi M et al (2022) A review on VANET research: perspective of recent emerging technologies. *IEEE Access* 1
15. Jayapal C (2022) Realtime congestion avoidance using vanet

Navbot—College Navigation Chatbot Using Deep Neural Network



M. Sobhana, A. Yamini, K. Hindu, and Y. L. Narayana

Abstract In terms of size, college campuses are massive. It is an uphill task to identify in which block a particular room is located or where the block is located within the campus. A chatbot has been developed to address this problem. A chatbot is an interactive artificial intelligence program that attempts to mimic human behavior by interpreting input and responding appropriately in text format. Deep learning model and natural language processing algorithms are used to develop the chatbot. The knowledge base for this chatbot will be in JavaScript Object Notation (JSON) format. Users can request college navigation-related queries. The query is processed using natural language processing techniques such as tokenization and lemmatization. The processed query after spell correction is given as an input to the Deep Neural Network (DNN) algorithm. The proposed chatbot will search the processed query in the knowledge base and respond with the corresponding answer using a sequential DNN model with five hidden layers. User interface of the chatbot is developed using Hyper Text Markup Language (HTML), Cascading Style Sheets (CSS) and Java Script. The proposed model will help in navigating people inside the college to different blocks through google map links and inside the blocks through textual directions. The model works with an accuracy of 98%.

Keywords Deep learning · Artificial intelligence · Natural language processing · Chatbot · Navigation

M. Sobhana · A. Yamini (✉) · K. Hindu · Y. L. Narayana
Department of Computer Science and Engineering, V R Siddhartha Engineering College,
Vijayawada, Andhra Pradesh, India
e-mail: yaminiaravapalli@gmail.com

M. Sobhana
e-mail: sobhana@vrsiddhartha.ac.in

K. Hindu
e-mail: 198w1a0586@vrsiddhartha.ac.in

Y. L. Narayana
e-mail: 198w1a0592@vrsiddhartha.ac.in

1 Introduction

A chatbot is a software application that allows a machine and a person to converse in the same way as people do. Nowadays chatbots are becoming very popular as they save time and respond quickly and efficiently to the questions asked by the users. Machine learning and artificial intelligence algorithms are used in developing conversations [1]. Chatbots are being used in many applications like public administrations, businesses, and non-profit organizations.

Natural Language Processing (NLP) is a process of making the machine understand the text or spoken words of the human and make the machine produce a meaningful response [2]. Tokenization and lemmatization are some of the techniques of natural language processing. Tokenization is the process of splitting the text into smaller pieces or tokens. Lemmatization is the process of eliminating inflectional endings from a word to get the base or glossary form of a word known as a lemma.

A deep neural network comes under the class of machine learning algorithms which can contain multiple hidden layers between the input and output layers [3]. The proposed model has five hidden layers. The first hidden layer is the embedding layer. It is used for converting words into vectors of required dimensions. A 16-dimensional vector has been created for each word in the vocabulary. The second layer is the Global Average Pooling1d layer. It is used instead of the flattening layer to reduce the computational complexity. Three dense layers are added with 64, 16, and 16 nodes respectively to improve the efficiency of the proposed model. Rectified linear unit (ReLU) activation is used in the dense layers. This activation function works on the logic $y = \max(0, x)$. For all values less than zero the output will be zero and for values greater than zero output is the value. The final layer is the output layer which is a dense layer with 70 nodes which is equal to the number of classes. A softmax activation function is used in the output layer. Deep neural network architecture is displayed in Fig. 1.

An “intent” indicates the idea the user has in mind when typing any query [4]. In the proposed methodology the database has been augmented with 70 intents to provide a response to the user navigation-related queries.

The following paper is organized as follows. The literature survey is discussed in the second section, the proposed model is explained in detail in the third section, the results are displayed in the penultimate section, and the final section concludes the paper.

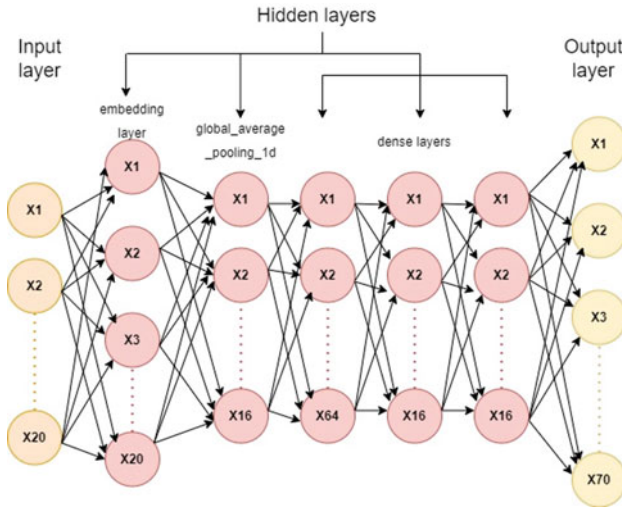


Fig. 1 Deep neural network layers

2 Related Works

C. Schultdt et al. [5] has introduced the level 5 Indoor Navigation project (L5IN). He suggested an indoor navigation technique just by using the 5G mobile orientation signal without additional infrastructure. He investigated how existing exterior route frameworks can be utilized for indoor navigation using the smart mobile 5G technology.

E. Lewandowicz et al. [6] proposed the Medial Axis Transformation (MAT) algorithm to improve the methods for developing indoor navigation models. The Node-Relation Structure (NRS) approach is used to generate corridor axes. The spots where the centre lines cross the modelled structure will be used to determine the structure’s axis. The suggested technique comprises a different approach to corridor space division. The method for creating Triangulated Irregular Systems through Delaunay triangulation is utilized in the conventional approach.

Rubio-Sandoval J. I. et al. [7] proposed an indoor navigation system based on augmented reality and semantic webs. He gave a detailed methodology to build. This was achieved in four steps. In the first step spatial modeling is done. In this they represented different places and points of interest. The other steps are data management and structuring, Positioning and Navigation. The final step is content visualization done through mobile application.

D. Khan et al. [8] suggested an indoor navigation framework that is less expensive and scalable. They employed basic markers written on paper on the building roofs. These markers are distinguished using the camera of a smartphone. The sound-related and visual information related to these markers are used to direct the client. This framework decides the shortest route between two markers. New markers can also be added as per requirement. This procedure helps in coordinating individuals, visitors, and newcomers in an inside environment.

V. Oguntosin et al. [9] worked on building Hebron, a web-based chatbot. This chatbot was built with Python and React.js, with MySQL as the database, to provide an overview of the e-commerce datasets as well as the Admin Portal procedure. This chatbot was developed for the Covenant University Community Mall. Recast.ai api was used to create, train and monitor the performance of the chatbot.

T. Nguyen et al. [10] introduced an AI-based chatbot through which students may receive daily updates on curriculum, new student admission, tuition rates, IELTS writing problem II score, and other topics. Rasa framework was used to create this chatbot. It is designed to support Vietnamese and English languages. Rasa BERT pipeline was used as it supports both languages instead of the convert pipeline which supports only English. It works with an accuracy of 97.1%. The chatbot was used on facebook for the official affirmation fan page for social organizations.

Villegas-Ch et al. [11] suggested adding artificial intelligence to the existing smart class infrastructure which improves the learning abilities of the students. He suggested adding a chatbot to the learning management system. The chatbot should have access to each individual's credentials which helps in authorizing. This chatbot can remind the student about the pending activities or achievements in the previous activities. Natural language processing, Natural language understanding and machine learning are used to build this chatbot.

S. Patil et al. [12] suggested a new crossbreed Long Short Term Memory-based Ensemble model for storing data in various contexts. It is important to store data in conversative applications to build the conversation. They discovered that both LSTM and GRU, which are ensemble techniques, are effective, perform well in a variety of dataset contexts and that techniques are particularly useful in chatbot applications.

Saraswat S. et al. [13] proposed Galgobot, a chatbot system that can be connected to college websites. To prevent an unauthorized user from acquiring confidential information, the system includes a Login and Signup System webpage. The program also used a Natural Language Processing (NLP) model to ask numerous inquiries to obtain exact responses to the inquiry. This chatbot used HTML, CSS, Ajax, JavaScript, jQuery for frontend and PHP, and Python for the backend. It uses the Rasa NLU Core framework.

Nithuna S. et al. [14] researched a variety of chatbot development tools are available on the market. But they lack the flexibility and agility needed to develop genuine discussions. Microsoft Cortana and Google Assistant are among the most popular intelligent personal assistants. These chatbots have constraints. Most of the chatbots built are rule-based, machine learning algorithm-based, or retrieval-based. But these didn't provide satisfactory results. In this study, he analyzed the critical sections of chatbots with the current technologies being applied to develop.

Koundinya H. et al. [15] studied the ways Artificial Intelligence is being utilized to improve many features focused on developing chatbots. A chatbot based on artificial intelligence and natural language processing was developed for this project where users can communicate with the college chatbot using natural language input, and the chatbot can be trained using relevant machine learning approaches to reply correctly. It used wordnet to find the matching response by finding the most closely matched input from the database.

Karri S. P. R et al. [16] studied different technologies and methods used by different people to develop chatbots. NLTK is a python package that can do natural language processing. It is capable of processing human voices and producing responses like humans. Some highly demanded virtual assistants are Google Assistant, Cortana, Siri, and Alexa. chatbots are primarily used in the commercial market to assist customer support and reduce human work. The bag of words technique was discussed to process the input.

Smys S. et al. [17] proposed a naïve bayes classifier to identify the difference between a human user and chatbots. The main idea of it is to stop the spread of malware and spam through this chatbots. They have collected samples of chats form different users. Entropy classifier and naïve bayes was used to identify between a chatbot reply and human reply. Factors like inter message delay were considered to classify.

Sungheetha A. et al. [18] proposed a way of assimilating iot sensors in the smart campus environment. The idea is to store the data of the sensors and make it visible in the smart campus environment interface. Through this students and faculty can know about the factors of indoor environment like temperature, humidity etc. DHT11 humidity and temperature sensor was used.

3 Proposed System Diagram

The proposed system college navigation chatbot is used to solve the queries asked by the users regarding directions of various blocks inside the campus. The chatbot makes use of HTML and CSS for the front end and a trained sequential model for the backend.

- The chatbot welcomes the user to ask the queries.
- The user enters the query, the corresponding query is processed.
- The model does the spell correction and identifies the suitable intents and entities from the database and predicts the best response using deep neural network.
- The response is given in the format of textual directions inside the block and google map links to navigate them outside the block.
- The conversation is continued till the user is satisfied with the response given. Figure 2 shows the flow diagram of the system.

The proposed system takes the users query as input and processes it. It then recognizes intents and entities, creates a response, and returns it to the user.

3.1 Database

The database stores various intents and entities to solve the queries asked by the users. In the proposed methodology the database has been augmented with 70 intents to provide a response to the user navigation-related queries. These intents have been prepared to mention all the directions required to direct people. Dataset was manually prepared to address the navigation queries of a specific college. It can be shared on request by the authors. The intent is divided into three parts such as tag, pattern and response. Tag helps us in classifying the query. Message patterns are used to match sections of user messages. When the query is matched then the response will be displayed to the user.

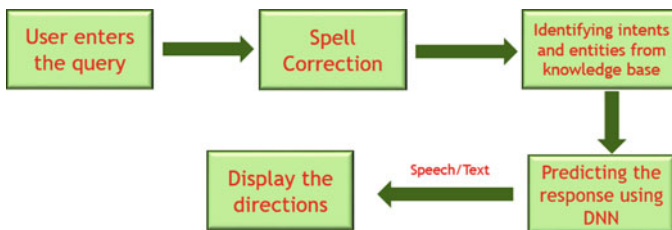


Fig. 2 Chatbot methodology

3.2 Preprocessing

Firstly the labels are converted into the form of lists and performed the encoding using the label encoder present in sklearn.preprocessing package. The chunks of text are divided into tokens by tokenization. The processing of the text is done after performing the tokenization process. By using the word_index method we assign a value to each word. Now by using a tokenizer.texts_to_sequences method each sentence is converted into numbered form. When the user enters the query the text is converted into lower case and spell correction is performed. The spell correction is performed by comparing every word with correctly spelled words and if the word matches then the related word with the close probability will be produced.

3.3 Model Development

The sequential model with five hidden layers is used. The model was trained on the patterns and labels. In Fig. 3 architecture diagram of the chatbot is presented.

The data pre-processed using tokenization and lemmatization technique is fed into the proposed model for training. The model is trained on labels and patterns to predict the correct label for a given pattern. Later, the query typed by the user is pre-processed and spell checked and sent to the trained sequential model. The deep neural network sequential model block in the architecture predicts the label

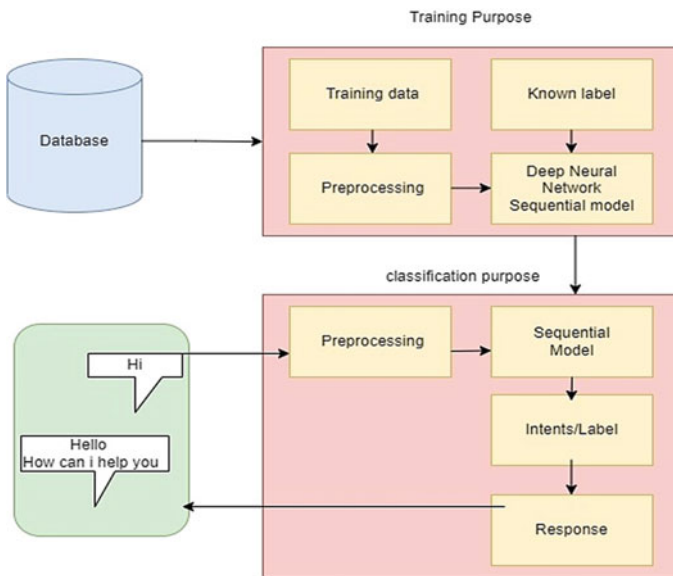


Fig. 3 Chatbot architecture

belonging to the query and generates a response to the user based on the predicted label. If the model fails to predict a suitable label it responds with a predefined answer informing the user to ask related query.

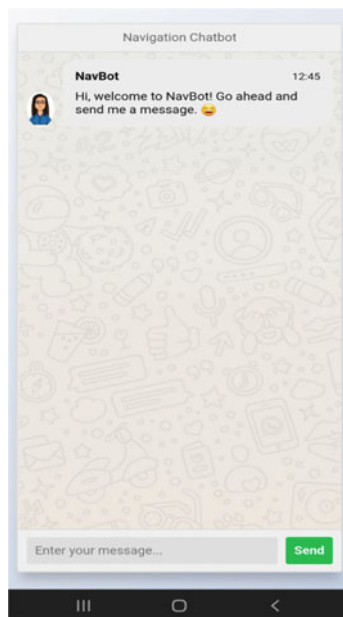
3.4 Algorithm for Model Training

The knowledge base containing the intents is loaded. The data obtained from this file is pre-processed. For training purpose we will store the all the tags of the intents in a list called labels. After converting the labels and patterns into numerical form train the sequential model. Model was trained on 80% of the data and tested on 20% data. Repeat this training until the model gives satisfactory result.

- Step 1: Start
- Step 2: Load the intents.json file
- Step 3: Perform data cleaning
- Step 4: Store the tags of the intents
- Step 5: Convert the labels and patterns into numerical form
- Step 6: Create a sequential model
- Step 7: Add the 5 hidden layers with ReLU activation
- Step 8: Compile the model
- Step 9: Train the model on labels and patterns
- Step 10: Now test it with some input.
- Step 11: Stop

3.5 Integration

Flask web framework is used to integrate the frontend of the chatbot with the trained model. All the html templates that are to be rendered are put in the templates folder and the cascading style sheets are put in the static folder.

Fig. 4 Navbot mobile view

4 Results

Screens of the developed chatbot are presented in Figs. 4 and 5. It is the mobile view of the chatbot. A web version of the chatbot is presented in Figs. 7 and 8. The directions link is redirected to google maps as shown in Fig. 6. In Fig. 8 spell correction feature of the chatbot is presented.

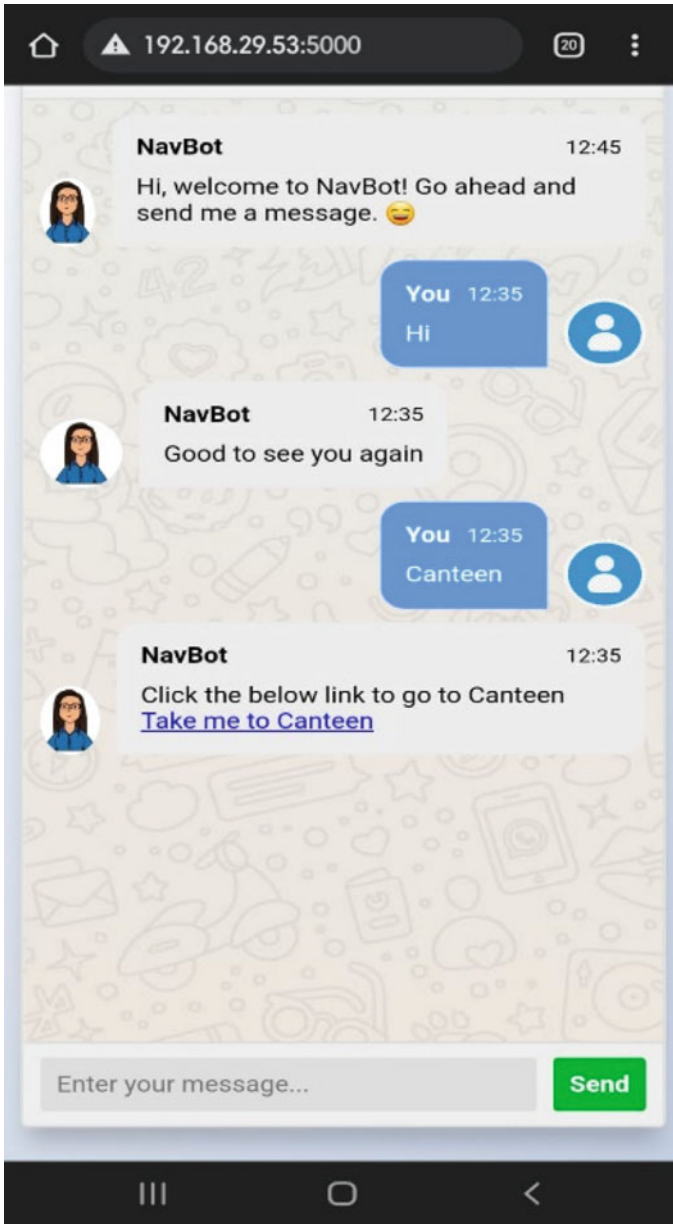


Fig. 5 Chat with Navbot

Fig. 6 Redirected to google maps

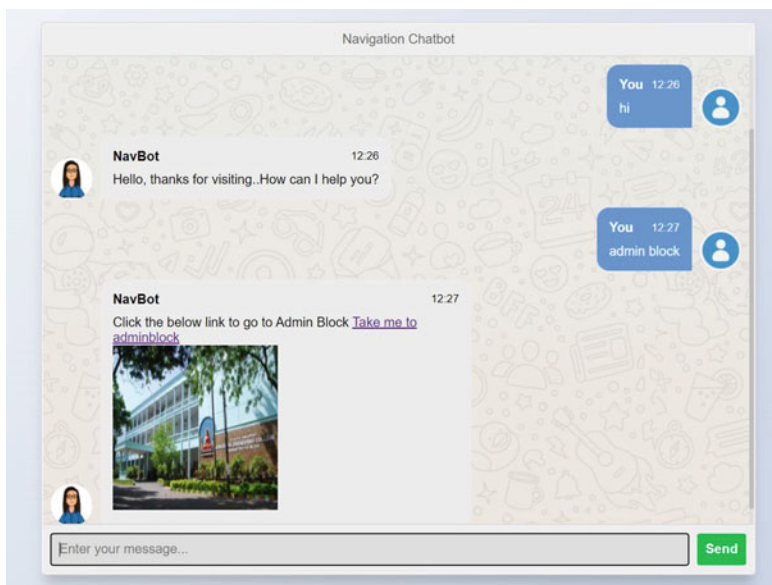
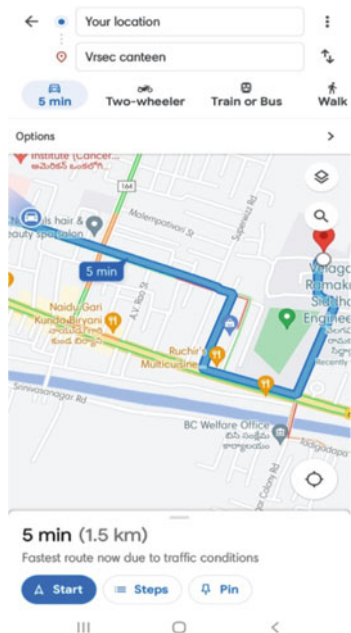


Fig. 7 Web view of Chatbot

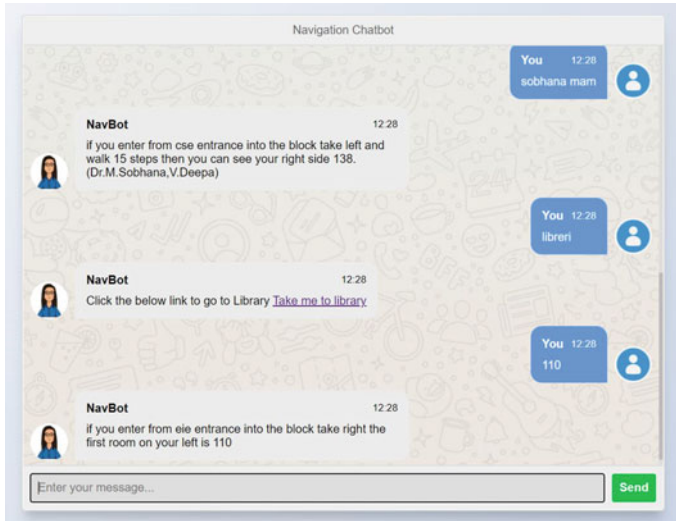


Fig. 8 Spell correction in Chatbot

5 Conclusion and Future Work

This navigation chatbot developed using a deep neural network with an automatic spell correction feature helps students and visitors in college to identify different blocks present on the campus. Directions and pictures are provided for user convenience. It provides directions to different rooms present in a block. Faculty rooms information is also included. This chatbot works 98% accurately when trained with 150 epochs. In future, we can include language options and voice features.

References

1. Arel I, Rose DC, Karnowski TP (2010) Deep machine learning-a new frontier in artificial intelligence research [research frontier]. *IEEE Comput Intell Mag* 5(4):13–18
2. Mathew L, Bindu VR (2020) A review of natural language processing techniques for sentiment analysis using pre-trained models. In: 2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC), pp 340–345. IEEE, March 2020
3. Jaiwai M, Shiangien K, Rawangyot S, Dangmanee S, Kunsuree T, Sa-nguanthong A (2021) Automatized educational Chatbot using deep neural network. In: 2021 Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunication Engineering, pp 5–8. IEEE, March 2021

4. Deepika K, Tilekya V, Mamatha J, Subetha T (2020) Jollity Chatbot-a contextual AI assistant. In: 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), pp 1196–1200. IEEE, August 2020
5. Schuldt C, Shoushtari H, Hellweg N, Sternberg H (2021) L5in: overview of an indoor navigation pilot project. *Remote Sens* 13(4):624
6. Lewandowicz E, Lisowski P, Flisek P (2019) A modified methodology for generating indoor navigation models. *ISPRS Int J Geo Inf* 8(2):60
7. Rubio-Sandoval JI et al (2021) An indoor navigation methodology for mobile devices by integrating augmented reality and semantic web. *Sensors* 21(16):5435
8. Khan D, Ullah S, Nabi S (2019) A generic approach toward indoor navigation and pathfinding with robust marker tracking. *Remote Sens* 11(24):3052
9. Oguntosin V, Olomo A (2021) Development of an e-commerce chatbot for a university shopping mall. *Appl Comput Intell Soft Comput* (2021)
10. Nguyen TT, Le AD, Hoang HT, Nguyen T (2021) NEU-Chatbot: chatbot for admission of national economics university. *Comput Educ Artif Intell* 2:100036
11. Villegas-Ch W, Arias-Navarrete A, Palacios-Pacheco X (2020) Proposal of an architecture for the integration of a Chatbot with artificial intelligence in a smart campus for the improvement of learning. *Sustainability* 12(4):1500
12. Patil S, Mudaliar VM, Kamat P, Gite S (2020) LSTM based ensemble network to enhance the learning of long-term dependencies in chatbot. *Int J Simul Multidiscip Des Optim* 11:25
13. Saraswat S, Mishra S, Mani V, Priya S (2021) GALGOBOT—the college companion Chatbot. In: 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), pp 1374–1378. IEEE, May 2021
14. Nithuna S, Laseena CA (2020) Review on implementation techniques of chatbot. In: 2020 International Conference on Communication and Signal Processing (ICCSP), pp 0157–0161. IEEE, July 2020
15. Koundinya H, Palakurthi AK, Putnala V, Kumar A (2020) Smart college chatbot using ML and python. In: 2020 International Conference on System, Computation, Automation and Networking (ICSCAN), pp 1–5. IEEE, July 2020
16. Karri SPR, Kumar BS (2020) Deep learning techniques for implementation of chatbots. In: 2020 International Conference on Computer Communication and Informatics (ICCCI), pp 1–5. IEEE, January 2020
17. Smys S, Haoxiang W (2021) Naïve Bayes and entropy based analysis and classification of humans and chat bots. *J ISMAC* 3(01):40–49
18. Sungeetha A (2021) Assimilation of IoT sensors for data visualization in a smart campus environment. *J Ubiquitous Comput Commun Technol* 2022(4):241–252

Heart Problems Diagnosis Using ECG and PCG Signals and a K-Nearest Neighbor Classifier



Youssef Toulni, Benayad Nsiri, and Taoufiq Belhoussine Drissi

Abstract At the present, the diagnosis of the cardiovascular diseases knew a great evolution, this progress was the result of the development of the techniques used in the diagnosis, which allowed a rapid and sometimes automatic detection of the symptoms of these diseases. Among these techniques we find the methods which are based on the processing of biomedical signals, the electrocardiogram and phonocardiogram signals (ECG and PCG) are the most used in this field, the difference in the nature of these signals implies a different processing for each of them. this paper proposes the processing of these two ECG and PCG signals from the same person and the extraction of the features of each of these signals by two different ways, the statistical features are extracted from the wavelet coefficients of the ECG signals and the Mel frequency cepstral coefficients were calculated from the PCG signals. the classification of the combination of these two types of features using a KNN classifier showed an improvement in the accuracy compared to the classification of the features of each signal separately.

Keywords Electrocardiogram · Phonocardiogram · Wavelet · Mel frequency cepstral coefficients · Statistical features · K-nearest neighbor classifier

Y. Toulni (✉)

Laboratory of Electrical and Industrial Engineering, Information Processing, IT and Logistics (GEITIIL), Faculty of Science Ain Chock, Hassan II University, Casablanca, Morocco
e-mail: youssef.toulni@gmail.com

B. Nsiri

Research Center STIS, M2CS, National School of Arts and Crafts of Rabat (ENSAM), Mohammed V University in Rabat, Rabat, Morocco

T. Belhoussine Drissi

Laboratory of Electrical and Industrial Engineering, Information Processing, IT and Logistics (GEITIIL), Faculty of Science Ain Chock, Hassan II University, Casablanca, Morocco

1 Introduction

The diagnosis of health problems is now an essential part of the treatment of these problems. Indeed, the importance of an accurate real - time diagnosis is illustrated in the choice of an appropriate therapeutic interventions as well as the efficiency with which it is executed. As a result, researchers are trying to make disease diagnosis more independent and autonomous by using new approaches, especially those aimed at data collection and processing, as well as artificial intelligence techniques [1, 2]. The information gathered comes from a variety of sources, including various forms of signals, which are seen as as a treasure trove of information used for diagnosis.

The identification of signals, particularly biomedical signals, has made considerable progress; this progress is the result of the development of various techniques and methods to obtain the necessary information; a great number of mathematical tools is used in the processing of these signals and the extraction of various types of features that help in their identification [3]. The use of these approaches has greatly aided in the diagnosis of cardiovascular diseases, and it has played an essential role in the development of these techniques.

One of these biomedical signals that help us in the diagnosis is the electrocardiogram. The electrocardiogram, often known as the ECG signal, is a cardiac activity that may be represented by an electrical voltage as a function of time. The ECG signal interprets the various contractions of the heart's parts, which are manifested by the periodic observation of the P wave, the QRS complex, and the T wave[4] (see Fig. 1). Also, sounds wich represent noises and murmurs of the heart can be captured and used in the diagnosis. The variations of this sounds over time is called phonocardiogram (PCG signal), in a phonocardiogram signal there are four types of sounds, the first two sounds S1 and S2 come from the normal functioning of the heart valves, while the other types S3 and S4 are abnormal sounds [5, 6] (see Fig. 1).

The automatic evaluation of cardiac signals (ECG and PCG) using identification approaches provides for the early detection of some cardiac irregularities or diseases employing a simple and non-invasive method without the need for a qualified person [7, 8]. The field of signal processing has proven to be successful in the analysis of biomedical signals over the years, due mainly to the development of new techniques, especially those used for removing noise from a given signal as well as for detecting and extracting features that help us in the identification of the signal [9, 10]. Among these techniques we find the wavelet analysis, which allows a simultaneous analysis of the signal in time and frequency, this method has several uses, for example this analysis makes it possible to locate the noises of high and low frequencies [11, 12], so wavelet analysis helps us locate and extract some features used in signal identification [13, 14]. In addition, the calculation of the cepstral coefficients at the Mel scale (MFCC) has an important role in the field of signal processing and in particular in the recognition of audio signals because these coefficients are considered as features of this type of signals [15, 16].

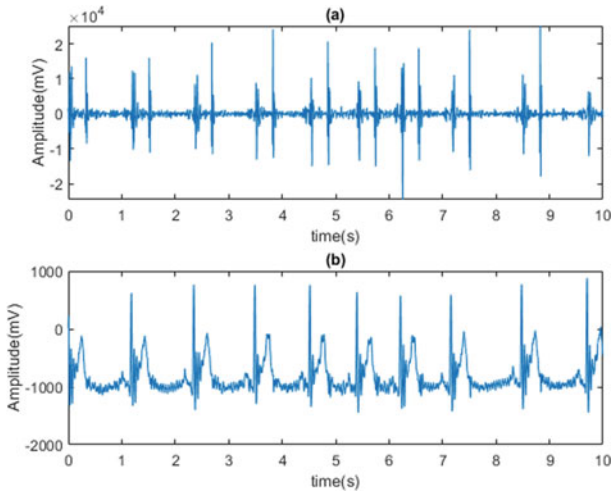


Fig. 1 a PCG signal b ECG signal

This work aims to identify certain cardiac problems of patients examined automatically through the processing of ECG and PCG signals from the same patient using signal processing techniques and classification methods. To do this, we will focus on the features extraction by calculating statistic features from the wavelet coefficients and also, we will compute the MFCC coefficient from the PCG signal, after that all these features will be used to distinguish sick persons from the from the healthy one, in the step of classification, at this stage we will use the KNN algorithm to classify the signals.

2 Proposed Method

The approach adopted in this work is summarized in the diagram shown in Fig. 2; in this paper we will simultaneously use the recordings of the ECG and PCG signals of each patient to extract the features which are used for the classification of the signals, the classification step allows us to distinguish between healthy people and sick people; in the classification phase we will use a KNN classifier.



Fig. 2 Proposed method

2.1 Data Collection

The database used in this work is used in PhysioNet/CinC Challenge 2016, this database contains five training sets classified from A to E which contains 3126 recordings in total [17], the participants in this data source are of all ages and include both healthy and sick people, we will limit ourselves to using only set A as a base of data, this decision is based on the fact that this set contains both the recordings of the ECG and PCG signals of the same patient. Database A contains recordings of 409 people, 405 of whom have recordings of both ECG and PCG signals which are sampled at 2000 Hz.

2.2 Processing and Features Extraction

Wavelet Analysis. Wavelet analysis is a method which allows the analysis of a given signal in time and frequency thanks to the convolution of the signal by a function called wavelet ψ which is concentrated around a frequency and a precise instant and verifies the following condition [4, 18]:

$$\int_{-\infty}^{+\infty} \psi(t)dt = 0 \tag{1}$$

The coefficients a of scaling and b of translation are respectively used to adjust the frequency and central instant, according to the expression [12, 18]:

$$\psi_{a,b}(t) = \frac{1}{\sqrt{a}}\psi\left(\frac{t-b}{a}\right) \tag{2}$$

The wavelet analysis of a given signal $x(t)$ can be done through the calculation of the continuous transform by the following formula:

$$c_{a,b} = \frac{1}{\sqrt{a}} \int_{-\infty}^{+\infty} x(t)\psi^*\left(\frac{t-b}{a}\right)dt \tag{3}$$

Or in a discrete way for practical reasons using the discrete wavelet transform DWT. In practice, this transform consists of calculating the wavelet coefficients by passing a given signal $x(t)$ through a series of high-pass (HPF) and low-pass filters (LPF) [19]; the coefficients extracted by the high pass filters are called the detail coefficients while the coefficients from the low pass filters are the approximation coefficients Fig. 3 shows the decomposition of a signal at three levels.

Wavelet analysis has several functionalities and applications in the field of signal processing, among the applications of this analysis technique we find the denoising and the extraction of the features which make it possible to identify the analyzed signal, where the interest of this technique in signal processing and in particular non-stationary signals [20, 21].

Mel Frequency Cepstral Coefficients MFCC. Widely used to identify audio signals, the cepstral coefficients at the Mel scale are considered as coefficients that give the opportunity to characterize this type of signals [22]. This technique was designed so that the recognition of the signal by the MFCC coefficients is similar to the recognition by the human ear of various audio signals, which explains the use of the Mel scale. The calculation of the MFCC coefficients follows the following steps (see Fig. 4) [23, 24].

Fig. 3 DWT at the 3rd level of scale

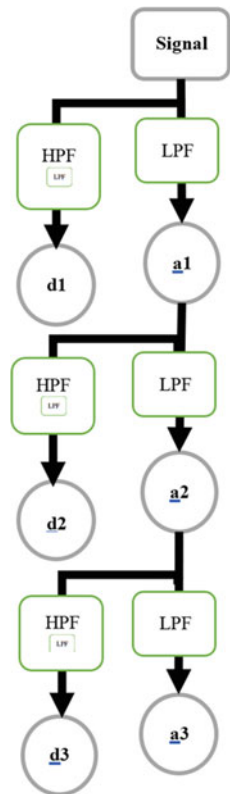


Fig. 4 MFCC calculation steps



Pre-emphasis. This step emphasizes the high frequency domain of the signal. To do this, the signal x_n is introduced into a first-order finite impulse transfer function filter:

$$H = 1 - kz^{-1} \tag{4}$$

Thus, the signal x'_n obtained after the pre-emphasis will be:

$$x'_n = 1 - kx_{n-1} \tag{5}$$

The coefficient k represents the pre-emphasis coefficient which is often between 0.9 and 1 [16, 24].

Segmentation and windowing. Segmentation consists of dividing the signal into frames of limited durations which vary between 40 and 60 ms, signal segmentation is used to easily apply signal processing techniques such as DFT on non-stationary signals. Also, the problem of the discontinuity that can occur between the different

frames of the signal can be solved by overlapping from 10 ms up to 30 ms and a windowing between these frames by multiplying the frames of the signal by the Hamming window $w(n)$ of expression:

$$w(n) = 0.54 - 0.46 \cos\left(\frac{2\pi n}{N-1}\right) \quad (6)$$

N is the number of samples in each frame. Thus, the windowed signal x_n'' will be:

$$x_n'' = w(n)x_n' \quad (7)$$

Discrete Fourier transform. The discrete Fourier transform is employed in this stage to move from the time domain to the frequency domain in order to obtain the information contained in the spectrum of the signal.

Mel's filter bank. After that, the spectrum from each frame is put into a series of triangle bandpass filters that overlap [25], we selected a bloc of twenty triangle filters for this work; these filters are based on Mel's scale, and their features are comparable to those of the auditory system. To convert frequency f to Mel's frequency mel , we use the formula below:

$$mel = 2595 \log_{10}\left(1 + \frac{f}{700}\right) \quad (9)$$

At the output we obtain the mel spectrum m_j .

Logarithm & Discrete Cosine Transform DCT. The MFCC coefficients are determined by taking the decimal logarithm of the energy measured at each filter's output and applying the discrete cosine transform as follows:

$$c_i = \sqrt{\frac{2}{N}} \sum_j^N \log(m_j) \cos\left(\frac{\pi i}{N}(j-0.5)\right) \quad (10)$$

With N is now the number of filters and m_j become the logarithm of the energy at each filter's output.

Liftering. The cepstral coefficients of higher order are too small. To solve this issue, we apply liftering to increase the cepstrum, which in turn increases the amplitudes, bringing them closer together by using the following formula [16, 25]:

$$c_n' = \left(1 + \frac{L}{2} \sin\left(\frac{\pi n}{L}\right)\right) c_n \quad (11)$$

Classification. The method that we will adopt in this work in the classification phase is the method of the k nearest neighbors KNN as in Fig. 6. This technique was chosen due to the fact that it is a simple supervised machine learning technique used for classification, unlike some learning techniques, this method does not create a prediction model after the data training. Indeed, it is enough only to calculate the distance which exist between the various vectors constituting the whole of the data and to assign to the vector which one wants to classify the most frequent class for the k vectors closest neighbors of the studied vector, thus the determination of the neighbors is done by a calculation which is very easy, fast and which does not need a powerful calculator [26, 27].

In the classification phase, the choice of the number of neighbors and the size of the training and test sets acts on the performance of the classifier and can lead to decisions that can sometimes be inaccurate, in order to avoid this kind of problem we Choose an odd number of neighbors to avoid equality in number of neighbors of different classes. Also, in this work we adopt in the determination of the sets of training and test the k -fold cross validation approach, k -fold cross validation makes possible to divide the data set into k parts, $k-1$ parts will be intended for the training while the remaining part is for testing, this process will be repeated k times the performance of the classifier is calculated by averaging the results obtained in these k repetitions.

3 Results

As we have already mentioned, in this article we will take two signals of a different nature, the ECG signal and the PCG signal. In the data processing phase and the extraction of the parameters, we will decompose the ECG signal by the discrete wavelet transform DWT and extract the first eight approximation coefficients from a_1 to a_8 , [28, 29] which will subsequently be used to calculate the parameters following statistics [13, 28]:

The mean value:

$$\mu = \frac{1}{N} \sum_{i=1}^N x_i \quad (12)$$

The root mean square:

$$rms = \sqrt{\frac{1}{N} \sum_{i=1}^N x_i^2} \quad (13)$$

The variance:

$$v = \frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2 \quad (14)$$

Standard deviation:

$$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \mu)^2} \quad (15)$$

The skewness:

$$\varphi = \frac{1}{N} \frac{\sum_{i=1}^N (x_i - \mu)^3}{\sigma^3} \quad (16)$$

The kurtosis:

$$\psi = \frac{1}{N} \frac{\sum_{i=1}^N (x_i - \mu)^4}{\sigma^4} \quad (17)$$

Wavelet entropy:

$$H = -\frac{1}{N} \sum_{i=1}^N x_i^2 \log(x_i^2) \quad (18)$$

where x_n is a sampled signal and N represents the number of samples.

These features will constitute the first set of data, the second set of data consists of the first twelve MFCC coefficients extracted from the audio recordings of the PCG signals. Initially these two sets of data will be introduced separately into the KNN classifier, then we will combine between these two sets. The results summarized in Tables (1, 2, 3) are obtained by choosing the following parameters for the evaluation of these results:

$$Acc = \frac{TN + TP}{TP + TN + FP + FN} \times 100 \quad (19)$$

$$Sen = \frac{TP}{TP + FN} \times 100 \quad (20)$$

$$Spe = \frac{TN}{TN + FP} \times 100 \quad (21)$$

With:

TP: True positive (represents normal signal who were correctly classified).

TN: True negative (represents abnormal signal who were correctly classified).

FP: False positive (represents normal signal who were incorrectly classified).

FN: False negative (represents abnormal signal who were incorrectly classified).

So, we took seven as the number of neighbors and divided the data set into twenty parts. Our algorithm has been implemented in Matlab R2020a. The tests were carried out on a PC with the following settings:

CPU: Intel(R) Xeon(R) CPU E5-1620 0 @ 3.60 GHz 3.60 GHz.

Memory: 16 Go.

Table 1 Performance metrics of statistical features extracted from wavelet coefficients of the ECG signal

Wavelet	Classifier performance	The first eight approximation scale coefficients							
		a1	a2	a3	a4	a5	a6	a7	a8
Sym7	Accuracy	80.32	79.53	79.33	79.23	78.90	78.23	75.33	77.82
	Sensitivity	51.30	50.89	50.50	47.84	49.00	45.25	41.99	43.84
	Specificity	91.97	91.02	90.90	91.83	90.90	91.47	88.71	91.91
Sym8	Accuracy	80.37	79.66	78.99	79.39	77.81	77.80	77.07	77.69
	Sensitivity	51.19	50.69	50.36	48.65	47.03	45.01	44.75	43.51
	Specificity	92.09	91.28	90.48	91.72	90.17	90.96	90.05	91.10
Coif4	Accuracy	80.34	79.57	79.14	79.49	78.62	78.49	77.57	78.09
	Sensitivity	51.20	50.55	50.31	48.73	48.47	46.25	45.88	44.29
	Specificity	92.03	91.21	90.71	91.83	90.73	91.44	90.29	91.26
Coif5	Accuracy	80.31	79.42	79.35	79.15	78.64	78.02	78.38	77.84
	Sensitivity	51.23	50.53	50.34	48.65	47.24	45.47	45.67	45.73
	Specificity	91.99	91.02	90.99	91.39	91.24	91.08	91.50	90.74

Table 2 Performance metrics of MFCC coefficients extracted from PCG signal

Accuracy	Sensitivity	Specificity
75.40	51.31	85.07

Table 3 Performance metrics when we combine the feature corresponding to ECG & PCG signals

Wavelet	Classifier performance	The first eight approximation scale coefficients							
		a1	a2	a3	a4	a5	a6	a7	a8
Sym7	Accuracy	79.78	79.68	79.60	79.46	79.29	79.18	79.48	79.40
	Sensitivity	58.50	57.92	57.10	56.71	56.41	56.80	58.73	57.64
	Specificity	88.32	88.42	88.63	88.59	88.47	88.16	87.80	87.91
Sym8	Accuracy	79.84	79.72	79.66	79.79	79.47	79.90	79.47	79.77
	Sensitivity	58.48	58.05	57.28	56.71	56.91	58.66	59.48	58.15
	Specificity	88.42	88.42	88.64	89.05	88.53	88.43	87.49	88.17
Coif4	Accuracy	79.77	79.76	79.63	79.54	79.48	79.42	79.27	78.95
	Sensitivity	58.31	58.09	57.02	56.94	57.75	57.99	58.72	57.94
	Specificity	88.38	88.46	88.71	88.61	88.20	88.02	87.51	87.23
Coif5	Accuracy	79.79	79.71	79.80	79.66	79.80	79.65	78.94	78.58
	Sensitivity	58.51	57.99	57.36	56.87	58.25	58.75	58.10	56.98
	Specificity	88.34	88.42	88.81	88.81	88.45	88.04	87.30	87.02

The choice of these mother wavelet families is based on the fact that they have a comparable shape to the ECG signal. (See Fig. 5), in addition the decomposition with these wavelets has given good results in previous works [8], compared to similar work already carried out, we see also that the way in which we train the database plays an important role in improving the performance of the model, in fact the accuracy of this model is obtained by averaging the accuracies obtained after having trained the model a thousand times this shows that the accuracy of the lower scale coefficients increases (see Table 4). We also note that the statistical features resulting from the approximation coefficients give better results compared to the MFCC coefficients extracted from the PCG signal. In terms of the results obtained by combining the features of the two signals, it can be seen that the results have clearly improved from the coefficients of level 2 compared to those obtained by the processing of the ECG signal only, especially for levels 7 and 8 as shown in Fig. 7, this can be explained by the fact that the Mel filter bands have a large resolution for low frequencies [15] which makes the MFCC coefficients more effective at low frequencies for feature detection.

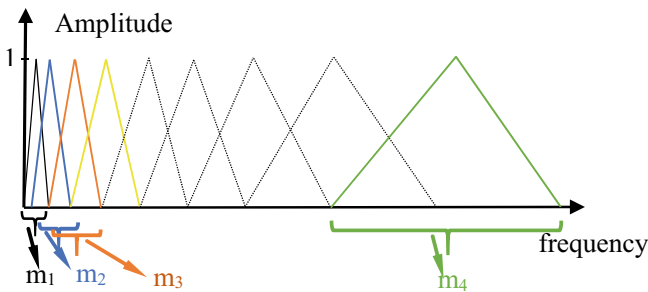


Fig. 5 Mel's filter bank

Fig. 6 Principle of the KNN algorithm

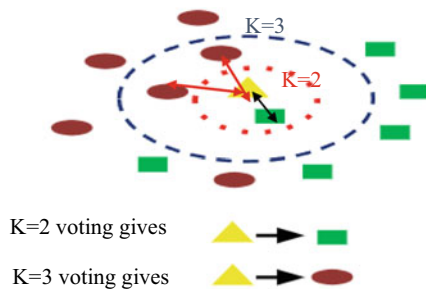


Table 4 Results obtained in a previous paper (SVM classifier) [29]

Wavelet	Classifier performance	The first eight approximation scale coefficients							
		a1	a2	a3	a4	a5	a6	a7	a8
Sym7	Accuracy	71,67	72,50	72,50	81,67	72,08	70,83	58,75	42,00
	Sensitivity	76,67	76,67	75,83	91,67	85,83	81,25	66,67	58,77
	Specificity	66,67	68,33	69,17	71,67	58,33	57,50	50,83	27,50
Sym8	Accuracy	73,33	72,92	73,75	79,17	68,75	67,50	60,00	42,98
	Sensitivity	78,33	75,83	77,50	90,00	80,00	77,50	69,17	63,21
	Specificity	68,33	70,00	70,00	68,33	57,50	55,83	50,83	25,83
Coif4	Accuracy	74,48	73,96	75,00	85,42	71,88	72,40	60,42	49,40
	Sensitivity	81,25	81,25	83,33	97,92	83,33	79,69	72,92	63,96
	Specificity	67,71	66,67	66,67	72,92	60,42	62,50	47,92	38,54
Coif5	Accuracy	75,00	73,96	75,00	87,50	73,96	73,96	55,21	42,80
	Sensitivity	82,29	80,21	82,29	100,00	84,38	82,29	65,63	55,31
	Specificity	67,71	67,71	67,71	75,00	63,54	63,54	44,79	33,33

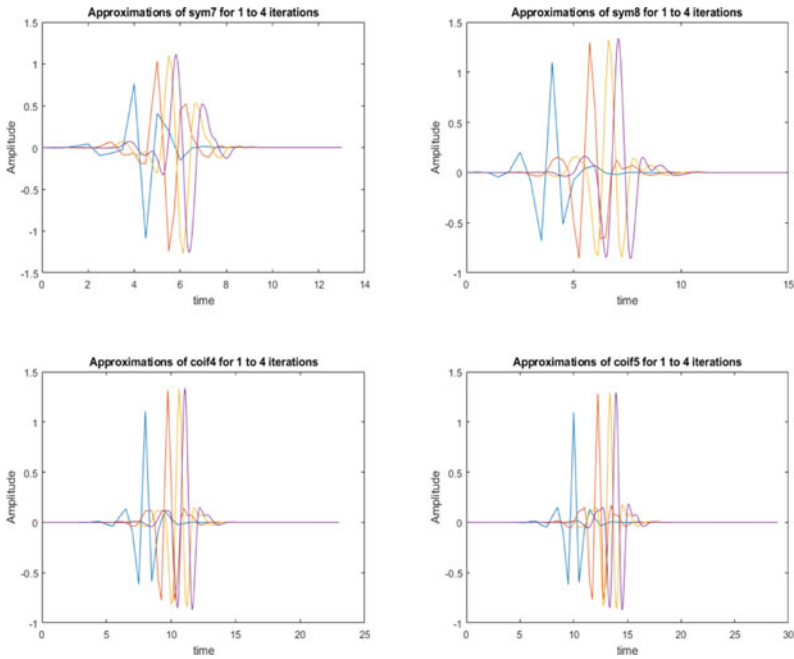


Fig. 7 Wavelet shapes at four level of scale

4 Conclusion

In this paper, we proposed to establish a comparison between two types of different features extracted from two different types of signals which are the ECG signal and the PCG signal, since we extracted the MFCC coefficients from the PCG signals and calculate the statistical features from the approximation coefficients resulting from wavelet decomposition of the ECG signal. The elaborate classification of these signals has shown us that the type of features and signals used have a considerable influence on the performance of the model. Also, we have tried to make the results obtained more credible by applying the k-fold cross-validation to the training and test bases, in addition the combination of the features of the ECG and PCG signals give promising results which can be better exploited in future work.


References

1. Manoharan S, Sathish P (2020) Patient diet recommendation system using K clique and deep learning classifiers. *J Artif Intell Capsul Netw* 2:121–130. <https://doi.org/10.36548/jaicn.2020.2.005>
2. Shakya S, Joby PP (2021) Heart disease prediction using fog computing based wireless body sensor networks (WSNs). *IRO J Sustain Wirel Syst* 3:49–58. <https://doi.org/10.36548/jsws.2021.1.006>
3. Nouhaila BO, Taoufiq BD, Benayad NS (2022) An intelligent approach based on the combination of the discrete wavelet transform, delta delta MFCC for Parkinson's disease diagnosis. *Int J Adv Comput Sci Appl* 13. <https://doi.org/10.14569/IJACSA.2022.0130466>
4. Addison PS (2005) Wavelet transforms and the ECG: a review. *Physiol Meas* 26:R155–R199. <https://doi.org/10.1088/0967-3334/26/5/R01>
5. Ismail S, Siddiqi I, Akram U (2018) Localization and classification of heart beats in phonocardiography signals—a comprehensive review. *EURASIP J Adv Signal Process* 2018:26. <https://doi.org/10.1186/s13634-018-0545-9>
6. Nabih-Ali M, El-Dahshan ES, Yahia AS (2017) A review of intelligent systems for heart sound signal analysis. *J Med Eng Technol* 41:1–11. <https://doi.org/10.1080/03091902.2017.1382584>
7. Babaei S, Geranmayeh A (2009) Heart sound reproduction based on neural network classification of cardiac valve disorders using wavelet transforms of PCG signals. *Comput Biol Med* 39:8–15. <https://doi.org/10.1016/j.compbiomed.2008.10.004>
8. Yusuf S, Hidayat R (2019) MFCC feature extraction and KNN classification in ECG signals, pp 1–5. <https://doi.org/10.1109/ICITACEE.2019.8904285>
9. Patro K, Kumar P (2017) Effective feature extraction of ECG for biometric application. *Procedia Comput Sci* 115:296–306. <https://doi.org/10.1016/j.procs.2017.09.138>
10. Rodrigues J, Belo D, Gamboa H (2017) Noise detection on ECG based on agglomerative clustering of morphological features. *Comput Biol Med* 87. <https://doi.org/10.1016/j.compbiomed.2017.06.009>
11. Karthikeyan P, Murugappan M, Yaacob S (2012) ECG signal denoising using wavelet thresholding techniques in human stress assessment. *Int J Electr Eng Inform* 4. <https://doi.org/10.15676/ijeei.2012.4.2.9>
12. Singh P, Pradhan G, Shah Nawazuddin S (2017) Denoising of ECG signal by non-local estimation of approximation coefficients in DWT. *Biocybern Biomed Eng* 37. <https://doi.org/10.1016/j.bbe.2017.06.001>

13. Sinha N, Das A (2020). Automatic diagnosis of cardiac arrhythmias based on three stage feature fusion and classification model using DWT. *Biomed Signal Process Control* 6. <https://doi.org/10.1016/j.bspc.2020.102066>
14. Rao K (2015) DWT based detection of R-peaks and data compression of ECG signals. *IETE J Res* 43:345–349. <https://doi.org/10.1080/03772063.1997.11416001>
15. Majeed SA, Husain H, Samad SA, Idbeaa TF (2015) Mel frequency cepstral coefficients (Mfcc) feature extraction enhancement in the application of speech recognition: a comparison study. *J Theor Appl Inf Technol* 79:38–56
16. Drissi TB, Zayrit S, Nsir B, Ammoummou A (2019). Diagnosis of Parkinson’s disease based on wavelet transform and Mel frequency cepstral coefficients. *Int J Adv Comput Sci Appl* 10. <https://doi.org/10.14569/IJACSA.2019.0100315>
17. Goldberger AL et al (2000) PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation* 101(23):e215–e220. <http://circ.ahajournals.org/cgi/content/full/101/23/e215>. [Circulation Electronic. Pages]
18. Feng H-Y, Wang J-P, Li, Y-C, Chen J (2011) 12 and application summarizing, pp 337–343. https://doi.org/10.1007/978-3-642-25255-6_43
19. Ramkumar M, Babu C, Kumar K, Hepsiba D, Manjunathan A, Kumar R (2021) ECG Cardiac arrhythmias classification using DWT, ICA and MLP neural networks. *J Phys Conf Ser* 1831:012015. <https://doi.org/10.1088/1742-6596/1831/1/012015>
20. Kumar A, Arumugam M, Bian G-B (2019) An intelligent learning approach for improving ECG signal classification and arrhythmia analysis. *Artif Intell Med* 103:101788. <https://doi.org/10.1016/j.artmed.2019.101788>
21. Qaisar S, Hussain F (2020) Arrhythmia diagnosis by using level-crossing ECG sampling and sub-bands features extraction for mobile healthcare. *Sensors* 20:2252. <https://doi.org/10.3390/s20082252>
22. Chandrashekhar V, Singh P, Paralkar M, Tonguz OK (2020) Pulse ID: the case for robustness of ECG as a biometric identifier, pp 1–6. <https://doi.org/10.1109/MLSP49062.2020.9231814>
23. Naing HM, Hidayat R, Hartanto R, Miyanaga Y (2020) Using double-density dual tree wavelet transform into MFCC for noisy speech recognition, pp 302–306. <https://doi.org/10.1109/ICI TEE49829.2020.9271737>
24. Arpitha Y, Madhumathi G, Balaji, N (2022) Spectrogram analysis of ECG signal and classification efficiency using MFCC feature extraction technique. *J Ambient Intell Humaniz Comput* 13. <https://doi.org/10.1007/s12652-021-02926-2>
25. Benba A, Jilbab A, Hammouch A (2015) Detecting patients with Parkinson’s disease using Mel frequency cepstral coefficients and support vector machines. *Int J Electr Eng Inform* 7:297–307. <https://doi.org/10.15676/ijeei.2015.7.2.10>
26. Veena K, Meena K, Teekaraman Y, Kuppasamy R, Radhakrishnan A (2022) SVM Classification and KNN techniques for cyber crime detection. *Wirel Commun Mob Comput* 2022:1–9. <https://doi.org/10.1155/2022/3640017>
27. Rahaman M (2019) A color and texture based approach for the detection and classification of plant leaf disease using KNN classifier
28. Toulni Y, Benayad N, Taoufiq BD (2021) ECG signal diagnosis using discrete wavelet transform and K-nearest neighbor classifier. <https://doi.org/10.1145/3454127.3457628>
29. Toulni Y, Benayad N, Taoufiq BD (2021) Electrocardiogram signals classification using discrete wavelet transform and support vector machine classifier. *IAES Int J Artif Intell (IJ-AI)* 10:960–970. <https://doi.org/10.11591/ijai.v10.i4.pp960-970>

Deep Convolutional Neural Network for Multi-class Brain Tumor Classification System in MRI Images



A. Jayachandran , M. A. Sreema, S. P. Anandaraj,
and T. Sudarson Rama Perumal

Abstract Brain tumour is a serious disease which can cause severe damage to the brain cells which eventually turns into a life threatening cancer. The tumour stages when identified early can help to increase the survival rates of the patients. The performance of the automated brain tumour diagnosis depends on the classification accuracy of the model. In this article, a deep convolutional neural network (DCNN) is developed for brain tumor classification of brain tumors in MRI images. Specifically, the auto-weight dilated convolutional unit utilized multi-scale convolutional feature maps to acquire brain tumor features at different scales and employed a learnable set of parameters to fuse convolutional feature maps in encoding layers. The AD unit is an effective architecture for feature extraction in the encoding stage. We used the advantages of the U-Net network for deep and shallow features, combined with AD units to multimodal image classification. In this model, the four-channel model inputs correspond to the MRI images of four modes, respectively. The main body of the network is composed of auto-weight dilated (AD) unit, Residual (Res) unit, linear upsampling, and the first and last convolution units. The network that applied Block-R3 had higher segmentation performance than the networks of Block-R1 and Block-R2. In the U-shaped network, feature extraction at the coding stage is the most important component. Designing the network to extract the features of interest efficiently is crucial. The proposed tumour diagnosis with the optimal feature extraction achieved better results with less time consumption.

Keywords Brain tumor · MRI · Segmentation · Deep CNN · Classification

A. Jayachandran (✉) · S. P. Anandaraj
Department of CSE, Presidency University, Bangalore, India
e-mail: ajaya1675@gmail.com

M. A. Sreema
Department of ECE, Arunachala College of Engineering for Women, Nagercoil, Tamil Nadu, India

T. Sudarson Rama Perumal
Department of ECE, Rohini College of Engineering and Technology, Nagercoil, Tamil Nadu, India

1 Introduction

Gliomas originate from intracranial tumors of glial cells and are highly lethal. It is the occurrence of mutations that are sufficient for carcinogenesis at the level of the cell's genetic material (DNA) and epigenetic material (EPI) through the interaction of internal genetic predisposing factors with external environmental pathogenic factors. Gliomas are mainly classified into the following categories: Astrocytoma, Oligodendroglioma, Mixed gliomas (such as oligodendro astrocytomas, which contain mixed types of glial cells), and Ependymoma. Patients with low-grade have a survival rate in months or even years, while the history of high-grade gliomas is often in weeks to months. Magnetic Resonance Imaging (MRI) has no ionizing radiation damage to human body and could be imaged without injecting radioisotopes. The soft tissue structure appears clear on MRI images. The multimodal gliomas sequences, including T1, T1c, T2, and T2-Flair, are advantageous for assessing health risks and clinical diagnosis [1, 2].

In recent years, more image segmentation methods are developed based the neural network technology to improve the deficiency of the traditional image segmentation algorithm, these methods are well applied in biomedical image segmentation [3]. The current image segmentation technology is mainly based on the color, gray, texture and other features of the image to extract the main part of the image. The image contains a lot of geometric information, and it is the most fundamental application to segment and extract the geometric features of the image. Its technology is widely used in autonomous driving, medical imaging, biometric recognition and remote sensing images. Traditional image segmentation methods are based on image threshold, edge detection and region based growth method [4]. However, these traditional methods have limitations for analyzing biological images. For example, the segmentation method based on image threshold is not appropriate to analyze images with no obvious difference in gray values or with different target gray value overlap, this method is also susceptible to noise producing false targets, when cell colors are not uniform in the images, using this method can produce holes during segmentation. Although the method based on edge detection can get the target profile, the structure inside the target is still missing. The segmentation method based on region growth is sensitive to noise and produces segmentation cavities or over-segmentation for complex images.

Manual segmentation of multimodal brain tumors is time-consuming and expensive compared to the automated method. It usually takes an expert radiologist about 3 h to the segment at the pixel level. The manual segmentation for the Dice Similarity Coefficient (DSC) score is 74%–85%. Therefore, accurate methods of tumor segmentation are of vital importance in clinical diagnostics and planning for treatment. The automated segmented methods aim to partition these multimodal MRI images into four tissues, including the normal tissues, the whole tumor (WT), the tumor core (TC), and the enhancing tumor (ET). TC describes that most tumors need to be removed. The WT describes the peritumoral edema (ED). And TC represents

the necrotic and the non-enhancing parts. Automatic gliomas segmentation accurately is still challenging work. In the 2010s, many methods have been proposed for automated segmentation in brain tumors [5, 6].

Since the powerful generalization, the deep learning method has gained a significant advantage compared to other approaches. Deep learning puts forward a way to let computers learn the features automatically based on data-driven to reduce the complexity of artificial design features. The deep learning model with essentially enlarged depth advances segmentation performance, such as CNN, FCN, GNN, RNN, GAN and other procedures of network. In brain tumor segmentation, the deeper neural network model is more and more important to advance the state-of-art performance. In recent years, deep learning methods gained significant interest in the segmentation of brain tumors. The U-shaped model is an efficient and straight-forward segmentation network in 3D medical images especially in brain tumors, learning features from deep and shallow neural units. The UNet model consists of four encoder layers and four decoder layers [7, 8]. The manuscript is organized as follows. Section 2 provides the related work. Section 3 illustrates the proposed work, Sect. 4 details the simulation results and Sect. 5 ends with conclusion.

2 Literature Review

The categorization of brain tumors has been the subject of empirical investigation. In this particular division, modern works on the classification of brain tumors are mentioned. Their proposed methodologies are also explained. Researchers have suggested many architectural models for the classification of brain tumors and the most eminent methods have been identified. Majib et al. (2021) [9] have proposed a VGG Net-Based deep learning framework. They considered 90 images, 81 of which have been labeled as YES and the rest labeled as No. They received the best training and also tested performance by the usage of CNN architecture. They achieved 97.8% F1 score using the VGG16 model and got the issue on the trade-off between the algorithmic performance and time complexity.

Noreen et al. (2020) [10] developed a concatenated approach for brain tumor classification, using Inception-V3 and DensNet201 deep learning models and extracted the features. They also concatenated those features using a softmax classifier for classification of the brain tumor cells. They achieved 99.34% and 99.51% accuracy for Inception-V3 and DensNet201 respectively. Kumar et al. (2019) [11] have suggested the combination of Deep Wavelet Autoencoder (DWA) with the Deep Neural Network (DNN). DWA adds the feature reduction property known for the autoencoder with the image decomposition property seen in the wavelet transform. They used all the images in the DICOM format and the python for processing the data. They achieved 96% accuracy by combining DWA with DNN. The entire research had been carried out using tenfold cross-validation. They also saw DNN as combination with other varieties of the encoder to compare these results. Liu et al. (2020) [12] have suggested deep Convolutional Long Short-Term Memory (C-LSTM) for

the detection of a tumor in the brain. They also made a comparative study with the other types of deep learning techniques and concluded that the proposed C-LSTM achieved satisfactory performance in the classification of the five classes of brain tumors. They ran the experiment 20 instances in order to get the average and standard deviation of the categorization findings. The main drawback of the model was its proneness to unanticipated noise and prototype errors. They saw computation as exorbitant, with datasets limited and so the training data sets as also limited requiring improvement.

Sultan et al. (2019) [13] have developed a brain tumor classification system or the classification of brain tumors using a Convolutional Neural Network (CNN). They worked on two different datasets. The first dataset was used for the classification of various types of tumors including glioma, meningioma, and pituitary tumor. The second dataset was used for the classification of different glioma grades. They found the proposed model requiring training with the use of a large number of datasets including samples from different age groups. Wang et al. (2019) [14] addressed the issues seen in the Deep Convolutional Neural Network (DCNN) considering the computational complexity as high. They have proposed an assignment process that can be used for correlation of the weights of the Fully Connected Layer (FCL), instead of the weight adjustment process for a reduction in complexity. According to the author's expertise, there exists some computational complexity in the existing system, and as also issues arising while handling large datasets. Time complexity can be reduced while handling huge datasets when the parameters are small.

3 Proposed Methodology

Deep learning method has gained a significant advantage compared to other approaches. Deep learning puts forward a way to let computers learn the features automatically based on data-driven to reduce the complexity of artificial design features. The deep learning model with essentially enlarged depth advances segmentation performance. Figure 1 depicts the summary of the proposed methodology.

3.1 Pre-processing

We used the randomization strategy as image preprocessing, which could ensure that the deep learning model still maintains strong generalization performance after a large number of repeated training. Multimodal brain images of the same patient use the same processing in one epoch training and different random measures in different epochs. It helps to learn the image features of different modes in the same brain while obtaining generalization. The figure shows the image preprocessing methods: 3D random clipping, 3D random rotation, 3D image intensity random enhancement, 3D image random mirror inversion, and normalization.

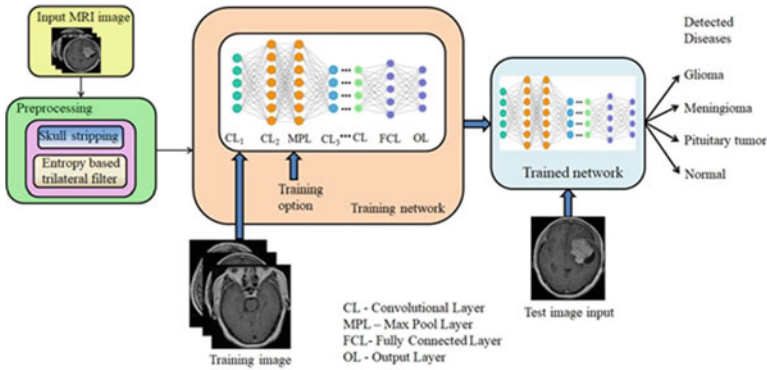


Fig. 1 Overall block diagram of proposed brain tumor segmentation model

Image normalization is a widely used technique in computer vision, pattern recognition and other fields. The z-score normalization was applied in this work. It is defined as per Eq. (1):

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where σ is the standard deviation, and μ is the mean value. Then, the 3D random clipping method randomly cuts the MRI image (240, 240, 155) into a matrix (144, 144, 128). The 3D random rotation method rotates the reduced image by the angle $U(-10, +10)$. The random intensity enhancement method of 3D image sets the image pixel value is defined as per Eq. (2):

$$x_{new} = x_{old} * U(0.9, 1.1) + U(-0.1, 0.1) \tag{2}$$

where U is the uniform distribution. The random mirror processing symmetrizes the image according to deep, height and width directions. We applied these image enhancement routines to extend the training data set to improve the performance and generalization ability of the deep neural network.

3.2 Deep Learning Model

Deep learning which is a subset of Artificial intelligence is gaining momentum each day by making different tasks much easier and more efficient. CNN which is a type of deep learning mechanism is an inevitable part of image vision problems. In recent years, deep learning methods gained significant interest in the segmentation of brain tumors. The U-shaped model is an efficient and straightforward segmentation network in 3D medical images especially in brain tumors, learning features from

deep and shallow neural units [15–17]. The UNet model consists of four encoder layers and four decoder layers. The proposed model is shown in Fig. 2. In this model, the four channel model inputs correspond to the MRI images of four modes, respectively. The main body of the network is composed of auto-weight dilated (AD) unit, Residual (Res) unit, linear upsampling, and the first and last convolution units. In the downsampling stage (feature coding extraction), we use 8 AD units to obtain multi-scale feature maps. In the upsampling stage (feature decoding), we use the AD unit, Res unit and a linear upsampling layer to form a primary decoding layer. Finally, a convolution unit outputs the results of the network model. Moreover, each convolution unit, AD unit and Res unit contains batch normalization and ReLU functions. We used extended convolution to extract fine-grained and multi-scale glioma features, and employed residual structure to obtain long-dependent glioma features.

As for the Res Unit layer, we used two convolution units to reduce and then enlarge the number of convolution kernels so as to realize feature learning and feature map reorganization. From an experimental point of view, this is an efficient coding method. Then, we used two group convolution units with stripe 1 and group 16, and the kernel size is $3 \times 3 \times 3$. Finally, we used a convolution residual element to obtain the characteristic graph of long dependence. As for the AD Unit layer, we used two convolution units firstly (like the Res unit). Then, we used three extended convolution units (the divided parameters are 1 and 2, respectively) and used two learnable parameters to adjust and fused the characteristics of the two group extended convolution units. Finally, a group convolution unit was used to output the result of the AD unit. We also set up residual calculations in the AD unit. The dilated convolution could expand the receptive field of the convolutional kernel without sacrificing computational resources, while normal convolution could provide a more

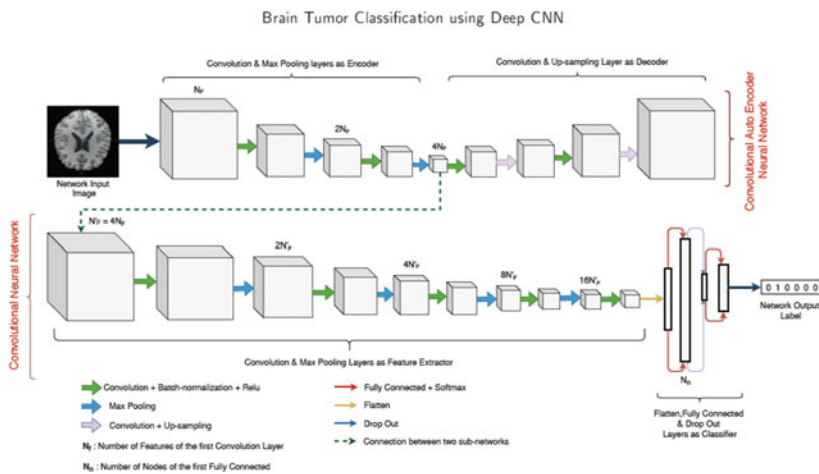


Fig. 2 An illustration of the proposed architecture for brain tumor segmentation

accurate feature map. The fusion of the two types of convolutions could strengthen the ability of the network to extract features.

In the encoder stage, each residual block is a dual-pathway structure. In this stage, we set channel depth to 32, 64, 128, and 256. The residual block is the critical structure of down-sampling. In the decoder stage, we connect a convolutional unit and a de-convolutional unit for upsampling. The stripe of the de-convolutional unit is $2 \times 2 \times 2$, and the kernel size is $3 \times 3 \times 3$. Batch Normalization and RReLU activation functions are connected behind the convolutional unit or the de-convolutional unit for all convolutional units and all de-convolutional units in this stage. Similarly, we set the channel depth in the decoder stage to $32 \times 64 \times 128 \times 256$. A combined deep neural network with the residual blocks enables the network to obtain more significant gradients in deep layers.

So, the phenomenon of gradient disappearance is relatively rare and gets more practical features of gliomas. The formula of the gradient propagation in the convolutional layer can be defined as per Eq. (3),

$$\delta_l = \sigma'(O_l) \cdot (w_{l+1})^T \delta_{l+1} O_l \tag{3}$$

where σ' means the first derivative of the loss function, w describes the weight, O indicates the output matrix vector, and l is the layer l . Then, the gradient in Block-R1, Block-R2 and Block-R3 can be defined as per Eq. (4),

$$\begin{aligned} O_{r1_{l+1}} &= f(\delta_2(f(\delta_1(O_{r1_l})) + O_{r1_l})) \\ O_{r2_{l+1}} &= f(\delta_2(f(\delta_1(O_{r2_l})) + O_{r2_l})) \\ O_{r3_{l+1}} &= \delta_2(f(\delta_1(f(O_{r3_l})) + O_{r3_l})) \end{aligned} \tag{4}$$

where f means the activation function, δ_1 and δ_2 represent the first and second convolution calculations, respectively. It is worth noting that the difference between Eqs. (3) and (4) lies in the order of normalization, which is not reflected in the equation

Multiplication is widely used in the calculation of series convolution, such as $\delta_2(f(\delta_1))$. The cumulative multiplication between $(-1, 1)$ makes it possible for the gradient to appear the result of approximate 0, so that the classical gradient disappears. The residual-connection weakens this problem through weight addition, and enhances the stability of the network. Obviously, it is a very effective way to use residual blocks to build architecture in very deep neural layers, especially in calculating the depth feature map.

We defined the convolution block (BN, RL, Conv) in AD unit as per Eq. (5)

$$\ell_{(c_i,k)} = w_{c_i}^T f(I_i) + b_{c_i} \quad (5)$$

where c_i is the convolution layer i , and k describes the kernel size, f is the activation function. The w and b represent the convolution weight and bias, I_i is the input data. Then, the AD block can be define as per Eq. (6)

$$\ell_{AD} = \ell_{(c_1,1)} \ell_{(c_2,1)} (a \ell_{(d_0,3)} + b \ell_{(c_3,3)}) \ell_{(c_4,3)} + \ell_{(c_5,1)} \quad (6)$$

where ℓ_{d_0} is the dilated convolution. In the gradient back-propagation process, a and b can automatically adjust the weight ratio of the convolution integral branch in the main path. In addition, the channel parameters of the AD-Net were set to 32, 64, 128, 256, and the skip connection adopted the 3D matrix concatenate method. The residual structure was a necessary element. The residual calculation ensured the stability of gradient in deep feature calculation.

4 Experimental Results and Discussion

4.1 Dataset

The MICCAI BraTS datasets consist of many pre-operative multimodal MR images from multi- institution [5, 15]. MICCAI is a comprehensive academic conference held by the international medical image computing and Computer Assisted Intervention society. It is a top-level conference in the field of medical image computing (MIC) and computer assisted intervention (CAI). The BraTS datasets already applied various preprocessing steps by the organizers. BraTS challenge has always focused on evaluating advanced methods for the multimodal brain tumor segmentation. The results of the validation set and test set required to be evaluated and returned by the online platform. In this study, two datasets (BraTS19 and BraTS20) were used. The BraTS19 training set consists of 335 multimodal MRI images, including high-grade gliomas (HGG) and low-grade gliomas (LGG), 259 HGGscans and 76 LGG scans. In addition, the validation set includes 125 MRI scans. The BraTS20 training set consists of 369 multimodal MRI images, including HGG and LGG and paying no attention to divide into HGG and LGG. In addition, the validation set contains 125 MRI scans and the testing set includes 166 MRI scans. The test set is for participants. Each scans contains five MRI images (T1, T1c, T2, Flair and segmentation label).

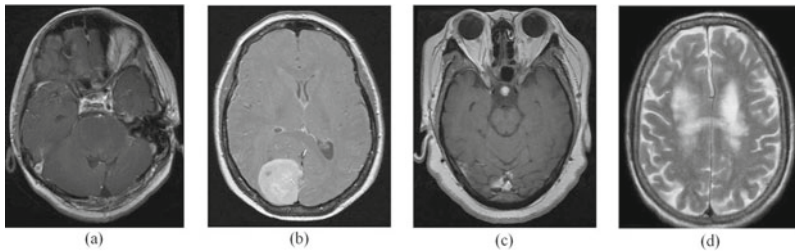


Fig. 3 Typical brain tumor images **a** Glioma **b** Meningioma **c** Pituitary tumor **d** Normal

4.2 Classification Results

The proposed brain MRI images segmentation method is simulated using MATLAB 2020b driven by a single Intel(R) Core(TM) i5-3570 64bit CPU with 8 GB RAM and 3.40 GHz clock frequency. Additional memory was not required during the entire process. In this work, three different publicly available Dataset is used for training and testing the classifier. The different types of brain tumor images are given in Fig. 3.

The following evaluation metrics are used for validating the classifier performance. It is given in Eq. (7)

$$\begin{aligned}
 Accuracy &= \frac{TP + TN}{Total\ no\ of\ images} \\
 Precision &= \frac{TP}{FP + TP} \\
 Recall &= \frac{TP}{FN + TP} \\
 F1score &= 2 * \frac{Precision * Recall}{Precision + Recall}
 \end{aligned} \tag{7}$$

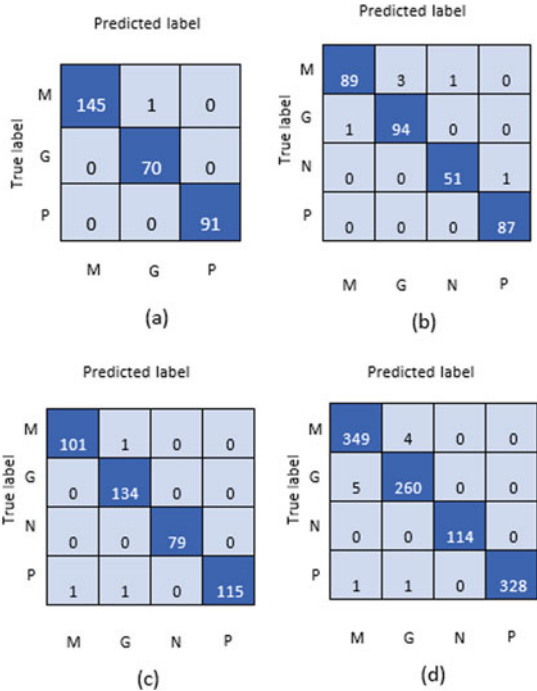
where accuracy explains how well the architecture can classify the images. It is given as the ratio of total correct prediction made to the total number of predictions made. Precision is given by the ratio of total number of correctly classified positive cases to the sum of total positive cases predicted. High value for precision is required in order to minimize the number of false positive classifications. Recall is defined as the ratio of total positive cases correctly classified to all correctly classified observations and F1 score gives the harmonic mean of precision and recall. It can be also defined as the weighted average of precision and recall [18–20]. The performance comparison of proposed model and other models are given in Table 1. The confusion matrix of the proposed system with different Dataset is given in Fig. 4 and the classification rate of various models using different statistics method is given in Fig. 5.

Table 1 Comparison of proposed system performance with state of the art methods

Types	Recall (%)				Precision (%)				F1-score (%)			
	P1	L1	L2	L3	P1	L1	L2	L3	P1	L1	L2	L3
Glioma	97	87.2	90.5	93.2	95	82.1	89.7	91.5	96.3	81.4	89.8	91.4
Meningioma	98.2	89.4	91.3	93.6	93.6	76.9	88.4	91.4	98.2	88.6	89.3	93.5
Pituitary tumor	96.3	88.3	89.7	92.1	94.2	81.4	89.3	92.4	99.3	83.7	91.3	94.6
No tumor	99.5	91.7	92.4	95.7	94.3	83.1	89.4	91.7	91.5	89.5	88.3	90.1

P1-Proposed method; L1-Wang et al. (2019) [14]; L2-Noreen et al. (2020) [10]; L3-Kumar et al. [15]

Fig. 4 Confusion matrix of proposed Deep CNN model (a) Dataset 1, (b) Dataset 2, (c) Dataset 3, and (d) Merged Dataset



5 Conclusion

In this paper, a new multi-scale approach to segment the brain tumor in MRI image was described and evaluated in several publicly available databases. This paper also presents an assessment of the most appropriate scales for the brain tumor segmentation, complementing previous work that defines these scales empirically. Furthermore, it was also demonstrated that a multi-scale analysis can improve the brain tumor segmentation. Although recent research has been focusing on deep learning

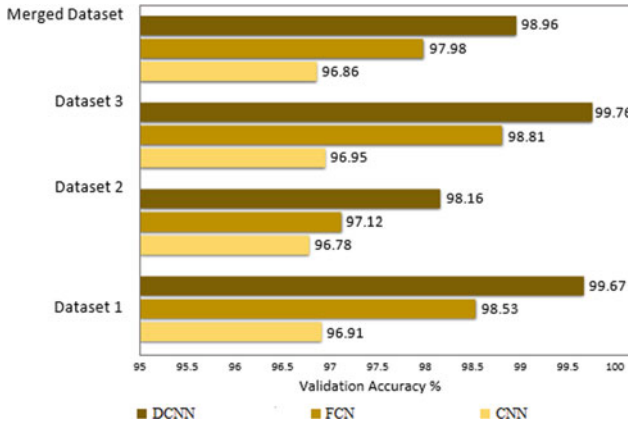


Fig. 5 Experimental results of accuracy of different models on all datasets

methods, rule-based methods can also be important for the definition of features, that can significantly improve the outcome of these methods. The achieved results show that the proposed approach is very competitive when compared with the current state of the art methods, particularly in high-resolution images. Our method still needs further improvement in the enhancing tumor region segmentation. It were practical tools in 3D brain tumor segmentation.

References

1. Pashaei A, Sajedi H, Jazayeri N (2018) Brain tumor classification via convolutional neural network and extreme learning machines. In 2018 8th International conference on computer and knowledge engineering (ICCKE). IEEE, pp 314–319
2. Irmak E (2021) Multi-classification of brain tumor MRI images using deep convolutional neural network with fully optimized framework. *Iran J Sci Technol Trans Electric Eng* 1–22
3. Rajasekaran KA, Gounder CC (2018) Advanced brain tumour segmentation from MRI images. *Basic physical principles and clinical applications, high-resolution neuroimaging*, pp 83–108
4. Hossain T, Shishir FS, Ashraf M, Al Nasim MDA, Shah FM (2019) Brain tumor detection using convolutional neural network. In 2019 1st International conference on advances in science, engineering and robotics technology (ICASERT). IEEE, pp 1–6
5. Jayachandran A, Dhanasekaran R (2013) Brain tumor detection using fuzzy support vector machine classification based on a Texton co-occurrence matrix. *J Imag Sci Technol* 7(1):10507-1-10507-7
6. Aurna NF, Anika FS, Rubel MDTM, Habibul Kabir K, Shamim Kaiser M (2021) Predicting periodic energy saving pattern of continuous IOT based transmission data using machine learning model. In 2021 International conference on information and communication technology for sustainable development (ICICT4SD). IEEE, pp 428–433
7. Jayachandran A, Kharmega Sundararaj G (2016) Abnormality segmentation and classification of multi model brain tumor in MR images using fuzzy based hybrid kernel SVM. *Int J Fuzzy Syst* 17(3):434–443

8. Mahiba C, Jayachandran A (2019) Severity analysis of diabetic retinopathy in retinal images using hybrid structure descriptor and modified CNNs. *Measurements* 135:762–767
9. Sajjad M, Khan S, Muhammad K, Wu W, Ullah A, Baik SW (2019) Multi-grade brain tumor classification using deep CNN with extensive data augmentation. *J Comput Sci* 30:174–182
10. Noreen N, Palaniappan S, Qayyum A, Ahmad I, Imran M, Shoaib M (2020) A deep learning model based on concatenation approach for the diagnosis of brain tumor. *IEEE Access* 8:55135–55144
11. Kumar Mallick P, Ryu SH, Satapathy SK, Mishra S, Nguyen GN, Tiwari P (2019) Brain MRI image classification for cancer detection using deep wavelet autoencoder-based deepneural network. *IEEE Access* 7:46278–46287
12. Liu Y et al (2020) Deep C-LSTM neural network for epileptic seizure and tumor detection using high-dimension EEG signals. *IEEE Access* 8:37495–37504
13. Sultan HH, Salem NM, Al-Atabany W (2019) Multi-classification of brain tumor images using deep neural network. *IEEE Access* 7:69215–69225
14. Balasooriya NM, Nawarathna RD (2017) A sophisticated convolutional neural network model for brain tumor classification. In: 2017 IEEE international conference on industrial and information systems (ICIIS). IEEE, pp 1–5
15. Wang W, Bu F, Lin Z, Zhai S (2020) Learning methods of convolutional neural network combined with image feature extraction in brain tumor detection. *IEEE Access* 8:152659–152668
16. Afshar P, Plataniotis KN, Mohammadi A (2019) Capsule networks for brain tumor classification based on MRI images and coarse tumor boundaries. In: ICASSP 2019–2019 IEEE International conference on acoustics, speech and signal processing (ICASSP). IEEE, pp 1368–1372
17. Prabhu AJ, Jayachandran A (2018) Mixture model segmentation system for parasagittal meningioma brain tumor classification based on hybrid feature vector. *J Med Syst* 42(12)
18. Namboodiri S, Jayachandran A (2020) Multi-class skin lesions classification system using probability map based region growing and DCNN. *Int J Comput Intell Syst* 13(1):77–84
19. Vijayakumar T (2019) Classification of brain cancer type using machine learning. *J Artif Intell* 1(2):105–113
20. Karuppusamy DP (2020) Hybrid manta ray foraging optimization for novel brain tumor detection. *J Soft Comput Paradigm* 2(3):175–185

Application of the Particle Swarm Algorithm to the Task of Image Segmentation for Remote Sensing of the Earth



Igor Ruban, Hennadii Khudov, Oleksandr Makoveichuk, Igor Butko, Sergey Glukhov, Irina Khizhnyak, Nazar Shamrai, and Temir Kalimulin

Abstract The article proposes the application of the particle swarm algorithm to the task of image segmentation for remote sensing of the Earth. The well-known methods of image segmentation are analyzed. They are ineffective for segmentation of remote sensing of the Earth data due to the peculiarities of space images. The proposed approach to image segmentation uses the particle swarm algorithm in each channel of brightness; divides into segments in each channel of brightness based on the operation of the particle swarm algorithm; combines segmentation results in each channel; gets the overall result of image segmentation. Type I and type II segmentation errors are calculated for the proposed approach and the k-means method. The proposed approach to segmentation estimation reduces type I segmentation errors by about 11% and type II segmentation errors by about 8%.

Keywords Remote sensing of the Earth · Image segmentation · Particle swarm optimization

1 Introduction

Every day, data obtained by remote sensing of the Earth are increasingly used in different areas of human activity [1–3]. These are national security, defense, military

I. Ruban · O. Makoveichuk · I. Khizhnyak · T. Kalimulin
Kharkiv National University of Radio Electronics, Kharkiv, Ukraine
e-mail: ruban_i@ukr.net

H. Khudov (✉)
Ivan Kozhedub Kharkiv National Air Force University, Kharkiv, Ukraine
e-mail: 2345kh_hg@ukr.net

I. Butko
Academician Y. Bugay International Scientific and Technical University, Kyiv, Ukraine
e-mail: butko_igor@ukr.net

S. Glukhov · N. Shamrai
Military Institute of Taras Shevchenko National University of Kyiv, Kyiv, Ukraine
e-mail: shamrainazar1981@ukr.net

intelligence, construction, cadastre, business, agriculture, forestry, mapping, etc. [3–6]. Modern technologies of remote sensing data application are actively implemented exactly on the basis of space systems. This is due to the availability of a large number of space satellites, with the possibility of obtaining space images of free high spatial resolution in the public domain and other advantages of remote sensing [7]. Therefore, it is necessary to develop and improve methods of data analysis, processing and interpretation using digital image processing methods for further use of remote sensing data.

Data mining methods of Earth remote sensing are divided into [8, 9]:

pre-processing methods (methods of color balancing, methods of noise suppression, methods of conversion to the required formats, methods of transition to other color models, etc. [1, 7, 9]);

thematic processing methods (methods of determining the contours of objects [10–12], methods of segmentation [13–15], methods of interpretation [16], etc.).

The stage of image segmentation is the main stage of processing of remote sensing data. The results of methods of segmentation affect the result of correct interpretation and analysis of data on the image [1, 2, 5].

Papers [17, 18] discuss trends and advances of the methods of segmentation of biomedical and medical images. These methods give good results of segmentation only on biomedical and medical images. It is when there are few objects of interest on the image and they occupy most of the whole image. The disadvantage [17, 18] is poor results of segmentation of images of remote sensing.

In [19, 20] segmentation methods using neural networks are proposed. The disadvantages [19, 20] are the long hours of creating an initial sample for the neural network and its training. The need to make changes when new objects of interest appear.

Paper [21] proposes segmentation of tomographic images. The advantage is real-time image processing. The disadvantage [21] is that this approach is used to segment only tomographic images.

Paper [22] proposes methods of image segmentation based on clustering. These approaches show good results in the segmentation of biomedical images. The disadvantage [22] is the good results of segmentation of only magnetic resonance images.

Paper [23] proposes image segmentation by method of k-means. The disadvantages [23] are the long hours costs, the unknown number of clusters and initialization, and dead centres of clusters.

Paper [24] proposes improved method of fuzzy c-means and improved method of k-means. The advantage [24] is automatic selection of centroids of clusters. The disadvantages [24] are the under-segmentation of remote sensing images with a large number of objects of interest on it and the long hours costs on segmentation of satellite image.

In [25, 26] it is proposed to use methods of parallel image segmentation. This approach is proposed in order to reduce the processing time. The disadvantage of the approaches [25, 26] is the high computational cost.

In [27, 28] methods of segmentation of the color images are proposed. The formation of the color images has some peculiarities. Therefore, the disadvantages [27, 28] is the need for special approaches for segmenting images of different color models.

Papers [29, 30] propose existing and new techniques of image segmentation. There are a large number of these techniques of image segmentation. However, most of them are ineffective for segmentation of space images due to the peculiarities of formation of such images.

2 Problem and Presentation Materials Researching

2.1 Image Segmentation in General Form

Image segmentation in general form can be represented as [31]:

$$g(x_1, x_2) \rightarrow g^s(x_1, x_2), \tag{1}$$

where $g(x_1, x_2)$ – the original image for segmentation;

$g^s(x_1, x_2)$ – the result of image segmentation.

The process of image segmentation (1) is the division of the original image into some groups of segments that satisfy to the next condition [31]:

$$\left\{ \begin{array}{l} \bigcup_{i=1}^N S_i = S, \quad i = 1, \dots, N; \\ S_i \cap S_j = \emptyset, \quad \text{for } j \neq i; \quad \forall j, i = \overline{1, N}; \\ LP(S_i) = 1; \quad \forall i = \overline{1, N}; \\ LP(S_i \cap S_j) = 0, \quad \text{for } j \neq i; \quad \forall j, i = \overline{1, N}, \end{array} \right. \tag{2}$$

where $S = \{S_1, S_2, \dots, S_N\}$ – image segments on the $g^s(x_1, x_2)$;

LP – the predicate;

N – total number of segments.

$LP = 1$, provided that the pair of pixels of S_i segment performs the next condition [31]:

$$LP(S_i) = \begin{cases} 1, & \text{if } g(x_{11}, x_{21}) = \dots = g(x_{1p}, x_{2p}); \\ 0, & \text{others,} \end{cases} \tag{3}$$

where $(x_{1p}, x_{2p}) \in S_i, p = 1, \dots, P$;

P – total number of pixels of the segment S_i .

In general, the result of the process of image segmentation is the division of the $g(x_1, x_2)$ into S_N segments, some of which are objects, the other is the background.

2.2 Image Segmentation by the Particle Swarm Algorithm in General Form

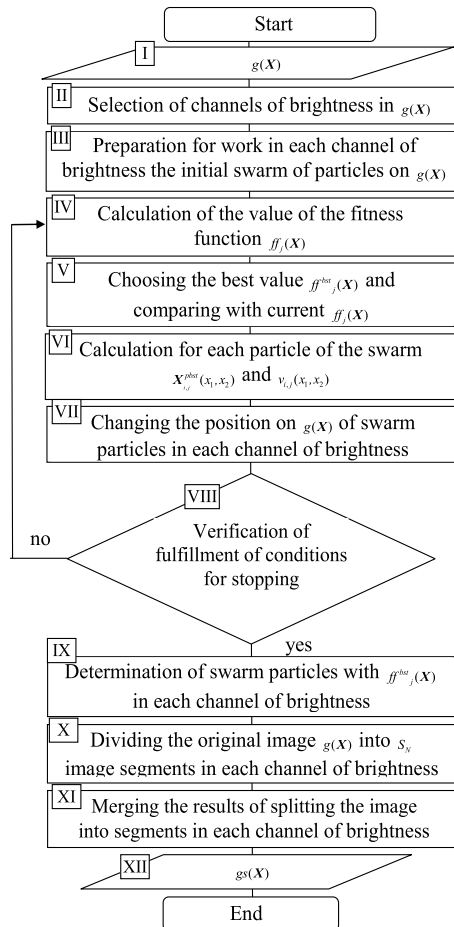
The block diagram of image segmentation by the particle swarm algorithm (PSA) is a sequence of the following main actions (Fig. 1) [32]:

I. Input the original image for segmentation: $g(X)$ – the original image for segmentation; $X(x_1, x_2)$ – coordinates of pixel on the original image.

II. Selection of channels of brightness on the original image.

If the original image $g(X)$ is a color image in the color space RGB, then this operation will result in three channels of brightness – R, G and B respectively. If the original image $g(X)$ is a tone image, then the result of this operation will be one channel of brightness.

Fig. 1 The block diagram of image segmentation by the particle swarm algorithm



III. Preparation for work in each channel of brightness the initial swarm of particles on the original image $g(\mathbf{X})$.

$\mathbf{X}_{i1}(x_{1i1}, x_{2i1})$ – the vector of the initial position of the swarm of particles;
 $i = 1, \dots, C$ – in the swarm a total number of particles.

Paragraphs I–III are executed only once at the first iteration of this algorithm.

IV. Calculation of the value of the fitness function.

The fitness function $ff_j(\mathbf{X})$ is calculated in each channel of brightness for each initialized swarm particle $\mathbf{X}_{i1}(x_{1i1}, x_{2i1})$ in the original image $g(\mathbf{X})$.

As a fitness function, we choose a function of the following form [10, 33]:

$$ff_j(\mathbf{X}) = \sum_{s=1}^Q \sum_{l=1}^L (R_l^s(j)), \quad (4)$$

where j – the number of the iteration of this algorithm;

L – the size of the image $g(\mathbf{X})$;

s – the current swarm particle number;

$R_l^s(j)$ – the function that defines a section of the route.

The function $R_l^s(j)$ is calculated as [10, 33]:

$$R_l^s(j) = |\Delta x_{1i}^s(j)| + |\Delta x_{2i}^s(j)| + c |\Delta d_i^s(j)|, \quad (5)$$

where $|\Delta x_{1i}^s(j)|, |\Delta x_{2i}^s(j)|$ – changing the position of the s particle of the swarm in the i pixel of the image at the j iteration of the algorithm along the axes;

$|\Delta d_i^s(j)|$ – the difference between the brightness values of neighbouring pixels in the i pixel of the image for the s swarm particle at the j iteration of the algorithm.

The $|\Delta d_i^s(j)|$ is calculated as [10, 33]:

$$|\Delta d_i^s(j)| = |g(x_{1i}^s(j), x_{2i}^s(j)) - g(x_{1i-1}^s(j), x_{2i-1}^s(j))|. \quad (6)$$

c – the coefficient that takes into account different units of movement and brightness.
 $c = 1$ provided that the brightness values are in the range from 0 to 255.

So, given expressions (5) and (6), the fitness function (4) can be calculated as the next expression:

$$ff_j(\mathbf{X}) = \sum_{s=1}^Q \sum_{l=1}^L (|\Delta x_{1i}^s(j)| + |\Delta x_{2i}^s(j)| + c |g(x_{1i}^s(j), x_{2i}^s(j)) - g(x_{1i-1}^s(j), x_{2i-1}^s(j))|). \quad (7)$$

V. Choosing the best value of the fitness function and comparing the current value of the fitness function for each particle in the swarm with the chosen best one. This operation is carried out in each channel of brightness of the original image $g(\mathbf{X})$.

The best position of the swarm particle at the j iteration is defined as:

$$\mathbf{X}_j^{gbst}(x_1, x_2) = \begin{cases} \mathbf{X}_{j-1}(x_1, x_2), & \text{if } ff(\mathbf{X}_{j+1}(x_1, x_2)) \geq ff(\mathbf{X}_j(x_1, x_2)); \\ \mathbf{X}_{j+1}(x_1, x_2), & \text{if } ff(\mathbf{X}_{j+1}(x_1, x_2)) < ff(\mathbf{X}_j(x_1, x_2)). \end{cases} \quad (8)$$

VI. Calculation for each particle of the swarm of the value of its new location on the image and the speed of its movement across the original image $g(\mathbf{X})$. These operations are also performed in each channel of brightness of the original image $g(\mathbf{X})$.

The value of the speed of the movement for each swarm particle is determined using expression:

$$v_{i,j+1}(x_1, x_2) = t \cdot v_{i,j}(x_1, x_2) + a_1 \text{ran}_{1,j} \left[\mathbf{X}_{i,j}^{gbst}(x_1, x_2) - \mathbf{X}_{i,j}(x_1, x_2) \right] + a_2 \text{ran}_{2,j} \left[\mathbf{X}_{i,j}^{pbst}(x_1, x_2) - \mathbf{X}_{i,j}(x_1, x_2) \right], \quad (9)$$

where $v_{i,j}(x_1, x_2)$ – the value of the speed of i swarm particles at the j iteration of the algorithm;

t – the inertia factor. It is an empirical coefficient that determines the change in speed of particle. It is introduced in order to find new areas on the image;

$\mathbf{X}_{i,j}(x_1, x_2)$ – the vector of the coordinates of i swarm particle at the j iteration of the algorithm;

$\mathbf{X}_{i,j}^{pbst}(x_1, x_2)$ – the vector of the coordinates of i swarm particle with the best value of the fitness function from all the values of the fitness functions at the j iteration of the algorithm.

The vector of coordinates $\mathbf{X}_{i,j}^{pbst}(x_1, x_2)$ is calculated as:

$$\mathbf{X}_{i,j}^{pbst}(x_1, x_2) \in \left\{ \mathbf{X}_{i,j}(x_1, x_2), \dots, \mathbf{X}_{i,j}(x_1, x_2) \right\}, \quad \text{for} \\ ff(\mathbf{X}_{i,j+1}^{pbst}(x_1, x_2)) = \min \left\{ ff(\mathbf{X}_{i,j}(x_1, x_2)), \dots, ff(\mathbf{X}_{i,j}(x_1, x_2)) \right\}, \quad (10)$$

where a_1, a_2 – the coefficients of acceleration;

$\text{ran}_1, \text{ran}_2$ – the coefficients that introduce randomness. Their value is chosen in the range from 0 to 1 inclusive.

VII. Changing the position on the original image of swarm particles in each channel of brightness.

The coordinates of the new location of the swarm particles are defined as:

$$\mathbf{X}_{i,j+1}(x_1, x_2) = \mathbf{X}_{i,j}(x_1, x_2) + v_{i,j}(x_1, x_2). \quad (11)$$

VIII. Verification of fulfilment of conditions for ending the iteration process. The conditions for the end of the iteration process are:

- execution of a given number of iterations;
- the increase in speed became almost equal to zero.

If one of the conditions is met, then the next action is paragraph IX. If none of the conditions is met, then the transition to the next iteration of the algorithm takes place – paragraphs IV–VII.

IX. Determination of swarm particles on the image with the best fitness function values in each channel of brightness.

X. Dividing the original image $g(X)$ into S_N image segments in each channel of brightness.

XI. Merging the results of splitting the image into segments by individual channel of brightness.

XII. Getting result of the segmentation $-gs(X)$.

2.3 The Result of Image Segmentation for Remote Sensing of the Earth by the Particle Swarm Algorithm

Let us apply the proposed approach to image segmentation for remote sensing of the Earth. As an original image, we take a color optical-electronic image from a spacecraft WorldView-2 [34]. This image of the remote sensing of the Earth in the public domain is provided by MAXAR (Fig. 2).

Size of the original image is 1757×1297 pixels. The color space is RGB. Visually, the original image contains objects of interest such as military equipment.

On Fig. 3 presented the result of image segmentation for remote sensing of the Earth (Fig. 2) by the PSA. This is a segmented image by the PSA after combining the channels of brightness.



Fig. 2 An original image from a spacecraft WorldView-2 [34]

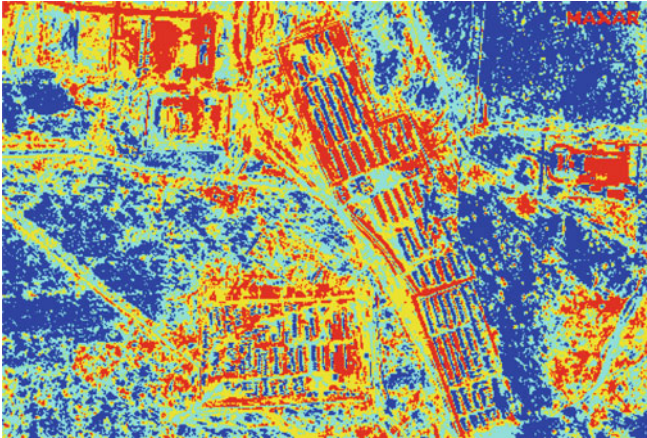


Fig. 3 The result of image segmentation by the PSA after combining the channels of brightness

On Fig. 3 the whole image is divided into four segments, which are marked with different colors. Objects of interest are marked in red color on the image. Visual analysis shows that this approach allows segmentation of images from space remote sensing systems.

2.4 Visual Assessment of Quality of Image Segmentation by the Particle Swarm Algorithm and k-Means

For quality comparison of segmentation of images from space remote sensing systems by the PSA was chosen the well-known k-means method [35]. The number of clusters in the method k-means was chosen to be four. The result of segmentation of the original image using the k-means method is shown on Fig. 4.

Visual assessment of quality of image segmentation by the PSA (Fig. 3) and k-means (Fig. 2) showed that:

- the k-means method relates objects of interest under snow cover (should be highlighted in red color) to snow (highlighted in blue color);
- the proposed approach to segmentation by the PSA classifies almost all units of military equipment on the image as objects of interest (highlighted in red color);
- the proposed approach to segmentation by the PSA refers to the snow cover some objects of military equipment that are completely snowed in (highlighted in blue color).

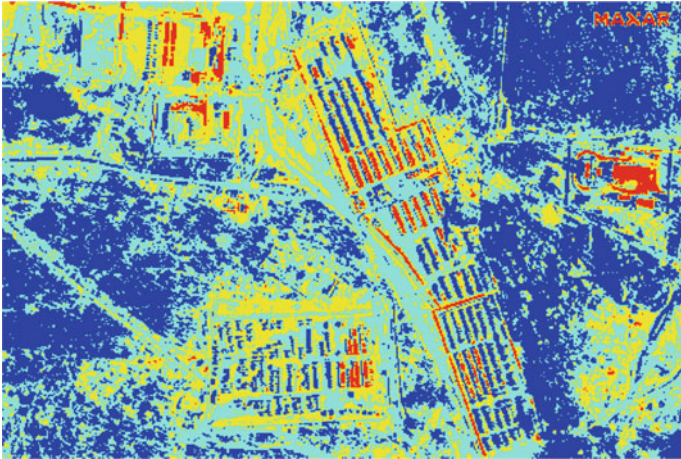


Fig. 4 The result of image segmentation by k-means

2.5 *Quantitative Assessment of Quality of Image Segmentation by the Particle Swarm Algorithm and k-Means*

For quantitative assessment of quality of image segmentation by the PSA and k-means, we will use classical type I and type II errors [36, 37].

The type I segmentation errors are calculated as (12) [4]:

$$\alpha_1 = \frac{A_1(gS(X))}{A_2(g(X))}, \quad (12)$$

The type II segmentation errors are calculated (13) [4]:

$$\beta_2 = 1 - \frac{A_3(gS(X))}{A_4(g(X))}, \quad (13)$$

where $A_1(gS(X))$ – the area of the background on a segmented image that is incorrectly identified as an object of interest;

$A_2(g(X))$ – the total area of the background on the original image;

$A_3(gS(X))$ – the area of objects of interest on the segmented image that are correctly identified as the object of interest;

$A_4(g(X))$ – the total area of objects of interest on the original image.

Results of calculating the values of the type I segmentation errors for these methods are shown in Table 1.

Results of calculating the values of the type II segmentation errors for these methods are shown in Table 2.

Table 1 Results of calculating the values of the type I segmentation errors

Method of image segmentation	$\alpha_1, \%$									
	Number of the iteration of the segmentation									
	I	II	III	IV	V	VI	VII	VII	IX	X
k-means (k = 4)	25,7	26,6	25,9	26,8	25,3	27,1	26,9	25,9	26,8	26,5
The proposed approach to segmentation by the PSA	13,7	15,6	14,5	14,6	13,8	15,5	14,9	14,1	15,3	15,1

Table 2 Results of calculating the values of the type II segmentation errors

Method of image segmentation	$\beta_2, \%$									
	Number of the iteration of the segmentation									
	I	II	III	IV	V	VI	VII	VII	IX	X
k-means (k = 4)	16,6	16,9	15,9	15,8	16,2	16,4	16,9	16,0	17,2	16,8
The proposed approach to segmentation by the PSA	11,7	11,1	11,0	10,6	10,4	10,6	11,0	10,3	10,2	11,2

Values of type I segmentation errors after 10 iteration steps of the considered methods are shown on Fig. 5.

Values of type II segmentation errors after 10 iteration steps of the considered methods are shown on Fig. 6.

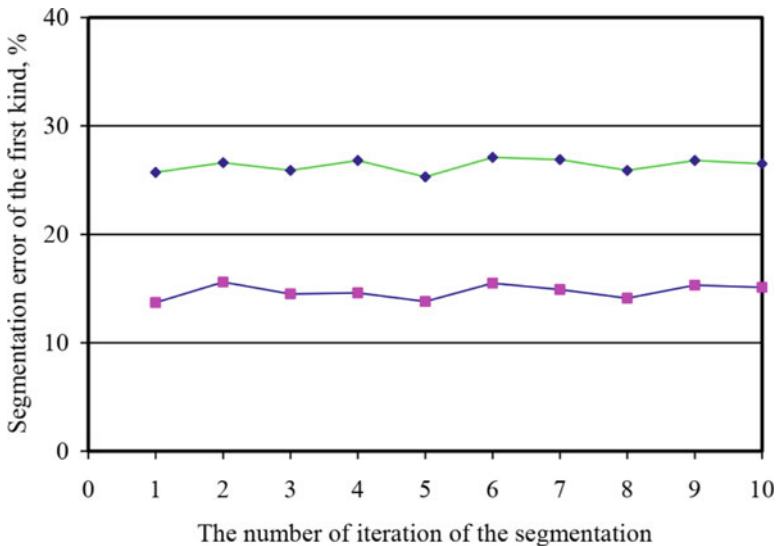


Fig. 5 Values of type I segmentation errors after 10 iteration steps of the considered methods

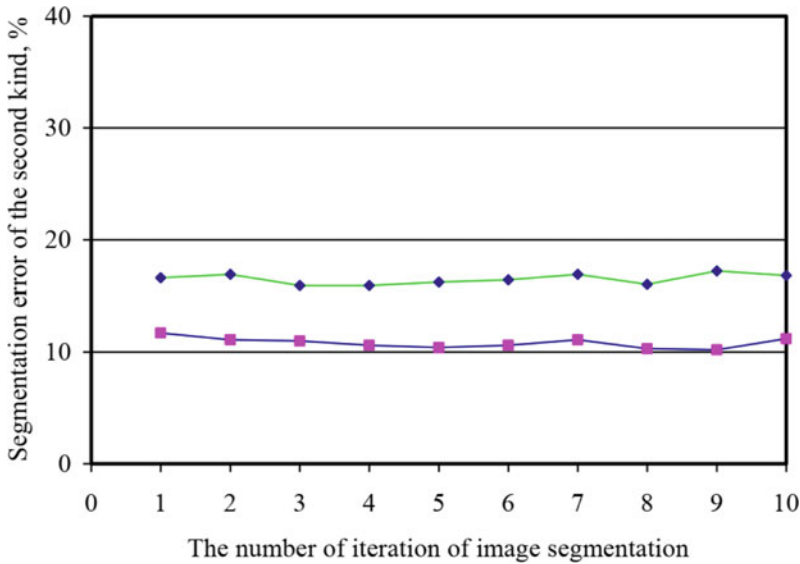


Fig. 6 Values of type II segmentation errors after 10 iteration steps of the considered methods

On Figs. 5 and 6, the lower curve line corresponds to the operation of the k-means method. The upper curve line corresponds to the image segmentation by the proposed approach.

Thus, analysis of Tables 1 and 2, Figs. 5 and 6 showing that the proposed approach to segmentation based on the PSA reduces:

- type I segmentation errors by about 11%;
- type II segmentation errors by about 8%.

3 Conclusions

Thus, the article proposes the application of the PSA to the task of image segmentation for remote sensing of the Earth. The proposed approach to image segmentation uses the PSA in each channel of brightness; divides into segments in each channel of brightness based on the operation of the PSA; combines segmentation results in each channel; gets the overall result of image segmentation.

The results of image segmentation for remote sensing of the Earth using the PSA are presented. Type I and type II segmentation errors are calculated for the proposed approach and the k-means method. The proposed approach to segmentation estimation reduces type I segmentation errors by about 11% and type II segmentation errors by about 8%.

References

1. Richards JA (2021) Remote sensing digital image analysis, 6th edn. Springer, Berlin, p 993
2. Chaminé HI, Pereira AJSC, Teodoro AC, Teixeira J (2021) Remote sensing and GIS applications in Earth and environmental systems sciences. *SN Appl Sci* 3(12):870
3. Pricope N, Mapes K, Woodward K (2019) Remote sensing of human-environment interactions in global change research: a review of advances, challenges and future directions. *Remote Sens* 11(23):2783
4. Ruban I et al (2019) Method for determining elements of urban infrastructure objects based on the results from air monitoring. *East Eur J Enterpr Technol* 4/9(100):52–61
5. Security Defence & Military: Satellite-Based Monitoring. <https://www.eos.com/industries/security-defence-and-military/>
6. Ruban I, et al (2019) The method for selecting the urban infrastructure objects contours. In: 6 International scientific-practice conference problems of infocommunications. Science and technology. Kiev, Ukraine, pp 689–693
7. Ruban I, et al (2021) The development of a forecasting model for the situation based on space images. In: XVI international scientific-technology conference computer science and information technologies (CSIT-2021). Lviv, Ukraine, pp 763–774
8. Fu W, Ma J, Chen P, Chen F (2020) Remote sensing satellites for digital earth. In: Guo H, Goodchild MF, Annoni A (eds) *Manual of digital earth*. Springer, Singapore, pp 55–123
9. Gonzalez RC, Woods RE (2017) *Digital image processing*. Prentice Hall, p 1192
10. Ruban I et al (2019) Construction of methods for determining the contours of objects of tonal aerospace images based on the ant algorithms. *East Eur J Enterpr Technol* 5(101):25–34
11. Cheng G, Han J (2016) A survey on object detection in optical remote sensing images. *ISPRS J Photogramm Remote Sens* 117:11–28
12. Rahman MZU, Jagadeeswar Goud M (2017) Lung cancer detection using marker-controlled watershed transform and K-means clustering. *Int J Magaz Eng Technol Manag Res* 4(1):113–123
13. Johnson BA, Ma L (2020) Image segmentation and object-based image analysis for environmental monitoring: recent areas of interest, researchers' views on the future priorities. *Remote Sens* 12(11)
14. Bhadoria P, Agrawal S, Pandey R (2020) Image segmentation techniques for remote sensing satellite images. *IOP Conf Ser Mater Sci Eng* 993(1):012050
15. Jasim W, Mohammed N (2021) A survey on segmentation techniques for image processing. *Iraqi J Electric Electron Eng* 17(2):73–93
16. Khudov H, et al (2021) The improved mathematical model for interpretation of satellite imagery. In: 8 International scientific-practice conference problems of infocommunications. Science and technology. Kharkiv, Ukraine, pp 384–388
17. El-Baz A, Jiang X, Suru JS (2016) *Biomedical image segmentation: advances and trends*. CRC Press, New York, p 546
18. Zhang B, Rahmatullah B, Wang SL, Zhang G, Wang H, Ebrahim NA (2021) A bibliometric of publication trends in medical image segmentation: quantitative and qualitative analysis. *J Appl Clin Med Phys* 22(10):45–65
19. Liu ZY-C, Chamberlin AJ, Tallam K, Jones IJ, Lamore LL et al (2022) Deep learning segmentation of satellite imagery identifies aquatic vegetation associated with snail intermediate hosts of schistosomiasis in Senegal, Africa. *Remote Sens* 14:1345
20. Saifi MY, Singla J, et al (2020) Deep learning based framework for semantic segmentation of satellite images. In: 2020 fourth international conference on computing methodologies and communication (ICCMC), Erode, India
21. Geetha K (2021) Root CT segmentation using incremental learning methodology on improved multiple resolution images. *J Innov Image Process* 3(04):347
22. Manoharan S (2020) Performance analysis of clustering based image segmentation techniques. *J Innov Image Process* 2(01):14–24

23. Kumar JM, Nanda R, Rath RK, Rao GT (2020) Image segmentation using K-means clustering. *Int J Adv Sci Technol* 29(6s):3700–4370
24. Annadurai P, Kumar LS (2020) Automatic cloud segmentation from INSAT-3D satellite image via improved K-means and improved fuzzy C-means clustering. *Image Process IET* 14(5)
25. Liu B, He S, He D, Zhang Y, Guizani M (2019) A spark-based parallel fuzzy c-means segmentation algorithm for agricultural image big data. *IEEE Access* 7:42169–42180
26. Körting TS, Castejon EF, Fonseca LMG (2013) The divide and segment method for parallel image segmentation. In: *International conference on advanced concepts for intelligent vision systems*. Springer, pp 504–515
27. Jain S, Laxmi V (2017) Color image segmentation techniques: a survey. In: Nath V (Ed) *Proceedings of the international conference on microelectronics, computing & communication systems*. Lecture Notes in Electrical Engineering, vol 453. Springer, Singapore, pp 189–197
28. Lucchese L, Mitra S (2001) Color image segmentation: a state-of-the-art survey, image processing, vision, and pattern recognition. In: *Indian National Science Academy (INSA-A)*, vol 67A, no 2, New Delhi, India, pp 207–221
29. Sarma R, Gupta YK (2021) A comparative study of new and existing segmentation techniques. *IOP Conf Ser Mater Sci Eng* 1022:012027
30. Annadurai P, Sebastian S, Rohith G, Kumar LS (2022) Significant full reference image segmentation evaluation: a survey in remote sensing field. *Multim Tools Appl* 81(6):17959–17987
31. Ruban I et al (2019) Segmentation of optical-electronic images from on-board systems of remote sensing of the earth by the artificial bee colony method. *East Eur J Enterpr Technol* 2/9(98):37–45
32. Kennedy J, Eberhart RC (2001). *Swarm Intelligence*. Morgan Kaufmann. ISBN 1-55860-595-9
33. Ruban I, Khudov V, Makoveichuk O, Khizhnyak I, Khudov H (2018) Swarm method for segmentation of images obtained from on-board optoelectronic surveillance system. In: *5 International scientific-practice conference problems of infocommunications*. Science and technology. Kharkiv, Ukraine, pp 613–618
34. *Satellite Imagery*. <https://www.maxar.com/products/satellite-imagery>
35. Wu T, Gu X, Shao J, Zhou R, Li Z (2021) Color image segmentation based on a convex K-means approach. *Electrical engineering and systems science*. Image and video processing
36. Khudov H et al (2022) Devising a method for segmenting complex structured images acquired from space observation systems based on the particle swarm algorithm. *East Eur J Enterpr Technol* 2/9(116):6–13
37. Costa H, Foody GM, Boyd DS (2018) Supervised methods of image segmentation accuracy assessment in land cover mapping. *Remote Sens Environ* 205:338–351

Overview of Data Center Link Load Balancing Technology Based on SDN



Feifan Hao, Shan Jing, and Chuan Zhao

Abstract The emergence of Internet industries such as big data and cloud computing promotes the rapid development of data centers. The traditional traffic scheduling method is easy to cause load imbalance and link congestion. The concept of elephant and mice flow brings a new idea to the design of load balancing (LB) in data center. In this paper, load balancing technology in data center network link based on software defined network (SDN) technology is summarized. Firstly, this paper classifies the relevant methods of elephant flow detection in data center, and analyzes the advantages and disadvantages of each model. Then, it makes a comprehensive investigation on the rerouting methods of elephant flow and routing optimization strategies. Finally, the paper emphasizes the challenges and future research directions of link load balancing technology in SDN.

Keywords Software defined network · Data center · Elephant flow detection · Link load balancing · Routing optimization

1 Introduction

With the booming of cloud computing, mobile Internet and Internet of Things (IOT), applications and businesses are increasingly diversified, bringing great pressure to the core network, and the traditional Internet has been overwhelmed. Facing this dilemma, SDN leads to a new network architecture: The traditional closed network system is decoupled into data plane, control plane and application plane, the centralized control and management of the network are realized logically [1], its structure is shown in Fig. 1. Nowadays, SDN technology has been applied in many fields, such as data center [2], virtualization [3] and cloud computing [4].

F. Hao · S. Jing (✉) · C. Zhao
School of Information Science and Engineering, University of Jinan, Jinan, China
e-mail: jingshan@ujn.edu.cn

C. Zhao
e-mail: ise_zhaoc@ujn.edu.cn

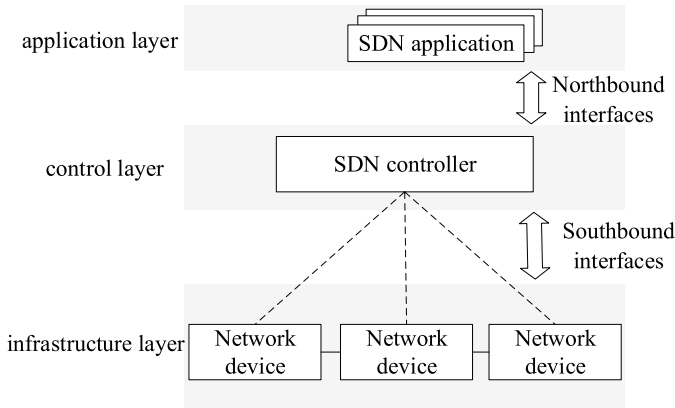


Fig. 1 Three layer structure of SDN

SDN is being widely used in data centers to improve their performance. However, the huge amount of data caused serious network congestion and latency. Therefore, routing optimization is a core problem in networks. The research found that more than 80% of its internal traffic carries a small amount of data and a short duration, while only about 10% of the traffic carries a large amount of information and a long duration. The former is called the mice flow and the latter is called the elephant flow [5]. Then, it was found that the main cause of data center link congestion was elephant flow, which carried large amounts of data and had a long life. Therefore, researchers have found a new direction for the problem of LB on data center links. By using machine learning (ML) and classifier technologies, a model mechanism is deployed in controllers or switches to classify traffic in data centers. The purpose is to detect elephant flow more accurately with the lowest cost. In addition, designed a new rerouting method and optimization algorithm for the detected elephant flow, which greatly alleviated the link congestion and delay, and improved the reliability and efficiency of the data center network.

Many researchers have studied the load balancing technology in SDN before. [6] summarized the classical LB technology in SDN. [7] introduced LB from a machine learning perspective. However, they did not discuss the LB technology in data center network based on SDN in detail, and did not summarize the design of routing optimization strategy from the perspective of discriminating elephant flow.

2 Related Technologies

2.1 SDN Architecture

The SDN architecture proposed by ONF is shown in Fig. 1, which is mainly divided into infrastructure layer, control layer and application layer [8]. The data plane of the infrastructure layer consists of several forwarding devices, it is only used to process and forward data. The SDN controller on the control plane has a global view of the network, it plays a global control role. Different northbound interfaces are provided between the controller and the application layer to implement the interaction between the controller and upper-layer applications. Southbound interfaces are deployed between the controller and the infrastructure layer to control the data plane. The application layer, composed of several SDN applications, interacts with the controller through northbound interfaces to achieve more User-Friendly operations.

At present, SDN has been widely used in data centers and other major service providers, so that the network can support the upper application well. Compared with the traditional network, which can only be configured but not programmed, SDN can flexibly respond to the changes of upper-layer applications and make dynamic adjustments according to the changes, which is the reason why it has been applied in more and more fields.

2.2 OpenFlow

OpenFlow [9] is a new network protocol, which is often used for the southbound interface in SDN. One of the functions of it is to decouples the control function from the forwarding function of basic network devices and centralize the control function to the remote logical controller, while the OpenFlow switch in the data plane is only responsible for local high-speed data forwarding. Figure 2 shows the architecture of the OpenFlow switch.

OpenFlow proposes the concept of flow and flow tables. A flow table consists of several flow entries. A flow entry is the smallest unit that represents a flow rule and the processing method for a particular flow. It consists of six parts: Match Fields, Priority, Counter, Instruction, Timeout and Cookie.

2.3 Data Center Network Based on SDN Technology

With the rapid development of the Internet, data centers are entering the peak of development. According to Cisco Global Cloud Index report, the number of large scale data centers worldwide has increased by nearly 2 times in 5 year [10]. However, due to the inflexible network architecture of traditional networks, SDN has been

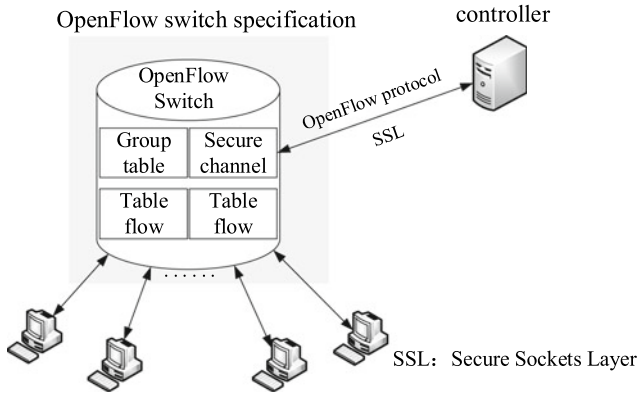


Fig. 2 Architecture of an OpenFlow switch

introduced into cloud data centers by more and more service providers. With the development of SDN, there have been many successful applications, such as Google's deployment of wide area Network B4 of SDN worldwide in 2012. The combination of SDN and data center improves the flexibility and reliability of data center network. Meanwhile, due to its programmable characteristics, SDN reduces operating costs and improves its commercial value.

3 Elephant Flow Detection Mechanism

This section introduces the classical method of detecting heavy traffic and the new elephant flow detection mechanism in data center network.

3.1 Traditional Traffic Detection Mechanism

Early classical flow detection model mechanisms include Hedera [11], sFlow [12], Mahout [13], etc. Hedera [11] is an extensible dynamic flow scheduling system, which detects elephant flow by regularly extracting flow information in switches, and adaptively schedules multi-level exchange structures to effectively utilize aggregated network resources. This model has a central scheduler to understand the global state, and its ability to schedule network resources is superior to the HASH based ECMP [14] load balancer at that time. But Hedera only considers bandwidth limits, which can lead to traffic imbalance. As the switch does not directly provide the sampling and detection functions of large traffic, Afek et al. [12] proposed sFlow, which sampled data packets flowing through SDN switches based on openFlow method. Their sets two thresholds by taking into account both the limited rule space in the switch and

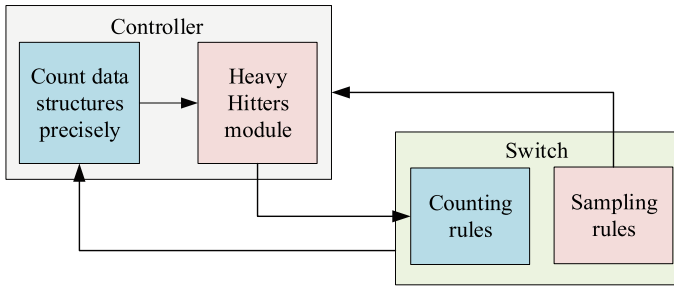


Fig. 3 Structure of the Sample&Pick

the time it takes for the controller to install the rule in the switch. The two threshold is used to determine the big flow and detecting potential flow, this method is called Sample&Pick, its structure as shown in Fig. 3. Among them, the Heavy Hitters module is used to receive samples from the controller to identify suspicious heavy flows. When the flow counter in the module exceeds a preset threshold, a rule is inserted into the switch to maintain the flow packet counter.

Curtis et al. [13] proposed Mahout, a mechanism for elephant flow detection in a terminal host, rather than in the network as described above. They consider that detecting traffic behavior in network will lead to high time cost, resource consumption and bandwidth cost, so they using socket buffer monitoring in terminal hosts to detect elephant flow. This avoids errors due to network congestion, and Mahout has been tested with low overhead on commercial servers. However, monitoring modules need to be installed to detect elephant flows at terminals, so the deployment cost is relatively high.

3.2 Mechanism and Model of Elephant Flow Detection

Many modern algorithms and theories are used to detect elephant flows. At present, many researchers designed LB algorithms by using supervised learning, unsupervised learning and reinforcement learning, so as to improve the overall performance of the network.

Chao et al. [16] adopted data stream mining technology to detect elephant flow and proposed the FlowSeer model. By extracting the features of the first five packets of the flow, the traffic classification model is trained. The CVFDT algorithm is used in the model, which does not need to store samples compared with the traditional decision tree algorithm and realizing the function of real-time detection. In addition, it enables the switch and controller to collaborate on detection, thus reducing latency and the burden on the controller. Yan [17] made improvements on the basis of the FlowSeer model, considering that the FlowSeer did not take into account the error and missed detection of elephant flow, and also ignored the problem of low accuracy

caused by the imbalance between elephant flow and mice flow. Firstly, SMOTE [18] is used for data balance sampling. Then, introduce dynamic weighting method to avoid error and missing detection, set the cost matrix for dynamic weighting detection of elephant flow, and finally get flow-completion time (FCT) as high as 90%. Huang et al. [19] is based on the idea of classification, through a switch and controller respectively the deployment of two classifiers to detect the elephant flows. In addition, the concept of application theory is introduced, so that the method has better FCT and recall rate.

Sampling-Based elephant flow detection is a common method to distinguish elephant and mice flow. Mori et al. [15] designed a scheme to identify elephant flow through periodic sampling based on Bayesian theory. The Bayesian theory is used to obtain the threshold of each flow packet, and the threshold is used to determine whether a flow belongs to elephant flow in unsampled packets. This method not only has high detection efficiency, but also ensures its simplicity of operation, so it can be easily implemented on most devices. Tang et al. [20] construct ESCA method, aiming to reduce time and bandwidth overhead as much as possible while detecting. The model search the most efficient sampling period and distinguishes elephant flow based on flow correlation. This method improves accuracy and reduces cost. Xiao et al. [21] introduced cost-sensitive learning methods to detect elephant flow. A cost-sensitive decision tree was constructed to implement the detection strategy, and decision tree indexes were proposed to minimize the cost of classification errors.

Throughput is the maximum rate a device can accept without frame loss. It is an important indicator of network performance. The performance of the entire network can be improved by improving the throughput through methods. Liu et al. [22] designed a new data structure to detect elephant flow using Bloom filter, namely RML-HCBF. In this model, elephant flow ids are constructed by overlapping hash bit strings, so that it can be identified without storing flow ids, thus improving the throughput.

Curtis et al. [23] proposed the DevoFlow mechanism, where they install triggers in switches to detect elephant flows by extracting key fields. DevoFlow improves the performance of equal-cost multipath routes on irregular topologies and uses wildcard rules to reduce the number of interactions between switches and controllers. This mechanism also has the function of traffic management [24]. However, DevoFlow requires the maintenance of flow table entries for each flow, thus placing a significant burden on the switch. Lin et al. [25] proposed the ESHSP mechanism by using the combination of single statistic information and aggregation in OpenFlow protocol to detect elephant flow. The function of elephant flow storage and range segmentation is added to improve efficiency by storing known elephant flow information to avoid double calculation.

Tang et al. [26] adopted an Autodetect Upload (ADU) mechanism to detect elephant flow. ADU mechanism consists of ADU-Client and ADU-Sever. The former is deployed on the host and the latter runs on the SDN controller. When detecting the elephant flow on the host, the ADU-Client generates a forged source IP address packet, and triggers the Packet_in message on the edge switch to report the elephant

flow to the SDN controller. Then, the ADU-Sever module completes the identification of the elephant flow. The detection time cost of this mechanism is obviously reduced and the efficiency is further improved.

SDN performs network operations by configuring corresponding network policies in the switch by the controller, and the storage of flow tables and traffic is usually done in Ternary Content Addressable Memory (TCAM) [27]. TCAM has high cost, energy consumption and limited storage space. Therefore, Liu et al. [28] considered using flow matrix and reasoning method to detect traffic in the network, so as to save storage space to a certain extent. Instead of monitoring each traffic, they only identify large traffic from multiple historical traffic matrices. GBM (Gradient Boosting Machine) [29] is an effective method to explore the correlation of sequential network traffic data through machine learning, which is used to learn features in multiple historical traffic matrices, so as to effectively detect large traffic. Although elephant flow detection is not mentioned in this paper, the idea of detecting and distinguishing large flow belongs to the same strategy as the former.

HAMDAN et al. [30] believe that current elephant flow detection technologies all adopt pre-set thresholds, which cannot be extended with the change of flow distribution. They proposed a dual-classifier detection method, each classifier is deployed on SDN switches and controllers respectively. When detecting elephant flow, CM Sketch classifier [31] will be used at the switch to filter mice flow, and decision tree VFDT classifier will be used at the controller to classify the flow. In the mininet experiment, the detection accuracy of this model is as high as 98.13%. Bi et al. [32] proposed an adaptive elephant flow detection system divided into two stages. Through the sampling detection method, the optimal threshold is designed according to the change of dynamic flow, which is more suitable for the dynamic network environment. Table 1 shows the methods for detecting elephant flow summarized in this section:

Table 1 Comparison of elephant flow detection methods

Model	Type	Methods	Advantages
sFlow [12]	Sampling detection	Sample&Pick	High accuracy, low overhead
The [15]	Sampling detection	Bayesian theory	Efficient and easy
Mahout [13]	Feature extraction detection	Detection at the terminal host	Faster and accurate
FlowSeer [16]	Feature extraction detection	Stream mining technology	Detect in real time
DevoFlow [23]	Feature extraction detection	Wildcard rules	Improved ECMP and reduce overhead
ADU [26]	Threshold detection	Through flow tables	Efficient and low cost
The [30]	Threshold detection	Decision Trees	High accuracy

4 Rerouting and Routing Optimization of Elephant Flow

The previous chapter mainly describes the mechanism of detecting elephant flow that causes link congestion. Starting from the detected elephant flow, this chapter discusses the design of rerouting strategy based on modern optimization algorithm and machine learning algorithm and other link routing optimization methods respectively.

4.1 *Routing Optimization Based on Modern Optimization Algorithm*

Modern optimization algorithms are designed to solve the problem of how to obtain the optimal solution in complex problems. These algorithms have been widely used in many commercial fields. Many optimization algorithms have been applied to route optimization in network LB.

Ant Colony Optimization Algorithm (ACO) simulates the behavior of ants searching for the optimal path. Ants release pheromones along multiple foraging paths, and the better the path, the higher the pheromone concentration, so as to find the optimal path to the food source. Wang et al. [33] proposed an ACO based on scheduling system—TSACO model. Firstly, elephant flow is detected through Open-Flow and sFlow. Then, ACO and K-Path algorithms are used to segment elephant flow, and finally the routing strategy of it is delivered to the switch. The remaining mice flow are forwarded according to the paths in the path database. Hamdan et al. [34] also proposed DPLBAnt model by using ACO. Different from the former, their model firstly obtains the global state of SDN and finds the congested path, and uses ACO to redirect the elephant flow in the path and forward it to the optimal path. As elephant flow is rerouted, link congestion is relieved and controller load is reduced.

Li et al. [35] proposed GA-ACO model. Firstly, congested links are found according to the global view, and then Genetic Algorithm (GA) is used to calculate multiple candidate paths according to the actual link utilization rate of the current network. As the input of the ACO, these paths have higher initial pheromone values than other paths. Finally, the ACO finds the optimal path to reroute the elephant flow. The combination of GA and ACO solves the problems of slow convergence speed and lack of early pheromone, and greatly improves the utilization rate of link. Hedera [11] also has the function of rerouting elephant flow. When elephant flow is detected, the controller uses Simulated Annealing Algorithm to schedule it to a new path. The algorithm performs probabilistic search to efficiently compute the flow path, and further optimizes it to reduce the search space to ensure fast convergence to near optimal solution.

In order to solve the problem of high processing delay in IoT, Darade et al. [36] proposed a new optimization strategy to solve the LB problem in terms of delay by using Whale Optimization Algorithm. By combining fog computing with

SDN, computing and storage functions are extended to the edge of the network, which greatly reduces latency. In order to optimize the selection of load distribution coefficient, they proposed a new threshold whale optimization algorithm–T-WOA, which improved the convergence speed of the algorithm by introducing dynamic threshold, thus optimizing the selection of load distribution coefficient. This paper tries to solve the problem of network LB from the angle of reducing latency, which is a good new idea.

The traditional multipath forwarding strategy cannot consider the link transmission status in real time. In view of this phenomenon, Xu et al. [37] proposed a LB algorithm based on Spider Monkey Optimization (SMO-LBA). It has the function of adaptive global exploration. Firstly, it obtains idle links in the data center network, then uses the algorithm to perform adaptive evaluation and update on the fitness values of each path, and finally selects the path with the lowest link occupancy rate as the global optimal path to deliver the flow table rules for forwarding. The algorithm has the advantage of real-time awareness of the details of the underlying network and adaptive global search ability, which improves the intelligence of the traditional algorithm to realize the local optimal feature, so as to improve the overall link utilization.

Route optimization can improve routing efficiency, but elephant flow error detection and traffic loss may occur due to fuzzy range between elephant and mice flows. As mentioned earlier, some methods take into account the problem of elephant flow error detection and missing detection, but these problems should still be taken seriously in the design of subsequent methods.

The methods in this section are summarized in Table 2.

Table 2 Routing optimization based on the optimization algorithm

Model	Optimization algorithm	Function	Advantages
TSACO [33]	Ant colony algorithm	Segmented flow	Find the optimal path
DPLBAnt [34]	Ant colony algorithm	Rerouted	Reduce controller load
GA-ACO [35]	Genetic algorithm Ant colony algorithm	Rerouted	Convergence speed is accelerated
Hedera [11]	Simulated annealing algorithm	Elephant flow reroute	Maximize utilization
T-WOA [36]	Whale optimization algorithm	Reduce latency	High load can be solved by reducing latency
SMO-LBA [37]	Spider monkey optimization algorithm	Improving link Utilization	Avoid local optimal

4.2 Routing Optimization Based on Machine Learning

After more than half a century of development and evolution, artificial intelligence has been widely applied in various fields. Machine learning, as a field of artificial intelligence, can learn from a wide range of data to make decisions, identify different patterns and perform corresponding operations without human intervention. This section will discuss three machine learning methods: supervised learning, unsupervised learning, and reinforcement learning (RL) to design network routing optimization strategies.

Cui et al. [38] designed a LB model that can adjust network state in real time by training BP artificial neural network model to predict load. By using the SDN global view, collected in each path bandwidth utilization, transmission delay, packet loss rate and the transmission hop these four load characteristics, to train the BP artificial neural network to forecast load in different paths, the final chosen a minimum load as the data flow transmission path. This strategy can select transmission path for new incoming streams and adjust network congestion in real time. In the experiment, compared with Static Round Robin strategy previously proposed [39], network latency is reduced by 19.3%.

Kumar et al. [40], from the perspective of unsupervised learning, they used k-means and cosine similarity method to create a model to select the least congested path in the list. Clustering is a kind of unsupervised learning. Its function is to classify the data with similar characteristics in the data set. K-means is a famous partition clustering algorithm, which has the characteristics of simplicity and efficiency. This model consists of two modules, training and deployment. The training module learns the behavior of the path from the latest network state. The deployment module search the optimal path in the network according to the time interval specified by the controller, and updates the optimal path in real time according to the received information. The advantage is that finite and constant comparison times are needed to get the optimal path, which solves the problem of different congestion factors in different links. By finding the path with the lowest congestion, the link status of the network is improved.

Fu et al. [41] considered that the routing strategy in SDN controller relies on manual design and is difficult to reach the optimal state in dynamic network environment, so they proposed a routing strategy based on deep Q-learning to automatically generate the optimal path. In view of the different requirements of elephant flow and mice flow in data center, this strategy carries out deep Q network training for them respectively, so as to realize high throughput and low packet loss of elephant flow, the low latency and low packet loss of mice flow. Ryu controller was selected for the experiment in mininet, and the experiment showed that the routing policy proposed by them could intelligently provide the optimized routing policy according to the specific situation, so as to improve the network performance.

Jha et al. [42] proposed a weighted multipath routing scheme based on deep reinforcement learning (DRL) on the basis of ECMP. RL framework can be modeled

as Markov decision process, and the training of agent is carried out by deep Q-learning algorithm. When the RL agent is configured with a link with the best weight, the controller uses Dijkstra shortest path first algorithm to select the best path. This scheme can estimate the future traffic periodically and reconfigure the network topology in the data center network. Compared with ECMP in experiments, the scheme improved latency and throughput clearly.

Todorov et al. [43] proposed a LB mechanism combining segmented routing algorithm and machine learning to reduce bandwidth and improve routing mechanism. In the prediction process, Q-learning algorithm is used for data prediction and training to cope with the possible changes of network load. When the load changes, the prediction module will send a signal to the path calculation module to update the flow table, and the optimal path will be sent back and installed to the switch. This model can minimize the traffic between controller and network device to improve network performance.

In addition to the LB methods discussed above, load balancers in the network may face single point of failure. The solution to this problem is to connect them to form a cluster, work cooperatively and supervise each other, so that can improve the stability of LB technology in data center links through this redundant method. Similarly, the routing optimization method based on ML should also consider the fault in the routing process, so that the fault node can be detected in time when problems occur.

5 Summary of Other Route Optimization Methods

This chapter summarizes and discusses some routing optimization strategies and load balancing models which are researched and designed in combination with other theories in many fields.

Zhang et al. [44] applied the stable matching theory to model the traffic scheduling and proposed the Fincher model. They improved the Gale-Shapley algorithm so that when a flow is rejected, any erratic flow will not be received by the switch, even if there is sufficient capacity. In the process of scheduling elephant flow, the stable matching theory is applied to alleviate the conflict between switch and traffic, so as to reduce the congestion degree. By obtaining the most stable matching between elephant flow and core switches, the flow is matched to the appropriate path to achieve the best performance of the network. Cui et al. [45] adopted the method of elephant flow segmentation to improve the average link utilization rate of data center network and improve network performance. This method improves the particle swarm optimization algorithm to alleviate the problem of falling into local optimum.

Song et al. [46] optimized network performance by dealing with network bandwidth fragments. Existing policies do not consider this, which reduces the transmission rate and bandwidth utilization. They proposed Ashman, which is a dynamic scheduling scheme based on flow. Ashman-Probfit in the scheme is an algorithm based on heuristic probability, which effectively balances bandwidth fragmentation

and overload traffic. Huang et al. [47] proposed EAshman, which based on Ashman and further improve performance. This method only rescheduled the elephant flow on the congested link and adjusted the polling period algorithm to adaptive, so that the rotation period could be adjusted effectively according to the dynamic load of the current network. So this approach further reduces overhead and improves throughput.

Quality of Service (QoS) is used to evaluate the ability of service providers to meet customer requirements. To ensure appropriate QoS to meet the needs of delay-sensitive applications in SDN networks, Dukkupati et al. [48] proposed a distributed solution, the Sieve architecture. Compared with querying all switches, this method only polls edge switches periodically to obtain network information, and reschedule only detected elephant flows, thus reducing the sampling and scheduling burden between data and control plane. This approach pays attention to FCT, which is an important user performance metric [49]. In mininet, Sieve outperforms the ECMP and Hedera in FCT of mice flow. HIKICHI et al. [50] deploying SDN controllers in a distributed manner to jointly handle the load of applications. The external load balancer uses the method of polling scheduling weight to distribute application requests between servers to reduce the amount of messages between servers and improve the overall performance of the controller.

Zang et al. [51] proposed the SFRM mechanism to reschedule congested elephant flows while maintaining high throughput. The elephant flow selection module is deployed in the controller to ensure the fast routing mechanism. This method improves the traditional static flow hashing method, thus improving bandwidth utilization and throughput. Zhang et al. [52] designed the DIFFERENCE method. This is a differential scheduling method, which can dynamically set the paths for elephant and mice flows according to the current link load state. This method reduces the path search space and ensures the bandwidth requirements.

6 Future Prospects and Conclusions

In this paper, the elephant flow detection mechanism in data center network based on SDN technology is summarized and discussed. In this process, the methods of sampling detection, elephant flow threshold detection and feature extraction detection are discussed from three perspectives. After the elephant flow is detected, this paper summarizes the routing optimization scheme by using modern optimization algorithm and machine learning algorithm. Finally, this paper summarizes the use of other field theories and new methods to design routing optimization strategy, so as to have a more comprehensive understanding of the data center network link mechanism.

With the rapid development of Internet technology, the traffic in the data center network will further increase, so the network load balancing technology has a long way to go. On the basis of ensuring efficiency, we should always keep up with the pace of the development of the times to make dynamic corresponding and innovation. Link load balancing will face the following challenges in the future:

(1) Security issues in load balancing

SDN has the advantage of centralized control, and the controller can adjust the network through the global network view. Many LB and routing optimization strategies make use of the centralized control capabilities of SDN, but network security and malicious attack prevention are not considered at the beginning of mechanism design, which also enables malicious attackers to find network flaws and loopholes. Despite the capabilities in the SDN, they hold certain limitations in terms of the information security and the quality of service [53]. Therefore, strengthening SDN network security is still an important topic.

(2) Combination of load balancing and virtualization

The structure of a large number of network virtualization platforms is centrally controlled, so SDN technology is more suitable for virtualization deployment in the network. When designing LB algorithms for data centers, it is also necessary to timely respond to network virtualization requirements and reserve extensible interfaces in advance for future upgrades.

(3) Load balancing in 5G networks

The data center network based on SDN is gradually realizing the transition from 4 to 5G network. In the process of designing the LB mechanism of SDN data center network, we should pay more attention to the integration with 5G network technology.

Acknowledgements This work was supported by Natural Science Foundation of Shandong Province (No. ZR2019LZH015, No. ZR2021LZH007), Projects of Ministry of Education Industry-University Cooperation Education (No. 202101103019, 201901234008, 201901166007), and Project of Shandong Postgraduate Education Teaching Case (No. SDYAL20119).

References

1. Wang M (2016) Software defined networking: security model, threats and mechanism. *J Softw* 4:24
2. Ghobadi M, Yeganeh SH, Ganjali Y (2012) Rethinking end-to-end congestion control in software-defined networks. In: *Proceedings of the 11th ACM Workshop on Hot Topics in networks*, pp 61–66
3. Mijumbi R, Serrat J, Gorricho JL et al (2015) Network function virtualization: state-of-the-art and research challenges. *IEEE Commun Surv Tutor* 18(1):236–262
4. Paul S, Jain R (2012) Openadn: Mobile apps on global clouds using openflow and software defined networking. In: *2012 IEEE Globecom Workshops*, pp 719–723. IEEE
5. Xi K, Liu Y, Chao HJ (2011) Enabling flow-based routing control in data center networks using probe and ECMP. In: *2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPs)*, pp 608–613. IEEE
6. Hamdan M, Hassan E, Abdelaziz A et al (2021) A comprehensive survey of load balancing techniques in software-defined network. *J Netw Comput Appl* 174:102856
7. Amin R, Rojas E, Aqduş A et al (2021) A survey on machine learning techniques for routing optimization in SDN. *IEEE Access* PP(99):1
8. ONF. SoftwareDefined Networking (SDN) Definition. <https://www.opennetworking.org/sdn-resources/sdn-definition>

9. Dave T (2014) OpenFlow: enabling Innovation in campus networks. *ACM SIGCOMM Comput Commun Rev* 38(2):675–690
10. Cisco (2016) Cisco global cloud index: Forecast and methodology, 2015–2020[EB/OL], 01 June 2016. <http://www.audentia-gestion.fr/cisco/white-paper-c11-738085.pdf>.
11. Al-Fares M, Radhakrishnan S, Raghavan B et al (2010) Hedera: dynamic flow scheduling for data center networks. In: *Proceedings of the 7th USENIX Symposium on Networked Systems Design and Implementation, NSDI 2010*, 28–30 April 2010, San Jose, CA, USA. DBLP
12. Afek Y, Bremner Barr A, Landau Feibish S et al (2015) Sampling and large flow detection in SDN. In: *Proceedings of the 2015 ACM Conference on Special Interest Group on Data Communication*, pp 345–346
13. Curtis A R, Kim W, Yalagandula P (2011) Mahout: low-overhead datacenter traffic management using end-host-based elephant detection. In: *2011 Proceedings IEEE INFOCOM*, pp 1629–1637. IEEE
14. Hopps C (2000) Analysis of an equal-cost multi-path algorithm. RFC 2992, Internet Engineering Task Force
15. Mori T, Uchida M, Kawahara R et al (2004) Identifying elephant flows through periodically sampled packets. In: *Proceedings of the 4th ACM SIGCOMM Conference on Internet Measurement*, pp 115–120
16. Chao SC, Lin KCJ, Chen MS (2016) Flow classification for software-defined data centers using stream mining. *IEEE Trans Serv Comput* 12(1):105–116
17. Yan K (2021) Research and implementation of SDN flow table optimization based on machine learning. *Beijing Univ Posts Telecommun*. <https://doi.org/10.26969/d.cnki.gbydu.2021.000472>
18. Bhowmick K, Narvekar M, Bhimdiwala A et al (2018) CDACI: concept drift detection and adaptation to classify imbalanced data streams. In: *Proceedings of the 2018 IEEE Punecon*, pp 1–5. IEEE
19. Huang YH, Shih WY, Huang JL (2017) A classification-based elephant flow detection method using application round on SDN environments. In: *2017 19th Asia-Pacific Network Operations and Management Symposium (APNOMS)*, pp 231–234. IEEE
20. Tang F, Zhang H, Yang LT et al (2019) Elephant flow detection and differentiated scheduling with efficient sampling and classification. *IEEE Trans Cloud Comput* 1
21. Xiao P, Qu W, Qi H et al (2015) An efficient elephant flow detection with cost-sensitive in SDN. In: *2015 1st International Conference on Industrial Networks and Intelligent Systems (INISCom)*, pp 24–28. IEEE
22. Liu W, Qu W, Liu Z et al (2012) Identifying elephant flows using a reversible multilayer hashed counting bloom filter. In: *2012 IEEE 14th International Conference on High Performance Computing and Communication & 2012 IEEE 9th International Conference on Embedded Software and Systems*, pp 246–253. IEEE
23. Curtis AR, Mogul JC, Tourrilhes J et al (2011) DevoFlow: scaling flow management for high-performance networks. In: *Proceedings of the ACM SIGCOMM 2011 Conference*, pp 254–265
24. Mogul JC, Tourrilhes J, Yalagandula P et al (2010) Devoflow: cost-effective flow management for high performance enterprise networks. In: *Proceedings of the 9th ACM SIGCOMM Workshop on Hot Topics in Networks*, pp 1–6
25. Lin CY, Chen C, Chang JW et al (2014) Elephant flow detection in datacenters using openflow-based hierarchical statistics pulling. In: *2014 IEEE Global Communications Conference*, pp 2264–2269. IEEE
26. Tang Q, Zhang H, Dong J et al (2020) Elephant flow detection mechanism in SDN-based data center networks. *Sci Program* 2020:1–8
27. Alsaeedi M, Mohamad MM, Al-Roubaiey AA (2019) Toward adaptive and scalable OpenFlow-SDN flow control: a survey. *IEEE Access* 7:107346–107379
28. Liu G, Guo S, Xiao B et al (2019) SDN-based traffic matrix estimation in data center networks through large size flow identification. *IEEE Trans Cloud Comput* 10(1):69–74
29. Friedman JH (2001) Greedy function approximation: a gradient boosting machine. *Ann Stat* 1189–1232

30. Hamdan M, Mohammed B, Humayun U et al (2020) Flow-aware elephant flow detection for software-defined networks. *IEEE Access* PP(99):1
31. Cormode G, Muthukrishnan S (2005) An improved data stream summary: the count-min sketch and its applications. *J Algorithms* 55(1):58–75
32. Bi C, Luo X, Ye T et al (2013) On precision and scalability of elephant flow detection in data center with SDN. In: *Proceedings of the 2013 IEEE Globecom Workshops (GC Wkshps)*, pp 1227–1232. *IEEE*
33. Wang C, Zhang G, Chen H et al (2017) An ACO-based elephant and mice flow scheduling system in SDN. In: *Proceedings of the 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA)*, pp 859–863. *IEEE*
34. Hamdan M, Khan S, Abdelaziz A et al (2021) DPLBAnt: improved load balancing technique based on detection and rerouting of elephant flows in software-defined networks. *Comput Commun* 180:315–327
35. Li H (2020) An optimal and dynamic elephant flow scheduling for SDN-based data center networks. *J Intell Fuzzy Syst* 38(1):247–255
36. Darade SA, Akkalakshmi M (2021) Load balancing strategy in software defined network by improved whale optimization algorithm. *J High Speed Netw* 2021(P reprint):1–17
37. Xu H (2021) Data center adaptive multi-path load balancing algorithm based on software defined network. *J Comput Appl* 41(04):1160–1164
38. Chen-Xiao C, Ya-Bin X (2016) Research on load balance method in SDN. *Int J Grid Distrib Comput* 9(1):25–36
39. Li Y, Pan D (2013) OpenFlow based load balancing for fat-tree networks with multipath support. In: *Proceedings of the 12th IEEE International Conference on Communications (ICC 2013)*, Budapest, Hungary, pp 1–5
40. Kumar S, Bansal G, Shekhawat VS (2020) A machine learning approach for traffic flow provisioning in software defined networks. In: *2020 International Conference on Information Networking (ICOIN)*, pp 602–607. *IEEE*
41. Fu Q, Sun E, Meng K et al (2020) Deep Q-learning for routing schemes in SDN-based data center networks. *IEEE Access* 8:103491–103499
42. Jha A, Singh KK, Devi KV et al (2021) Reinforcement learning based weighted multipath routing for datacenter networks. *Materials Today: Proceedings*
43. Todorov D, Valchanov H, Aleksieva V (2020) Load balancing model based on machine learning and segment routing in SDN. In: *2020 International Conference Automatics and Informatics (ICAI)*, pp 1–4. *IEEE*
44. Zhang Y, Cui L, Chu Q (2015) Fincher: elephant flow scheduling based on stable matching in data center networks. In: *2015 IEEE 34th International Performance Computing and Communications Conference (IPCCC)*, pp 1–2. *IEEE*
45. Cui X, Meng Q, Wang W (2020) A load balancing mechanism for 5G data centers. In: *2020 International Wireless Communications and Mobile Computing (IWCMC)*, pp 812–815. *IEEE*
46. Song T, Liu Y, Wang Y et al (2017) Ashman: a bandwidth fragmentation-based dynamic flow scheduling for data center networks. *Comput J* 60(10):1498–1509
47. Huang B, Dong S (2020) An enhanced scheduling framework for elephant flows in SDN-based data center networks. In: *2020 IEEE Symposium on Computers and Communications (ISCC)*, pp 1–7. *IEEE*
48. Zaher M, Alawadi AH, Molnár S (2021) Sieve: a flow scheduling framework in SDN based data center networks. *Comput Commun* 171:99–111
49. Dukkupati N, McKeown N (2006) Why flow-completion time is the right metric for congestion control. *ACM SIGCOMM Comput Commun Rev* 36(1):59–62
50. Kenji H, Toshio S (2016) Dynamic application load balancing in distributed SDN controller. *IEICE Proc Ser* (25):TS4–4
51. Zang W, Jin Z, Lan J (2017) An SDN based fast rerouting mechanism for elephant flows in DCN. In: *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, pp 363–366. *IEEE*

52. Zhang H, Tang F, Barolli L (2019) Efficient flow detection and scheduling for SDN-based big data centers. *J Ambient Intell Humaniz Comput* 10(5):1915–1926
53. Anand JV (2019) Design and development of secure and sustainable software defined networks. *J Ubiquitous Comput Commun Technol (UCCT)* 1(02):110–120

Wearable Band for Safety in Chemical Industries



D. Diana Josephine, R. Ajay Kumar, M. Ganesamoorthi, A. Meshwin, and M. Athiq Ahmed

Abstract Industrial gas leakage causes accidents and poses several threats to the environment and, many innocent human lives. It is essential to detect the gas leaks in time, and help employees from losing their hope; getting frightened and the industrialists from economy loss. This paper aims for the safety and protection of the employees working in the industry and make them to escape during the adverse situation. The wearable band comes with emergency alarms interfaced with Node MCU, hazardous gas detection sensors, navigators to navigate to the nearest exit path and navigate to different places inside the industry based on his/her requirements using inbuilt GPS. A panic button in the band helps the people in the industry during an abnormal condition and alert information will be sent to the authorities to take safety precautions. The RF-ID tag helps to register the employee attendance during the entry to and exit from the industry using barcode. The wearable band provides a complete guide for an employee and visitors to the industry with guaranteed safety measures.

Keywords Wearable band · Node MCU · GPS · RF-ID

1 Introduction

A chemical accident is the unintentional release of one or more chemical hazardous substances that could harm human health and the environment. Such events include fires, explosions, leakages, or the release of toxic or hazardous materials. With economic growth, the use of chemicals has continually increased, resulting in an increase in chemical accidents. Gas leaks are a major problem and today it is common in many places such as residential areas, industries, and vehicles such as Compressed Natural Gas (CNG), buses, cars, etc. It is noteworthy that because of gas leaks, dangerous accidents occur. Liquid petroleum gas (LPG) is a combustible mixture of

D. D. Josephine (✉) · R. A. Kumar · M. Ganesamoorthi · A. Meshwin · M. A. Ahmed
Department of Electronics and Communication Engineering, Coimbatore Institute of Technology,
Coimbatore, Tamil Nadu, India
e-mail: dianajosephine@cit.edu.in

hydrocarbon gas and it is very hot and can cause burn even from a leak source. This power source is mainly composed of propane and butane which are highly flammable chemical combinations. These gases can easily catch fire. The gas leaks have caused a lot of economical loss and most people have lost their life. So it is very important to detect the gas leaks and prevent people from groaning. Due to improper man-made operation or equipment aging, a large number of gas leaks have occurred. Some leaks that are detected and repaired in time have prevented great economic losses, environmental pollutions, and even huge casualties.

The paper is outlined as follows. In Sect. 2, works related to safety measures in chemical industry is presented. In Sect. 3 the proposed system model is illustrated. The results and discussions are mentioned in Sect. 4 followed by conclusion and future scope in Sect. 5.

2 Related Works

Determination of the bearing angle of unobserved ground targets by use of seismic location cells [1]: In this paper, location of the moving unobserved armored trucks (1 + 2) kilometers away is considered based on the seismic location method. The bearings are computed from the north direction which is 0° , 90° is east, 180° is south, and 270° is west. The shortest distance between source and destination is calculated by using the Haversine formula. Results show that it has less measurement errors but takes quite a long time for detection.

Detection of hazardous gas using wearable Wireless Sensor Networks for Industrial Applications [7]: In this paper, a unique framework based on Wireless Sensor Networks (WSN) has been deployed. This helps to collect the physical parameters and to overcome the difficulties involved in the conventional system in which the precision rate of the gas leakage was not appropriate. The rate of accuracy is increased by employing priority leach and Fuzzy logic algorithms. The system monitors and measures the concentration level of hazardous gas like carbon monoxide (CO), methane (CH₄) and provide necessary alerts to the wearer when level becomes vulnerable.

Barcode Recognition System [3]: In this paper, a method to scan a barcode using contrast enhancement, image processing and edge detection is proposed. Image that contains barcode information is captured using a camera. The colour image contains complete and useful information. The image is then converted to the grayscale format. The image is converted to a pre-processed grey image to reduce noise and improve image brightness between bars and spaces. An edge-finding algorithm is used to determine the boundaries. Slider is used to control the brightness of the image. The demerit of this paper is that it does not have read or write capabilities.

3 Proposed Model

Figure 1 shows the proposed model for the wearable band. The model is divided into three sections:

- i. Hazardous gas detection
- ii. Navigation
- iii. ID scanner

3.1 Hazardous Gas Detection

The hazardous gas detection section detects the gas leakage using MQ-2, MQ-3, and MQ-9 sensors [9]. These sensors will detect the gas leaks and send an alert message to the safety manager of the industry [2, 5].

A switched-mode power supply is used to convert 230 V alternating current to 12 V alternating current. Power board version 2 is used to convert 12 V alternating current to 12 V direct current. 3 sensors are used; MQ-2 sensor for detecting smoke; MQ-3 sensor for detecting alcohol; MQ-9 sensor for detecting methane, and a buzzer. The sensors are interfaced with Node MCU. 2 of the sensors are connected to the 12 V pin through voltage regulators and the remaining sensor to the 5 V pin. The 3.3 V input for the NODE MCU is given through the power board. When any of the given sensors gets activated, the buzzer alarms and alert message will be sent to the safety manager using the push bullet application. This application is available in play store of all Android phones. Log in has to be created in the official push bullet website using username and password. It will generate an Application Programmable

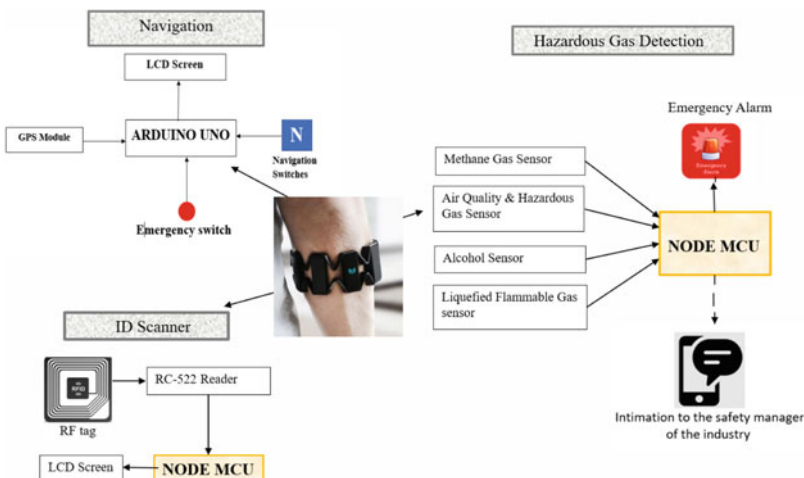


Fig. 1 Proposed model

Table 1 Power consumption rate and sensitivity level of each sensor

Sensor type	Power consumption	Detecting concentration scope
MQ-3	5 V \pm 0.1	0.05–10 mg/L
MQ-2	5 V \pm 0.1	300–5000 ppm
MQ-9	5 V \pm 0.1	500–10000 ppm

Interface Id (API ID) with 16-digits of characters and numbers. The same API-ID should be used in the code to receive messages. Table 1 shows the power consumption rate and sensitivity level of each sensor.

3.2 Navigation

The second is the navigation section. Arduino UNO microcontroller is used in this section as it is cost effective compared to other MCU boards. GPS module is used to find the location of the person for navigation [4, 6].

The current location is tracked by the GPS module, and it is then compared with the shortest exit gate. LCD screen will display the shortest exit among the available exit gates. The shortest path is calculated using $\text{haversine}(\theta) = \sin^2(\theta/2)$, where θ is the angle between the points. The haversine formula determines the distance between two points given their longitudes and latitudes. If the current location longitudes and latitudes are determined through GPS the between the two exit points can be calculated using the formula and the shortest one is displayed on the LCD screen. In case of an emergency when a person presses the panic button in the wearable band it will provide the shortest exit path. The purpose is that during life-threatening conditions our brain will trigger and release stress hormones. The brain becomes hyper alert, so it is difficult for a person to think wisely.

Navigation also aids new visitors to the industry to navigate to different locations without the help of others.

3.3 RF-ID Scanner

The third section is the ID scanner section. RF-ID tags are used for this purpose. The entry and exit of the employee to the industry is recorded. An RFID reader is powered by an external power supply. Thus, once it's ON, the generator in it generates a signal with the specified frequency but as the signal strength is terribly less (which might result in weakening of the signal if it's transmitted directly) it's to be amplified which is done using an amplifier circuit, so as to propagate the signal to a long-distance and modulate the signal by a modulator [3]. With all these enhancements the signal is

currently able to be transmitted by an antenna that converts the electrical signal into an electromagnetic signal.

4 Results and Discussion

This section presents the hardware module, and the obtained results.

Hazardous Gas Detection Section: Figure 2 shows the functional module (1) in which MQ-2, MQ-3 and MQ-9 sensors along with a buzzer are interfaced with node MCU [10].

When any one of the given sensors gets activated, i.e. at times of gas leakage or if any of the gas concentration is more or above normal, the buzzer gives an alert to the person wearing the band and it sends an alert message will be sent to safety manager of industry [8].

Figure 3 shows the functional module (2) in which the RC-522 reader is interfaced with NODE MCU. RC522 RFID reader creates an 13.56 MHz electromagnetic field that helps to communicate with the RF tag. Figure 4 shows the alert message sent to safety manager of industry. The message is displayed in the home screen of the mobile so that the officials can take necessary actions immediately to prevent any adverse happenings.

RF-ID Scanner Section: Figure 5 shows the functional module (3) in which the RF tag is placed above the reader. When the RF tag gets activated it sends a signal to the reader through antenna and tries to decode the data encoded in the tag.

Figure 6 shows the results of employee attendance details. Here two RF tags are used and random numbers 1804071 and 1,804,066 are assigned. The reader reads it and decodes it. The decoded data from the reader is fed into adafruit.com which displays the data in the dashboard. Upon scanning of RF tag by the employee the date and time of entry/exit to/from the industry will be displayed.

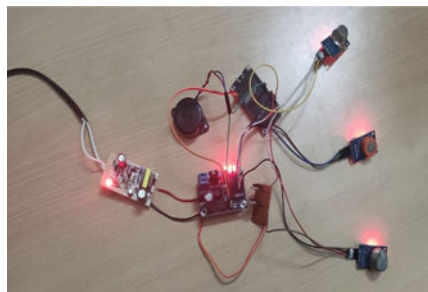


Fig. 2 Functional module (1)

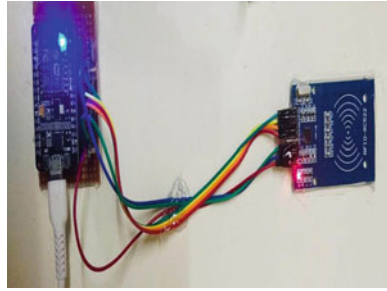


Fig. 3 Functional module (2)

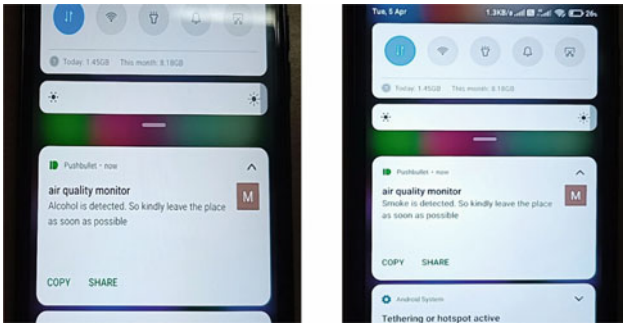


Fig. 4 Alert message

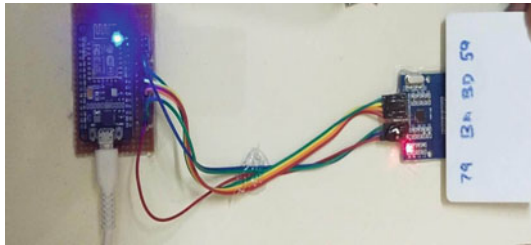
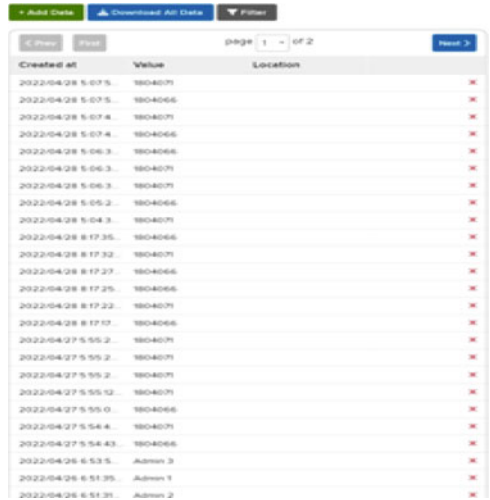


Fig. 5 Functional module (3)

Navigation Section: In this pandemic period permission to visit industry is denied. So, our college campus is taken as a model layout as an alternative to Industry. Figure 7 shows the satellite view of our college campus. College campus has several individual blocks and two exit gates. The marked points A and B represent main gate and polytechnic gate respectively. X is the location of the person wearing the wearable band. It is measured by comparing the latitudes and longitudes of the exit gates and current location of the user using GPS module.



Created At	Value	Location
2022-04-28 5:07:5...	1804071	
2022-04-28 5:07:5...	1804066	
2022-04-28 5:07:4...	1804071	
2022-04-28 5:07:4...	1804066	
2022-04-28 5:06:3...	1804066	
2022-04-28 5:06:3...	1804071	
2022-04-28 5:06:3...	1804071	
2022-04-28 5:05:2...	1804066	
2022-04-28 5:04:3...	1804071	
2022-04-28 8:17:35...	1804066	
2022-04-28 8:17:32...	1804071	
2022-04-28 8:17:27...	1804066	
2022-04-28 8:17:25...	1804066	
2022-04-28 8:17:23...	1804071	
2022-04-28 8:17:17...	1804066	
2022-04-27 5:55:2...	1804071	
2022-04-27 5:55:2...	1804071	
2022-04-27 5:55:12...	1804071	
2022-04-27 5:55:0...	1804066	
2022-04-27 5:54:4...	1804071	
2022-04-27 5:54:43...	1804066	
2022-04-26 6:53:5...	Admin 3	
2022-04-26 6:51:35...	Admin 1	
2022-04-26 6:51:31...	Admin 2	

Fig. 6 Employee registration details



Fig. 7 Satellite view of campus

The distance between A and X is 190 m. The distance between B and X is 400 m when the user is at X. So the nearest exit gate will be A, the Main gate.

Figure 8 shows the Arduino UNO interfaced with GPS and LCD screen. GPS tracks the live latitude and longitude coordinates of the user and compares it with the exit gate coordinates to find the nearest exit gate. The calculated value and the exit gate information are displayed in the LCD as shown in Fig. 9. For the case discussed before the nearest exit distance is calculated as 187.71 m.



Fig. 8 Functional module (4)



Fig. 9 LCD screen showing nearest exit

5 Conclusion and Future Scope

Detection of smoke, methane, and alcohol is done using MQ-2, MQ-3 and MQ-9 sensors. Since this model uses digital output, accuracy is more. Attendance registration is done using RC522 reader and RF tag. The nearest exit of the industry at times of emergency was successfully determined. Navigating people to different places inside the industry based on his/her requirements is done. Had difficulty in integrating the sensors with the Arduino as it gets short circuited often. Also GPS tracking was difficult as the location is with very tall buildings and more obstacles.

In future, the system model with the sensors related to hazardous gas detection, micro controllers, RF tags and readers has to be integrated as a wearable band for use by employee and visitors to the real industry to enhance their safety with proper BYOD policy, so it doesn't impact anything negatively.

References

1. Hashimov EG, Bayramov AA, Sabziev EN (2017) Determination of the bearing angle of unobserved ground targets by use of seismic location cells. *IEEE Int J Military Technol* 5:185–188

2. Dong L et al (2019) The gas leak detection based on a wireless monitoring system. *IEEE Trans Ind Inf* 15(12):6240–6251
3. Hashim NMZ, Ibrahim NA, Saad NM, Sakaguchi F, Zakaria Z (2017) Barcode recognition system. *J Recognit Syst* 2(4):2278–2285
4. Liang LC, Samsudin NA, Mustapha A, Abd Wahab MH (2018) Getting places on time: smart map orientation guide. In: *Fourth International Conference on Information Retrieval and Knowledge Management*, vol 10, no 5, pp 1–5
5. Jelicic V, Magno M, Brunelli D, Paci G, Benini L (2013) Con-text-adaptive multimodal wireless sensor network for energy efficient gas monitoring. *IEEE Sens J* 13(1):328–338
6. Brunelli D, Rossi M (2014) Enhancing lifetime of WSN for natural gas leakages detection. *Microelectron J* 45(12):1665–1670
7. Vallathan G, Shiny XA, Loga, K (2020) Detection of hazardous gas using wearable wireless sensor networks for industrial applications. In: *2020 International Conference on System, Computation, Automation and Networking (ICSCAN)*, pp 1–6. IEEE, July 2020
8. Kong X, Tong S, Gao H, Shen G, Wang K, Collotta M, You I, Das SK (2020) Mobile edge cooperation optimization for wearable internet of things: a network representation-based framework. *IEEE Trans Industr Inf* 17(7):5050–5058
9. Hao Y, Wu Y, Jiang J, Xing Z, Rao Y (2021) The method for leakage detection of urban natural gas pipeline based on the improved ITA and ALO. *J Loss Prev Process Ind* 71(4):104506
10. Anusha M, Nagesh V, Venkata Sai B, Srikanth K, Nanda R (2020) IoT based LPG leakage detection and booking system with customer SMS alerts. *Int J Mod Trends Sci Technol* 6(5):1–5

A Deep Learning Framework for Social Distance Monitoring and Face Mask Detection



Meghana Pamarthi, Sri Latha Injam, Osman Khan Zeeshan Md.,
and T. Lakshmi Surekha

Abstract The Covid outbreak has caused a worldwide calamity with its poisonous spreading. It has become very important to protect ourselves and the people around us from this infection. The dangers of contagiousness can be limited only by following Covid rules such as wearing facemask and keeping up social distance. This paper proposes a system to distinguish whether the person is wearing a facemask or not and also if the people are maintaining a social distance. The framework used is MobileNetV2 for object recognition. The model is prepared on an image dataset and tested with live real time video with a decent precision. The precision is represented by red and green bounding boxes which indicates facemask accuracy as well as the depth for social distance. Red bounding box appears when the particular object is not wearing a mask or not following social distance and green bounding box displays if the object is following the criteria.

Keywords COVID-19 · Social distance · Face mask detection · Deep learning · Mobile NetV2 · Bounding box

1 Introduction

Considering the present situation, social distancing has been a powerful friendly measure as far as normalizing the virus spread. It keeps away from any immediate contact of people and helps in lessening the transmission of the droplets containing the infection by means of the human breath. Johns Hopkins University [1] studies show the outrageous calamity destruction that is affecting human health all around the world. R. Ellis research depicts the pandemic outbursts in Sweden and the health issues of different age group people in that country [2]. What type of norms are

M. Pamarthi · S. L. Injam (✉) · O. K. Z. Md. · T. Lakshmi Surekha
Department of Information Technology, Velagapudi Ramakrishna Siddhartha Engineering
College, Vijayawada, India
e-mail: injamsrilatha01@gmail.com

T. Lakshmi Surekha
e-mail: lakshmisurekha@vrsiddhartha.ac.in

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
P. P. Joby et al. (eds.), *IoT Based Control Networks and Intelligent Systems*,
Lecture Notes in Networks and Systems 528,
https://doi.org/10.1007/978-981-19-5845-8_43

613

needed to be taken and how important it is to follow the rules of facemask and social distance have also been mentioned.

Factors like organic, clinical and designing should be considered to address every one of the inquiries regarding the spreading of infection. Accordingly, taking into account this large number of elements, this project proposes a Deep Learning model that aims in identifying any sorts of violations. This testing helps in building a model that can be applied in real-time frameworks and subsequently help in staying away from the spread of the infection.

The main study of this project is to implement an artificial intelligence system, which can monitor a person wearing face mask and maintaining social distance in a crowd. During the pandemic, these two are the most important rules to comply with. However, some person may not follow the regulations. Monitoring manually is the tough job, and thus this system has been introduced to monitor automatically. The trained model is tested, and the accuracy obtained is 99.72%.

2 Literature Survey

M. S. Islam, M. R. Bhuiyan and S. A. Khushbu [3] introduced a concept using a video illustration with YOLOv3 which is handled by a unique topic in which people benefit from a naturally occurring sickness. Adding to that, the YOLOv3's face mask detection functioned admirably, and it measured the real-time performance with a strong GPU. Abdellatif Mtibaa et al. [4] detected the social distance and face mask using pre-trained models like MobileNet, ResNet Classifier, and VGG. W. Jian and L. Lang [5] presented their detection model such as PP-YOLO that focused on different methods such as mobilenetV2 i.e., used to enhance the available PP-YOLO model and resulted in an accuracy of 86.69% and at a speed of 11.842 ms.

R. N. S and M. N [6] proposed a model using Adam Optimizer algorithm. It was created with Keras, TensorFlow, and OpenCV and was ideally suited for deep learning models. The suggested model achieved 99% accuracy for various training to testing and calculated various accuracy metrics. Pun N.S. et al. [7] approached using deepsort technique which is a method for tracking recognized people using bounding boxes and issued IDs. Their solutions by using YOLOv3 model were then differentiated with other top pre-trained models to detect social distancing in real-time obtained through bounding boxes around the object after the classification. L. Liu et al. [8] represented their work using deep learning techniques to classify generic object detection. The survey included more than 300 research submissions that span a wide range of general object detection topics. Few of the techniques have provided major improvements in object detection. The basic deep learning method used was CNN. Yash Chaudhary D.G., and Mehta M. [9] presented a framework that used bounding box information to identify people.

Qayum F, et al. [10] developed a number of person tracking methods, all of which have produced good tracking solutions which is for normal frontal view photos and video sequences. Depending on the camera distance, the shape and size of the figure

changes. The visibility of a person was also altered by other people or objects during tracking. U. Kumar, et al. [11] proposed the detection of face mask as one of the processes being monitored. The model was created with the help of a convolutional neural network/mobilenet. They included additional features and trained the model on multiple variants, ensuring that the dataset was vast, diverse, and enhanced such that the model could clearly recognise and detect face masks in real-time recordings. The trained model was evaluated based on other models using different scenarios.

Balasubramaniam, and Vivekanadam [12] introduced their ideology, Facemask Detection Algorithm on COVID Community Spread Control using EfficientNet Algorithm. The face mask detection was performed using methods namely R-CNN and YOLO V3. Here the average precision of R-CNN was 62% and the average precision of YOLO V3 was 55%. Dhaya R [13] introduced a technique “Efficient Two Stage Identification for Face mask detection using Multiclass Deep Learning Approach”. It was an efficient model with best accuracy for face mask detection.

3 Proposed System

The proposed system detects the face masks and social distancing with the usage of MobileNetV2 and DNN algorithms. The face mask detection is represented in boundary box using red and green colors and an alert message with buzzer. The aim of this work is to alert the people around with a buzzer, when a person violates either face mask or social distancing, as detected by the digital web camera. The violation count is also displayed so that those who are nearby can take the respective measures to maintain social distancing and wear a facemask.

Red color boundary box means the person standing in front of the webcam doesn't have a face mask and green color boundary box represents the detected person has the face mask, and this proposed system also calculates the social distance of people standing in front of webcam, whether they are maintaining social distance or not by using red and green colors boundary boxes. The red color boundary box represents the person is not maintaining a safe distance and green color boundary box means the person is maintaining a safe distancing from others, and the model calculates the distance between the individuals and is displayed on the upper side of the boundary box.

For detecting face mask and social distance, a dataset which consists 853 pictures has been taken from the MakeML, and also to expand the dataset, different images which consists of 1000 pictures are collected and added to the dataset (with mask:690, without mask:686).

The preprocessing methods such image resizing, image cropping and converting the gray scale to color pictures if necessary have been applied. Then, the face mask detection using MobileNetV2, and the social distance monitoring using DNN model have been performed.

Convolution layer is used for filters, and techniques are carried out to the original image. ReLU which is an activation feature, is used for the working of the convolution

layer used in this method and these may be carried out for a variety of filters. The 1×1 dimensioned kernel is used for convolution layer. There after the feature maps are generated through the global pooling layer. The most advantageous feature about global pooling layer is that it can enforce similarities between categories and feature maps.

While detecting the face of the person, the model can calculate the social distance and detect the presence of mask and represent it in the boundary boxes.

The general cycle in the proposed framework is as follows. First, the dataset is loaded, and then it is pre-processed, and the mask is distinguished. Later, individual identification is done and distance among individual for keeping social separation is calculated as shown in Fig. 1.

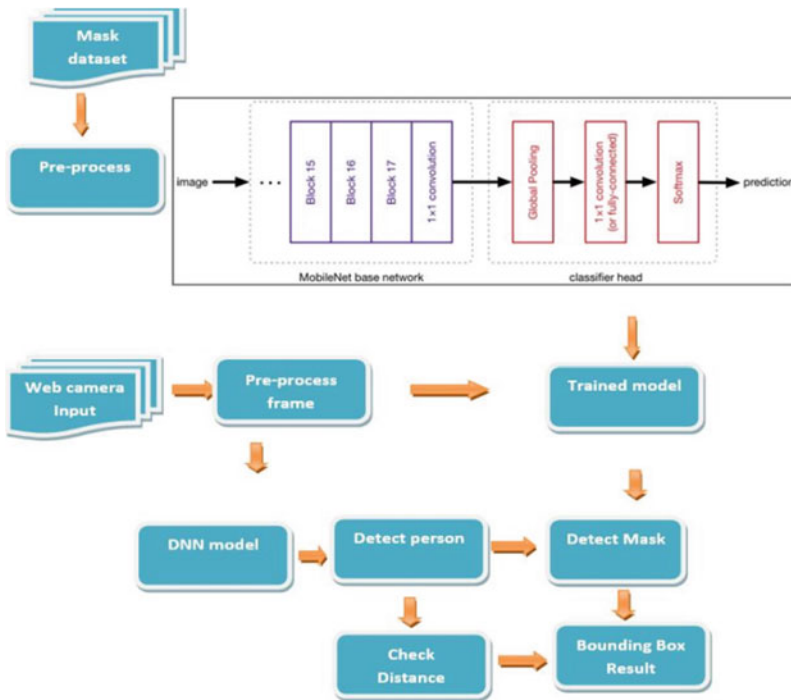


Fig. 1. Architecture diagram

4 Methodology

4.1 Data Collection

The aim of the project is to detect the social distance and face mask by using MobileNetV2 and DNN models. For this, a dataset which consists of the 853 pictures from the MakeML. Another dataset which consists of the 1000 different pictures such as with mask and without mask has also been considered. Among all the collected images, 973 images are set for training and 880 images are used for testing.

4.2 Data Preprocessing

There are four steps of preprocessing that can be evaluated. The first step is image resize, where all the images are converted into equal dimensions such as 224×224 . The second step is image cropping that is done if necessary and all images can be converted into images of equal size. The third step is converting the gray scale images to color if necessary and then the final step is mobileNetV2, applied to the dataset and to the images added further to the dataset, to build the model and to increase the accuracy of the model.

4.3 Face Detection

Input for face detection is taken through web camera which is live video streaming. Human face detection is done using Deep neural network in OpenCV. The input frame is detected for face and if face is identified with a confidence value >0.4 , then the person is detected.

4.4 Facemask Detection

After detecting the faces of human, the pre-trained model MobileNetV2 is applied and checked whether the human face has put on a mask or not. If the mask exists on the face, then the bounding box appears in green color and in case there is no mask it displays red bounding box with accuracy represented on top of the bounding box.

4.5 Social Distance Calculation

After taking the input through camera, the frame is pre-processed. People from the pre-processed frame are identified by applying deep neural network algorithm. Thereafter, if the distance between the tracked persons is more than 2meters green bounding box appears or else a red bounding box is displayed.

$$d(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2} \quad (1)$$

5 Results

The dataset downloaded from MakeML and images added to the dataset, consist of 1000 different pictures such as with mask and without mask as shown in the Fig. 2. Pre-processing methods such as image resize, image cropping and converting gray scale to colour picture if necessary are applied. For this, a model is built to detect the face mask and the social distance, wherein an accuracy of 99.72% is achieved.

In Fig. 3, clearly wearing facemask and/or social distancing have been violated in the three frames of the figure and are detected by the proposed system. Facemasks may be worn whereas social distance may be breached or vice versa. On the other hand, it also represents the accuracy of no mask and the depth of the social distance. Here, the system calculates the distance between individuals and displays it as depth.

The model's accuracy and loss when trained with the Adam optimizer with ReLU as the activation function have been determined and illustrated in Fig. 4.

The experiment is performed by training the model using Adam optimizer with an activation function ReLU. The model is trained with 100 epochs to get more accuracy in detecting facemask and social distance. Table 1 shows the model accuracy and model loss by using Adam optimizer with ReLU activation function.

Fig. 2. Image dataset



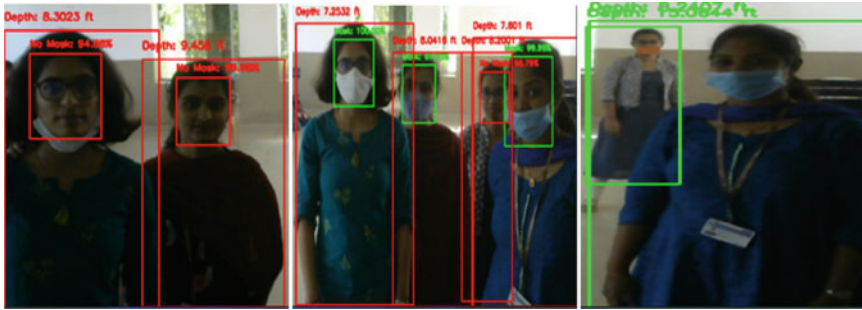


Fig. 3. Various frames showing social distance and face mask detection

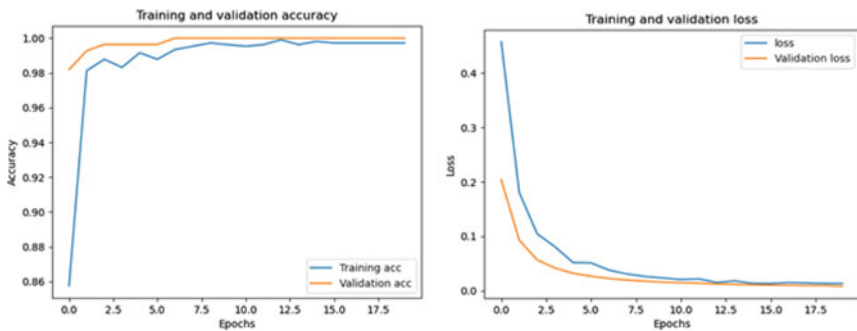


Fig. 4. Model accuracy and model loss graphs with Adam optimizer

Table 1. Model accuracy and loss

Optimizer	Accuracy		Loss	
	Training	Validation	Training	Validation
Adam	99.72	95.98	0.80	1.57

6 Conclusion

In general, people in public places like roads, parks, offices, hospitals etc. do not follow the covid rules such as wearing facemask and maintaining social distance, and as a result of this, the spreading of virus increases rapidly. In such case, the proposed method helps in identifying those violators and categorises them as individuals with or without facemask and with or without social distancing, so that people behind the camera get a chance to warn the people who breach the covid rules. This method is more useful to monitor at crowded areas through CCTVs. The algorithm most suitable for facemask detection and social distance identification is MobileNetV2 with an ideal analysis. This model can also be employed at workplaces for the safety of workers.

References

1. Johns Hopkins University. COVID-19 Map -Johns Hopkins Coronavirus Resource Center, Johns Hopkins Coronavirus Resource Center (2020). <https://coronavirus.jhu.edu/map.html>. Accessed 30 July 2020
2. Ellis R (2020) WHO Changes Stance, Says Public Should Wear Masks. <https://www.webmd.com/lung/news/20200608/who-changes-stance-says-public-should-wear-masks>. Accessed 31 July 2020
3. Bhuiyan MR, Khushbu SA, Islam MS (2020) A deep learning based assistive system to classify COVID-19 face mask for human safety with YOLOv3. In: 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp 1–5. <https://doi.org/10.1109/ICCCNT49239.2020.9225384>.
4. Teboulbi S, Messaoud S, Hajjaji MA, Mtibaa A (2021) Real-time implementation of AI-based face mask detection and social distancing measuring system for COVID-19 prevention. *Sci Program* **2021**:8340779. <https://doi.org/10.1155/2021/8340779>
5. Jian W, Lang L (2021) Face mask detection based on transfer learning and PP-YOLO. In: 2021 IEEE 2nd international conference on big data, artificial intelligence and internet of things engineering (ICBAIE), pp 106–109. <https://doi.org/10.1109/ICBAIE52039.2021.9389953>
6. S RN, N M (2021) Computer-vision based face mask detection using CNN. In: 2021 6th international conference on communication and electronics systems (ICCES). pp 1780–1786. <https://doi.org/10.1109/ICCES51350.2021.9489098>.
7. Punn NS, Sonbhadra SK, Agarwal S (2020) Monitoring COVID-19 social distancing with person detection and tracking via fine-tuned YOLO v3 and Deepsort techniques. *arXiv:2005.01385*
8. Liu L et al (2020) Deep learning for generic object detection: a survey. *Int J Comput Vis* **128**(2):261–318. <https://doi.org/10.1007/s11263-019-01247-4>
9. Yash Chaudhary DG, Mehta M (2020) 22nd international conference on E-health networking, applications and services (IEEE Healthcom 2020), Shenzhen, China, 12–15 December 2020
10. Ahmad M, Ahmed I, Khan FA, Qayum F, Aljuaid H (2020) Convolutional neural network-based person tracking using overhead views. *Int J Distrib Sens Netw* (2020). <https://doi.org/10.1177/1550147720934738>
11. Sakshi S, Gupta AK, Singh Yadav S, Kumar U (2021) 2021 International conference on advance computing and innovative technologies in engineering (ICACITE), pp 212–216. <https://doi.org/10.1109/ICACITE51222.2021.9404731>.
12. Balasubramaniam V (2021) Facemask detection algorithm on COVID community spread control using EfficientNet algorithm. *J Soft Comput Paradigm* **3**(2):110–122
13. Dhaya R (2021) Efficient two stage identification for face mask detection using multiclass deep learning approach. *J Ubiquit Comput Commun Technol* **3**(2):107–121

Information Security and Privacy in Smart Cities, Smart Agriculture, Industry 4.0, Smart Medicine, and Smart Healthcare



Sanjana Prasad, Arun Samimalai, S. Rashmi Rani, B. P. Pradeep Kumar,
Nayana Hegde, and Sufia Banu

Abstract Internet of Things (IoT) is used to interconnect various things, devices, technologies in a network in order to perform various tasks at a higher speed, less loss of information as well as for device-cloud/device-device communication. IoT works by transmission of data collected by the large number of sensors, devices which communicate to the cloud through the internet connectivity. Software processing is done after the data gets to the cloud. Some of the IoT devices used are smart mobiles, smart door locks, medical sensors, smart refrigerators, smartwatches, smart bicycles, smart fire alarms, fitness trackers and smart security systems. When a number of smart gadgets are connected with the Internet, there are various security threats, attacks and concerns when a malicious user enters the network. There is a possibility of the data transmitted/received being modified, lost or misused. In order to prevent those threats, there are various solutions introduced by various researchers for securing the networks. In this article, we analyse the importance, characteristics, security issues and privacy concerns involved in various applications such as Smart Agriculture, Smart Cities, Smart Healthcare and Smart Medicine. Frameworks proposed to reduce the impact of security and privacy issues are also discussed along with future research directions and scope.

Keywords Internet of Things · Security · Sensors · Smart city · Smart medicine · Smart healthcare · Smart agriculture

S. Prasad (✉) · S. R. Rani · B. P. P. Kumar · S. Banu

Department of Electronics and Communication Engineering, HKBK College of Engineering,
Bangalore 560045, India

e-mail: sanjanaprasad@hkbk.edu.in

S. Banu

e-mail: sufiabanu.ec@hkbk.edu.in

A. Samimalai

Department of Electronics and Communication Engineering, CMR Institute of Technology,
Bangalore 560037, India

e-mail: arun.samimalai@gmail.com

N. Hegde

School of Electronics and Communication Engineering, Reva University, Bangalore 560064, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

621

P. P. Joby et al. (eds.), *IoT Based Control Networks and Intelligent Systems*,

Lecture Notes in Networks and Systems 528,

https://doi.org/10.1007/978-981-19-5845-8_44

1 Introduction

IoT was introduced in the year 1999 by Auto-ID Lab, MIT, USA. 27 billion number of IoT devices are present in the current scenario and it might increase in the future. A network of physical devices such as electronic devices, sensors, and various softwares forms IoT. The “Things” on IoT means anything and everything around us that can be connected in order to create a better digital world. Things can include machines, buildings, devices, animals, and human beings., etc.

Some of the underlying technologies which makes IoT includes Radio Frequency Identification Device (RFID), Sensors, Smart Technologies and Nano technologies. Internet of Things consists of three primary types of networks such as ubiquitous, grid, vehicular and distributed networks. Some of the applications where sensors are primarily used includes entry of patient details to post-surgery in the database, caring children and elderly people, and smart cards. Though it has many advantages, it can also be affected with various security threats and attacks. Some of the phases in IoT system includes Data collection, Storage, Process, Data transmission and delivery. Different stages and layers in IoT applications can be affected with different kinds of threats and attacks.

Data in the IoT networks can be considered robust if they have the capacity to cope up with any kind of error sources, abnormalities from the input side and calculations during the period of execution. This is considered as one of the main concerns while transferring information in IoT networks. Scalability is considered as the capability of IoT to handle a large number of multiple users, app features and analytics capabilities without degradation in QoS. By getting an idea about the structure of community, adding few edges in-between the communities, using techniques employing scale-free networks, and by selecting appropriate algorithms, scalability can be achieved. By using appropriate IoT softwares, architectures like Context Broker, Kubernetes and databases, vertical scalability can be achieved by making the multicore processors to handle more number of requests at the same time. Another way of increasing scalability is by maximizing the number of cores and memory. Other solutions including building a interference immunity zone, by choosing star network topology design for IoT networks, leveraging of open interfaces, by employing a powerful network and device management tool. By employing these kind of features in IoT networks, we can use, deploy, operate and expand softwares to adapt to the required needs.

There are various ways in which security can be achieved. To state a few, Information in IoT networks can be secured by using anti-virus and anti-malware softwares like Firewalls, Intrusion Prevention and Detection Systems, using Embedded Tools and by facilitating Remediation. For each application, security schemes will vary. Based on the application and the concerns, information security scheme can be chosen.

Some of the phases in IoT networks includes – Data collection, data storage, data processing, data transmission and delivery end-to-end. The issues faced by each phase are described below.

- In the *data collection* phase, the issues faced are Data Sovereignty, Data leakage, Data authentication and Data breach.
- During the *data Storage* phase, some of the concerns to be considered includes Denial of Service (DoS), Access control, Attack on Availability, Integrity, Modification of sensitive data and Impersonation.
- During the *data processing* phase, Authentication attacks are prone by the IoT networks.
- During the *data transmission* phase, Routing protocols, Hijacking of session, flooding, and channel Security are the few things to be considered.
- In the *delivery end-to-end phase*, Man-Machine or Maker-Hacker interaction has to be taken into consideration.

IoT networks has various layers which will be involved during transmission such as Sensing (Perception) layer, Network layer, Transport and Application layers respectively. Each layer in the architecture are prone to various attacks. Table 1 illustrates the issues in various layers - Sensing, Network, Transport and Application layer of the IoT architecture.

Security and privacy issues preserving of locating the IoT sensor data from the cloud were described by Mugunthan in [3].

The organization of the paper is as follows. Section 2 describes the objective and the motivation behind the proposed work. Section 3 illustrates briefly about the information security and privacy in smart cities and its classification, architecture, attacks related to smart cities, factors affecting smart city applications and framework used to address the issues related to security and privacy. In Sect. 4, a brief introduction about Smart Agriculture (SA) along with the role of each layer in SA, security and privacy issues involved in it are discussed. In Sect. 5, a brief introduction about Industry 4.0 along with security and privacy issues related with it are also put forth. Section 6 discusses about smart medicine, architecture of IoMT-based healthcare systems, along with security and privacy issues are discussed. Section 7 briefs about smart healthcare, security and privacy issues associated with it are discussed, along with security and privacy issues. In Sect. 8, scope of the future research work along with concluding remarks are discussed.

Table 1 Issues in various layers of the architecture [2]

Layer	Attacks
Sensing/perception layer	Worm hole, Selective forwarding, Witch attack, External attack, Access control, HELLO Flooding, Link layer, Broadcast Authentication and Flooding,
Network layer	Routing Protocol, Address compromise
Transport layer	Man-in-the middle, Cross heterogenous and Masquerade, Distributed denial of service, Denial of Service
Application layer	Revealing sensitive data, User authentication, Intellectual property and Data destruction

2 Objective of the Proposed Work

IoT is one of the key advancements in the recent times. Almost all the devices used in our daily lives involves IoT. There are so many advantages in using IoT networks such as speed of data transmission, minimization of efforts required by humans, good utilization of existing resources, and data quality. Though it has so many advantages, there are few drawbacks as well in IoT networks due to the interconnection of devices to the network. Due to this, any malicious user can enter the network and hack the data, do modifications and affect the existing data. The motivation of this particular work is to address all the existing security and privacy concerns involved in various domains such as Smart Agriculture, Smart cities, Industrial IoT, Smart Healthcare and Smart Medicine and also discuss about the suitable ways to address the existing issues. Future research direction in various topics are also discussed.

3 Information Security and Privacy in Smart Cities

Smart cities are government initiatives consisting of e-services brought in order to integrate a number of systems and to handle urbanization. This, in turn, will also drive economic growth and improvise the style of living of an individual [4]. Smart cities are also named as intelligent cities comprising various factors. The 3 main features of smart city include instrumented, interconnected and intelligent.

These include increasing economy of a state/city/country, financial status of people residing in the region, governance, enhance the mobility, improve environmental conditions and by living a smarter life [1]. Some of the smart cities across the world are New York, Singapore, Amsterdam, Dubai, London, Hong Kong, Copenhagen, and Barcelona.

Ismagilova et al. [5] did a study based on security and privacy attacks along with its outcomes. The role of security attacks impacting engagement of individuals and data privacy and the challenges were also included.

Turjman et al. [6] described about the role of each layer in smart cities such as Physical, Network, Database, Virtualization, Data analytics/Mining and Application. Applications of smart cities includes the following mentioned in the Fig. 1.

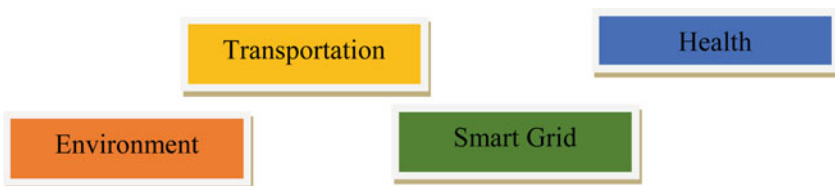


Fig. 1 Applications of smart cities

Some of the issues faced by smart cities includes the following—Cyber security, activities related to botnet and autonomous vehicles and leakage of private data.

Smart cities can be classified into 3 types and they are Digital city, ICT city and Compound city. Digital city is connected to the Internet and is considered as the technological platform for information and managing communication thereby enabling Internet of Things. Information and Communication Technologies Infrastructure is considered as lifeline of a smart city, which enables participatory governance and management, and interpretation of large data. This can be used to improve the interactivity of urban services, performance, quality and also to reduce costs, and improve communication gap between citizens, and resource consumption. Some of the characteristics of ICT city includes the following factors such as [7]

- Robustness and scalability
- IoT enabled networks
- Automatic security threat isolation and remediation
- Automated and simplified network management

The unique architecture of smart city has four layers such as Sensing layer, Data collection layer, Data processing, and smart processing and application layer. Sensing layer is the first layer consisting of sensors, actuators and cameras which is used to sense, collect and delivers data to data collection layer. Transmission of information from sensing layer to local databases is done in the data collection layer. After this, the local database information is transferred to data processing layer in order to perform pre-processing techniques. Application layer takes care of data exchange between operators and smart applications. Pillars of smart cities include institutional infrastructure, physical infrastructure, social infrastructure, and economic infrastructure. Figure 2 shows the privacy and security issues in smart cities.



Fig. 2 Attacks related to Cyber-Security faced by Smart cities [6]

Smart cities comprises of Smart government, Smart Health, Smart Grid, Smart Security, Smart Buildings, Smart Farming, Open Data, Smart Home, Smart Transportation, Digital Citizens, and Wi-Fi. The future of smart cities is driven by E-government. Smart city uses technology, urban growth, human capital, education, social capital, and environmental issues. The main challenges of smart cities involve the following factors such as Physical connectivity, Wi-Fi security, Hardware Security, Bandwidth Consumption, and Application risks in Smart vehicles. Security threats in smart cities and E-governance include Identity theft, DoS, Malware, IoT node security, Attacks on IoT devices, RFI conflict collision, Side channel attack, and Man-in the middle attack. Smys et al. [8] simulated 5G networks in smart cities using algorithm based on neural networks.

In order to preserve and protect the individual's information/privacy e-government infrastructure need to be safeguarded by carrying out respective methods. Some of the security and privacy solutions for securing and privacy in smart environments include Blockchain, Cryptography, Biometrics, Machine learning and data mining, game theory, ontology and non-technical supplements. Security and Privacy requirements for smart cities services are privacy by design, testing and verification, privacy architecture, data minimization, secret sharing, system security and access control, secure multiparty computation [9].

Future research should be chosen such that it is used to solve trust challenges within smart cities which will benefit initiative of smart city as shown in Fig. 3. [5] Characteristics of smart city includes Heterogeneity, Resource constraint, user involvement, connectivity/scalability and mobility. There are different components comprising a secure architecture.

Black networks for preserving privacy of data, SDN controller such as TTP for efficient routing and availability, Unified registry for mobility, validation, and acts as database for various nodes, gateway, and Key management system. Some of the solutions to security and privacy can be solved by using mobile crowdsensing, big data, Network security based on IoT, less-weight security solution, authentication, integrity, availability, and confidentiality and by involving cloud/fog computing.

Other techniques which can be used to improvise security and privacy in smart cities are Machine Learning, Game Theory, Cryptography, Biometrics, and blockchain [10]

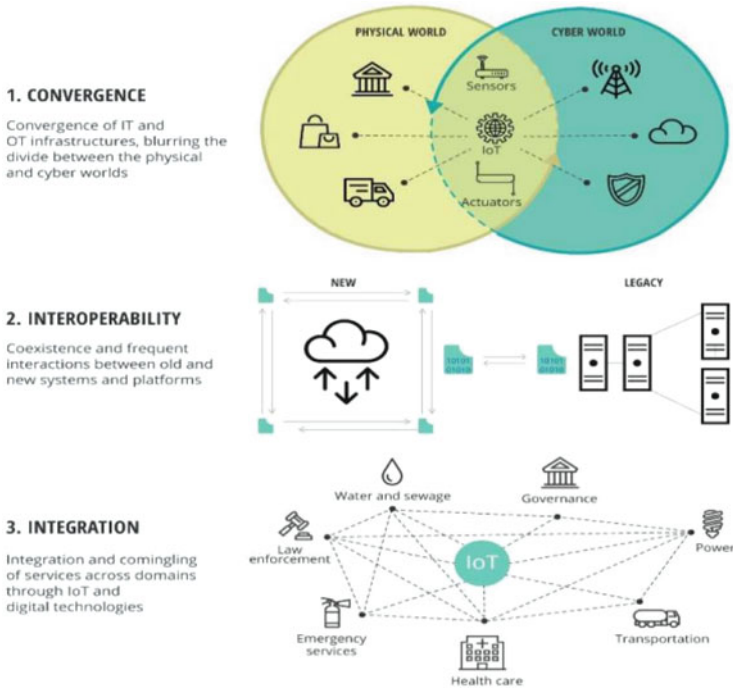


Fig. 3 Factors affecting smart cities applications [10]

4 Information Security and Privacy in Smart Agriculture

Agriculture is the primary provider of food as well as plays a major role in growth of an economy. Food production needs to be increased by 70% by 2050 to meet consumer needs. In order to improve the current agricultural processes, smart agriculture integrating a bunch of processes, components, rules and computation paradigms is integrated. To preserve, and to analyze the huge amount of information created by the components, different solutions like edge computing, cloud, artificial intelligence and big data are employed. Figure 5 shows the components involved in Smart agriculture.

Different issues exist in different layers. Issues existing in different layers are described in Table 2.

Security challenges in smart agriculture include Harsh Environment, threats from agricultural equipment. The authors [12] outlined a multi-layered architecture and described issues in cyber-environment along with scenarios, challenges and directions related to cyber attacks.

Yazdinejad et al. [14] categorizes security threats within smart farming/precision agriculture areas, taxonomy of cyber security threats, survey on risk mitigation strategies.

Table 2 Issues existing in different layers [2]

Layer	Attacks
Perception	Sleep deprivation, fake node, node capture, sensor weakening, irregularity, Hijacking, optical deformation, autonomous system disruption, Autonomous system and Random sensor incidents
Network	DoS/DDoS, Data transit attacks, Routing attacks, Signal disruptions
Edge	Flooding, signature wrapping, man-in-the middle, unauthorized access, Booting, request forgery, Gateway-cloud and Forged control for actuators
Application	DoS/DDoS, Malicious scripts and Phishing

Zanella et al. [11] described the overview of smart agriculture along with security threats in perception, network, edge and application layer. Current scenario in smart agriculture along with resources required to improve security were also briefly described.

5 Information Security and Privacy in Industry 4.0

Industry 4.0 is the major digital transformation that the manufacturing and production industry is going through in today's scenario. Various challenges related to security and privacy will pave a way during the implementation of Industry 4.0. There are 3 types of categories in which the Industry 4.0 revolution will come under and they are people, processes, and technologies.

Jhanjhi et al. [15] discussed about the various attacks in different layers of Industry IoT architecture along with countermeasures. A framework based on Industry IoT security was proposed by Sadeghi et al. Some of the solutions to address security and privacy risks includes the following—Industrial rights management, Secure Engineering, Security and Identification management and Platform security. Some of the major factors characterizing Industry 4.0 are automation, decrease human interaction, enhancing closed loop data models. Challenges in Industry 4.0 include. Lack of IT/OT security expertise and awareness due to the humans involved in manufacturing processes.

- Lack of policies and funds to focus on security
- Liability over products
- Lack of uniform standardization
- Technical constraints of the devices

Figure 4 shows the block diagram of Industrial Internet of Things.

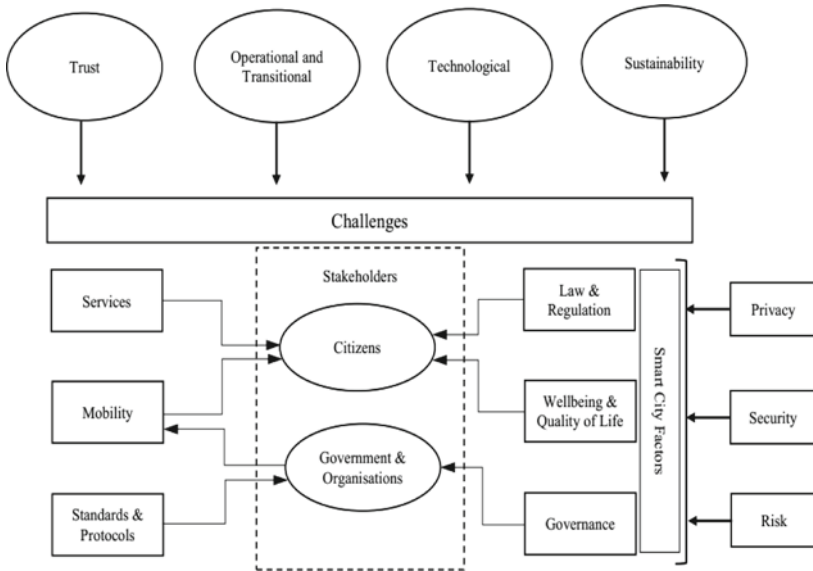


Fig. 4 Framework related to smart cities security and privacy [5]

6 Information Security and Privacy in Smart Medicine

Smart medicine/Smart drugs refer to a group of pharmaceutical agents which is consumed in order to improve the intellectual capacity of people suffering from neurological diseases and psychological disorders. Smart drugs are used to improve the cognitive function Smart drugs Smart medicine consists of a silicon sensor that is attached to the pills which will be taken by the patient. When the patient consumes the pill along with the attached sensor, it gets activated when it touches the gastric juice in the stomach. After that, the medicine’s identity and the period of consumption of the drug is determined by the sensor. Later, a throwaway patch in contact with the skin measures the different parameters of the body. This measured data is then transferred to any device like computer or phone via Bluetooth which can be accessed by clinicians and caregivers. The system allows users to set up alarms to remind them to take the medicines or go off when they are inactive for a certain time.

7 Information Security and Privacy in Smart Healthcare

Patient care is one of the most primary things in healthcare industry as in Figs. 6 and 7. By employing IoT in healthcare, we can improvise the lives of patients admitted in hospitals. Security and Privacy plays a major role when manufacturing devices, interconnect things, communication, while handling and storing data. Some of the



Fig. 5 Components of Smart Agriculture [11]

primary functionalities of Healthcare IoT include Health monitoring remotely, wearable devices such as sleep trackers, smart-shirt, smart-bracelet, and smart-watch on the body for monitoring and self-assistance, infusing medicine into the patient in a personalised way, and maintaining medical equipments.

When the data is transmitted over the network and cloud, there are some possibilities that the data can be hacked by external users due to which the individual is susceptible to risk. In order to overcome this drawback, solutions such as tagging of data, zero knowledge proof k-anonymity model, IoT enable cloud solution are proposed [19].

Some of the security schemes/solutions are proxy-based protection, distance bounding, ECG based encryption, analogue shielding, zero power communication, anomaly detection.

Elhoseny et al. [18] described the root causes related to ransomware, insider threat, unsecure database, environmental misconfiguration. The authors discussed about the

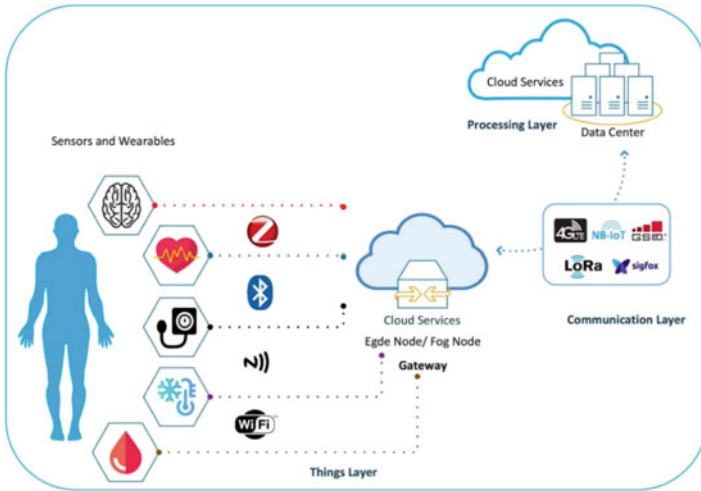


Fig. 6 Three tier architecture of IoT healthcare system [13]

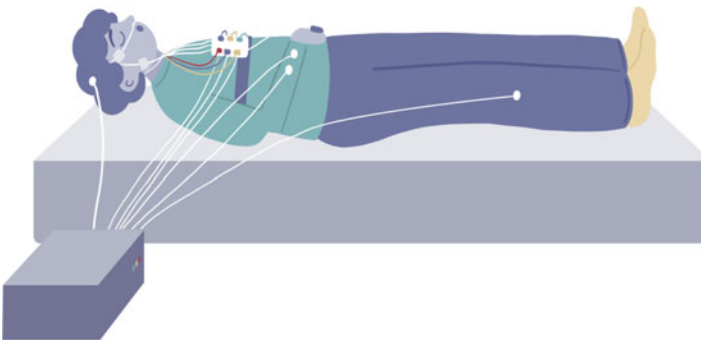


Fig. 7 Example of Sleep tracker

countermeasure concerns related to security and privacy in MIoT, challenges and future directions.

The challenges associated with HIoT include biocompatibility, interoperability, security, reliability, power sources and scalability, vulnerability and regulatory environment. Few requirements for security include maintaining confidentiality, integrity, using authentication, non-repudiation, authorization and freshness.

Data confidentiality

Problems such as Eavesdropping, unauthorized users can be addressed by using ciphers for encrypting data.

Data Integrity

By doing cryptography based Integrity checks, meddling by malicious attackers, errors occurring due to accidental communication during data transmission can be minimized. Some of the algorithms used are AES128/256, MD5, SHA and S-box.

Authentication. Digital Signatures and certificates and Authentication Keys are used to overcome drawbacks such as forgery, Masquerading healthcare and Personal healthcare records as soon in Figs. 8, 9 and 10. Public key infrastructures paired with Digital signatures can be used to avoid Non-Repudiation attacks.

Authorization ensures approval to the correct nodes giving the access rights.

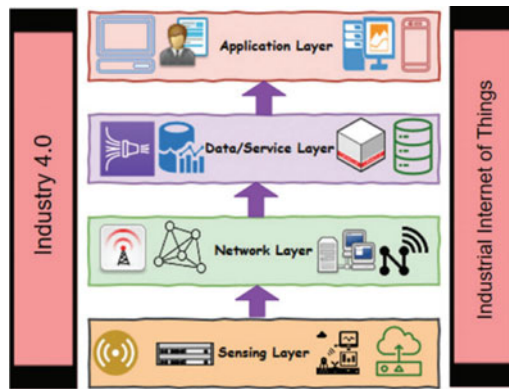


Fig. 8 Architecture of Industry 4.0 [15]

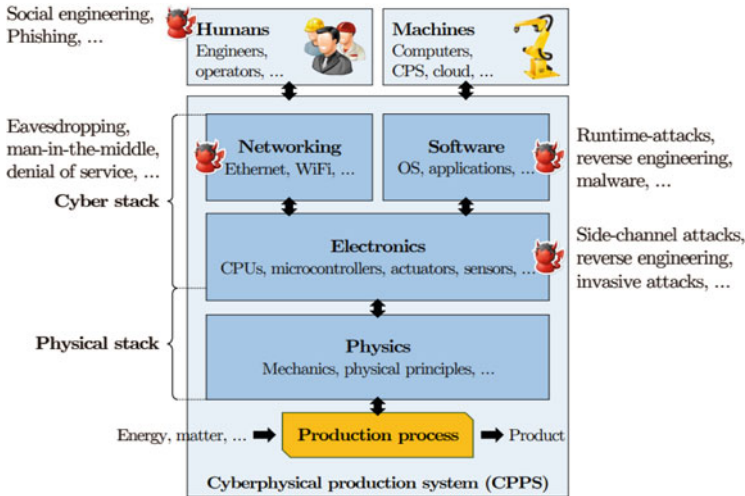


Fig. 9 Architecture of cyber-physical production system (CPPS) and attack surfaces [16]

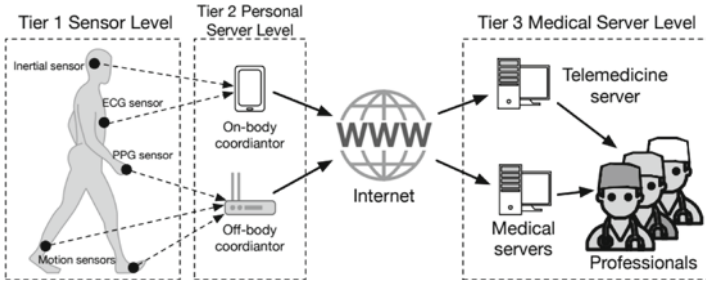


Fig. 10 Architecture of IoMT-based healthcare systems [17]

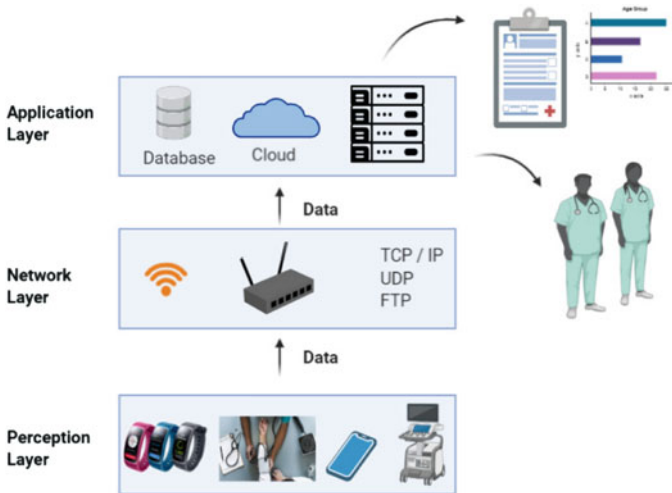


Fig. 11 Three layer architecture of Medical Internet of Things [18]

Solution: Assign the required access control to the correct user.

Freshness

Replay attacks can be reduced by ensuring freshness of received data by doing various processed such as verification of received data from medical sensors and ensuring they are current, ordered and non-duplicated [20].

Remote Patient Monitoring/Telehealth at smart home allows patients to use smart medical diagnostics along with mobile devices to capture vitals, real-time before getting transmitted to medical healthcare cost reduction and also it limits the number of times a patient is visiting a hospital thereby reducing the overall costs. In order to secure smart homes and healthcare as well as to identify threats and for maintaining privacy concerns in various scenarios, a method called STRIDE is employed. S-Spoofing; T- Tampering; R – Repudiation; I – Information Disclosure; D – Denial of Service; E – Elevated Privileges.

Various kinds of Using IoT based mini-CPR patch sensors and IR temperature check, a Smart medicine kit providing which monitors heart rate and blood glucose level was designed by Nalini et al. [21]. This kit also reminds them to take medicines by employing a timer circuit, thereby easing medical diagnosis.

8 Conclusion and Future Scope

In this paper, we discuss briefly about the importance, advantages of IoT in smart applications. Furthermore, the information security and privacy issues in Smart cities, Smart agriculture, Industry 4.0, Smart healthcare and smart medicine are also discussed. Future research directions and scope of the research work in the domain are also discussed. By having an idea about the various security and privacy issues in Smart applications, we can design intrusion detection systems to monitor inbound and outbound traffic on the network and also the information can be sent to the person in-charge in order to ensure the safety and protection of the network.

References

1. Gaire R, Ghosh RK, Kim J, Krumpholz A, Ranjan R, Shyamasundar RK, Nepal S (2019) Crowdsensing and privacy in smart city applications. In: Smart cities cybersecurity and privacy. Elsevier, pp 57–73
2. Jeyanthi N (2016) Internet of things (IoT) as interconnection of threats (IoT). security and privacy in intern. J ISMAC IoT. CRC Press, London, pp 23–42
3. Mugunthan SR (2019) Security and privacy-preserving sensor data localization based on the internet of things. J ISMAC 1(02):81–92
4. Yang L, Elisa N, Eliot N (2019) Privacy and security aspects of e-government in smart cities. In: Smart cities cybersecurity and privacy. Elsevier, pp 89–102
5. Ismagilova E, Hughes L, Rana NP, Dwivedi YK (2022) Security, privacy and risks within smart cities: literature review and development of a smart city interaction framework. Inf Sys Front 24:393–414, 1–22
6. Al-Tudjman F, Zahmatkesh H, Shahroze R (2022) An overview of security and privacy in smart cities' IoT communications. Trans Emerging Telecommun technol 33(3):e3677
7. Sookhak M, Tang H, He Y, Yu FR (2018) Security and privacy of smart cities: a survey, research issues and challenges. IEEE Commun Surv Tutorials 21(2):1718–1743
8. Smys S, Wang H, Basar A (2021) Journal of information security and applications, 5G network simulation in smart cities using neural network algorithm. J Artif Intell 3(01):43–52
9. Deebak BD, Fadi AT (2021) Privacy-preserving in smart contracts using blockchain and artificial intelligence for cyber risk measurements. J Inf Secur Appl 58:102749
10. Varfolomeev AA, Alfarhani LH, Olewi ZC (2021) Overview of five techniques used for security and privacy insurance in smart cities. In: Journal of physics: conference series, vol 1897, no 1. IOP Publishing, pp 012028
11. de Araujo Zanella AR, da Silva E, Albini LCP (2020) Security challenges to smart agriculture: current state, key issues, and future directions. Array 8:100048
12. Gupta M, Abdelsalam M, Khorsandroo S, Mittal S (2020) Security and privacy in smart farming: challenges and opportunities. IEEE Access 8:34564–34584

13. Karunaratne SM, Saxena N, Khan MK (2021) Security and privacy in IoT smart healthcare. *IEEE Internet Comput* 25(4):37–48
14. Yazdinejad A et al (2021) A review on security of smart farming and precision agriculture: security aspects, attacks, threats and countermeasures. *Appl Sci* 11(16):7518
15. Jhanjhi NZ, Humayun M, Almuayqil SN (2021) Cyber security and privacy issues in industrial internet of things. *Comput Syst Sci Eng* 37(3):361–380
16. Sadeghi AR, Wachsmann C, Waidner M (2015) Security and privacy challenges in industrial internet of things. In: *ACM/EDAC/IEEE design automation conference (DAC)*. IEEE, pp 1–6
17. Sun Y, Lo FPW, Lo B (2019) Security and privacy for the internet of medical things enabled healthcare systems: a survey. *IEEE Access* 7:183339–183355
18. Elhoseny M et al (2021) Security and privacy issues in medical Internet of Things: overview, countermeasures, challenges and future directions. *Sustainability* 13(21):11645
19. Sadek I, Rehman SU, Codjo J, Abdulrazak B (2019) Privacy and security of iot based healthcare systems: Concerns, solutions, and recommendations. In: Pagán J, Mokhtari M, Aloulou H, Abdulrazak B, Cabrera MF (eds) *How AI Impacts Urban Living and Public Health*, vol 11862. *Lecture Notes in Computer Science*. Springer, Cham, pp 3–17. https://doi.org/10.1007/978-3-030-32785-9_1
20. Karunaratne SM, Saxena N, Khan MK (2021) Security and privacy in IoT smart healthcare. *IEEE Internet Comput* 25(4):37–48
21. Nalini M, Abirami V, Lakshmi GA, Harini D (2021) IoT based smart medicine kit. *Mater Today: Proc* 46:4125–4127

Virtual Machine and Container Live Migration Algorithms for Energy Optimization of Data Centre in Cloud Environment: A Research Review



Shridevi Soma and S. Rukmini

Abstract The growing pace of cloud computing technology needs to optimize the energy consumption of the data centres. Virtualization is one such technology that helps in live migration of Virtual Machines and Containers and thus help reduction in energy consumed. An effort is made in this paper to explore different approaches for energy optimized live migration using Virtual Machine and Container. Majority of the work has been carried out by researchers considering CPU utilization as the parameter to optimize the energy consumption and some works have used other parameters like memory, disk space, application execution time for active server along with CPU utilization. The survey also depicts majority of the work are implemented using CloudSim and also there is more scope for developing solutions for optimal migration in Virtual Machine and Containers.

Keywords Virtual Machine · Physical machine (PM) · Live migration · Energy consumption · Container migration

1 Introduction

Live migration has recently attracted many approaches to optimize the energy consumption in data centres. An attempt is made in this section to understand the basic approaches to live migration.

Minimizing energy consumption of data centres has become a challenging issue due to increase in number of data centres, cloud service providers and cloud users. From the previous research we can observe that an email service provided by the organization with 500 user cost an average energy consumption of the server of 16 kW

S. Soma

CSE, Poojya Doddappa Appa College of Engineering, Kalaburagi, Karnataka, India
e-mail: shridevisoma@pdaengg.com

S. Rukmini (✉)

CSE, Government Women's Polytechnic, Kalaburagi, Karnataka, India
e-mail: rsatyal9scp@gmail.com

[21]. The energy efficiency efforts in data center were originally focused on the infrastructure changes such as using energy star non-IT equipment, flooring choices like perforated tiles [22], further focuses was on establishment of data centres at the location where electricity and water supply was easy, for example Google built their huge data centre on the banks of Columbia river [24]. Further research has been carried out to optimize energy by considering efficient hardware such as microblade [26]. Though there was development in hardware infrastructure and better selection of data centre location challenges still exists to optimize the energy consumption due to poor software performance [25]. Virtualization and migration of cloud computing technology plays a vital role in energy consumption. Many research focused on software-based solutions where majority of the works considered server/CPU utilization as the performance parameter to reduce the energy consumption of data centre.

1.1 Virtual Machine vs Container

Containers are micro-Virtual Machine, the essential difference between Virtual Machines and Containers is: Virtualization in Virtual Machine is at OS, kernel and hardware level whereas, Containers have OS level virtualization. Containers based on Docker [15] have been gaining attraction in recent years due to their reduced resource usage and less virtualization overhead. Container provides near-native performance to the application because of its execution within the Host-OS environment.

It can be observed in Fig. 1 the difference between Virtual Machine and Containers. Each Virtual Machine has different guest operating system, and a hypervisor for creating, maintaining and deleting the Virtual Machines whereas Containers

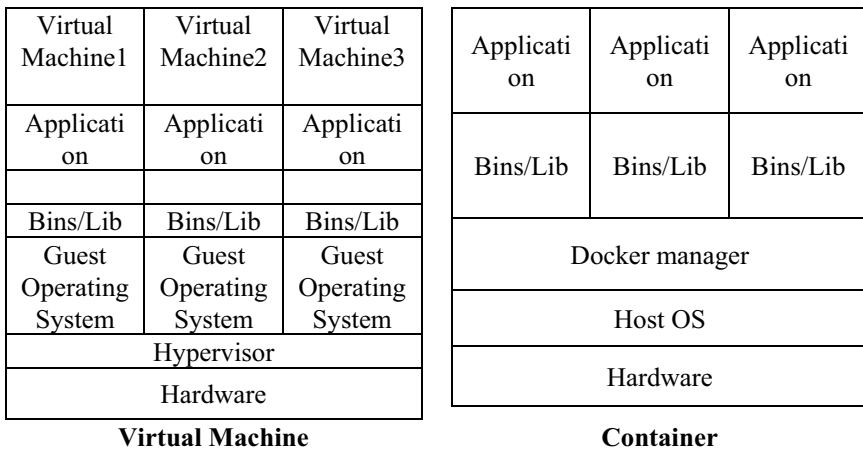


Fig. 1 Virtual machine and containers

are applications or lightweight Virtual Machines having own libraries and environment required for the execution of applications. Virtual Machines are managed by hypervisor. Containers also have Docker manager instead of hypervisor.

1.2 Live Migration

Live migration in Virtual Machine or Container basically involves shutting down the system for few seconds and migrate it from one server to another. This section gives an approach of live migration used in OpenVZ, which is OS level virtualization for Linux. OpenVZ is a open source virtualization implemented in Linux which allows to run multiple isolated Virtual machines and containers or virtual environments.

Live migration process is initiated whenever a physical machine requires maintenance, updating or load balancing. During live migration the information stored in virtual machine's memory is initially transferred to the destination physical host. Next CPU state, memory and storage is established on the destined physical host. In the last step the Virtual Machine is suspended at the source physical machine and then copied and initiated on the destination physical machine along with its application with a minimal downtime.

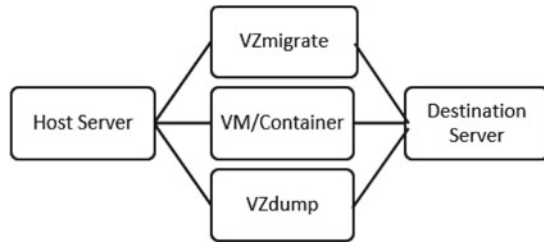
1.3 Live Migration Approaches

The basic approaches to live migration can be classified as.

All the approaches discussed are based [1] on transferring the contents memory of the Virtual Machine or Container to be migrated. The Fig. 2: illustrates a basic approach of live migration in Virtual Machine/Container in OpenVZ.

1. Stop and copy: In this approach the working Virtual Machine/Container is stopped and migrated to the destination server and restarted at the destination. The disadvantage of such approach is that the downtime is more and hence there is service violation to the cloud user.
2. Pure demand migration: This approach is a short stop and copy approach in which essential data structures required will be copied to the destination, and then Virtual Machine/Con is started at the destination. The rest of the memory is transferred later. Though short downtime, disadvantage the migration time is very longer.
3. Pre-copy: It is a iterative method, which consists of pre-copy rounds where very small pause and copy phases are there. These rounds are the ones which have pages to be transferred during round 'm' are modified in round 'm-1'. This method has the shortest down time compared to previous method.

Fig. 2 Live migration process in OpenVZ



1.4 Energy Optimization and Motivation for Survey

The motivation for this work is to survey on the approaches done to reduction in energy consumption of data centres by using live migration technology in Virtual Machine/Container. Energy optimization helps building an economical, eco-friendly cloud data centre. Data centre energy is the energy used by the hardware components like switches, NIC, servers and software components include Cloud management system. According to the survey [3] the cloud data centers use more energy than traditional requirement and also by the year 2025 [2] there would be more access to data centre which would cost more energy consumption and hence more energy efficient solutions are required for an efficient future.

The survey is divided into Sect. 2.1 listing the survey on energy optimized Virtual Machine live migration, Sect. 2.2 listing the survey on energy optimized Container live migration, Sect. 3 gives conclusion on work done in energy optimization of Virtual Machine and Container live migration.

2 Survey

This section gives the survey of the recent work done on energy efficient live migration in Virtual Machine and Container. The first part of the section includes the works done in Virtual Machine migration (Table 1) and the second part include the work done in Container migration (Table 2). The survey is basically done on parameter used, merits achieved and future directions.

2.1 Virtual Machine

The following section summarises a short survey description of the work done in Virtual Machine migration.

Andrew Toutov. et. al. [26] in 2021 have used a multicriteria method for Virtual Machine migration. The work considers two main parameters overload and overheat. Overload will indicate the more tasks assigned to the server or the CPU utilization is

more (over heat) in both the cases virtual machine is migrated to different physical server. Assigning the migrating Virtual Machines to physical server is done using Hungarian method.

Sreenivasa B. L et. al. [9] in 2020 have explored mean and minimum utilization of Virtual Machine is used in selecting algorithm. Their approach calculates the mean value of the CPU utilization of all the host that has Virtual Machine to be migrated, The host with min CPU utilization is selected with Virtual Machine of min MIPS is migrated. Implementation done using CloudSim has reduce the energy consumption.

Qiheng Zhou et. al. [14] in 2020 makes survey by comparing and evaluating on energy efficient techniques based on Virtual Machine Consolidation for Cloud Computing. The authors do simulated analysis of state-of-art energy efficient Virtual Machine consolidation algorithm and discusses merits and demerits of the algorithm under different workloads. The comparisons are done by implementing and analysing all the algorithms using CloudSim and different parameters.

Xiaojun Ruan et. al. [12] in 2020 have explored a novel approach as, Virtual Machine allocation and migration based on performance-to-power ratio in energy-efficient clouds. Their approach proposes a unique PPR ratio, where Virtual Machines are allocated to the host computer based on the most efficient PPR ratio. The PPR ratio is calculated based on number java operations completed on server, when the power consumption of host is active average in that duration. The implementation is done using CloudSim. The implementation shows 69.3% reduction in energy consumption.

Youssef Saadi1 et. al. [13] in 2020 have addressed the issue of energy optimization by considering the energy-efficient approach for Virtual Machine live migration in cloud environment. The authors examine the overall Virtual Machine allocated to different PM's. The underloaded PMs are shifted and the overloaded PMs are allowed to perform periodic Virtual Machine consolidation to make an energy efficient data centre with proper resource utilization. The implementation is carried out using CloudSim and results show reduced energy consumption. The future direction of this work is to implement in real cloud environment.

Sandeep G. Sutar et. al. [4] in 2020 have proposed a novel framework called, resource utilization enhancement through live Virtual Machine migration in cloud using ant colony optimization algorithm. The authors in their approach have used the resource allocator that checks memory and CPU usage of the PM (physical machine) and migrated Virtual Machine from overloaded or underloaded to another PM which can accommodate the migrated Virtual Machine. The optimizer uses ant colony algorithm and receives the overloaded or underloaded of Virtual Machine's using resource allocator, which then uses to generate a new list of Virtual Machine and PM mapping. The work is analysed using CloudSim and results show improvement in energy optimization. The author further suggests implementing the algorithm by removing unnecessary file during Virtual Machine migration.

Bhanu Pratap Singh et. al. [6] in 2020 performed a study on Energy Consumption of DVFS (Dynamic Voltage and Frequency Scaling), and Simple Virtual Machine Consolidation Policies in Cloud Computing Data Centres Using CloudSim Toolkit. This study compares the live Virtual Machine migration using power aware

DVFS, THR-RS (static Threshold Random Selection), (LR-MU) Local Regression Minimum Utilization, to know the energy consumption and concludes that LR-MU policy efficiently optimizes energy consumption when compared to other algorithms.

Jenia Afrin Jeba et. al. [7] in 2019 developed an approach that use static and dynamic energy consumption of the host machine to reduce the power consumption in the algorithm and further live migration is performed based on the resource utilization of the host machine. The system model uses a random or fitness function for Virtual Machine-PM mapping. The implementation is done using CloudSim and Cloudera. The work improves energy minimization by 30% and the work further suggest for hardware level optimization also.

Getzi Jeba et. al. [5] in 2018 designed a future prediction based Virtual Machine migration in cloud data centres. Their approach aimed to migrate Virtual Machine based on load aware. A combined forecast model involving weighted moving average (WMA), Exponential Smoothing Average (ESA) and Hatt Winter's Method (HWM) and AR model are used to forecast the resource utilization like CPU, memory, disk. Then the load aware method is used to allocate appropriate Virtual Machine to PM. The algorithm was analysed using CloudSim and the results showed 40% reduction in energy consumption. The work future directs to consider network bandwidth.

Amany Abdelsamea et. al. [8] in 2017 approached a Virtual Machine consolidation method which use regression algorithms such as multiple regression host overloaded detection algorithm and hybrid local regression host overloaded detection algorithm and compares both the methods using CloudSim. The methodology used in calculating an overloaded host involves predicted values of CPU utilization, bandwidth used and RAM utilization then OLS multiple Regression function [26] is used to determine the Virtual Machine to be migrated. The results showed 25% reduction in energy usage. The author suggests a future implementation of algorithm in real time cloud environment.

F´abio D. Rossi et. al. [10], in 2016 proposed ACPI (Advance Configuration Power Interface), an OS feature to support power management was implemented in their work. ACPI identifies different levels of states such as global state, device state, sleep state and processor state. The methodology uses these different level of power states to determine which PM to be shut down and to select Virtual Machine to be migrated to a PM. The implementation was done in real time and CloudSim both and results show 25% reduction in energy consumption. The author suggests a future work to improve communication between Virtual Machines during migration.

Mehiar Dabbagh et. al. [11] in 2016 proposed, A framework for energy-efficient Virtual Machine Prediction and Migration. The author in their algorithm avoids overloading of PM by Virtual Machine resource usage with prediction method and also efficiently migrates Virtual Machine to appropriate PM based on prediction. The experiments were carried out using CloudSim and the results showed improvement in minimizing energy.

Table 1 Summarizes the above work with a highlight on objectives, methods, advantages and disadvantages of each research work.

Table 1 Survey of live migration in virtual machine

Sl.no	Paper	Objective function	Simulation/realtime	Merits achieved	Future directions
1	MOP[26](2021)	CPU Utilization and resource overload	Simulator	Reduction in energy consumption and resource management	NA
2	VMS [9] (2020)	Mean value of CPU utilization	CloudSim	Reduction in energy consumption	NA
3	Energy Efficient Algorithms: Comparisons and Evaluations [14] (2020)	CPU	CloudSim	Comparing the state -of-algorithms from multiple parameters perspective for energy reduction	Comparison can be made on google trace or use network bandwidth as parameter comparison
4	Performance-to-power ratio [12] (2020)	PPr ratio using application	CloudSim	69.3%reduction in energy consumption	NA
5	Energy-efficient strategy for Virtual Machine consolidation [13] (2020)	CPU	CloudSim	Reduce energy consumption	Implement in real time cloud environment
6	Energy Consumption of DVFS [6] (2019)	Power	CloudSim	DVFS is best method among all the traditional methods used for energy consumption	NA
7	Ant colony optimization algorithm [4] (2020)	CPU and memory usage	CloudSim	Energy optimization and reduction in migration time	Removing unnecessary files in Virtual Machine to reduce downtime
8	Green Cloud Computing [7] (2019)	Host idle and active power	CloudSim and Cloudera	30% reduction in energy consumption	Consider hardware level optimization

(continued)

Table 1 (continued)

Sl.no	Paper	Objective function	Simulation/realtime	Merits achieved	Future directions
9	Forecast-based Virtual Machine migration [5] (2018)	CPU, memory and disk size	CloudSim	40% reduction in energy consumption	Analyze the influence of network bandwidth on migration
10	Hybrid regression algorithms [8] (2017)	CPU, Memory and bandwidth	CloudSim	25% reduction in energy consumption	Apply algorithm to real cloud environment
11	Virtual Machine Prediction and Migration [11] (2016)	CPU and memory	Real traces using google	Reduce SLA and energy consumption	Analyze using other realtime traces

2.2 Container

This section explores the works done in Container live migration.

Niloofer Gholipoura et. al [16], in the year 2020 proposed a novel method that implements resource management which uses energy consumption method using joint Virtual Machine and Container live migration novel way for optimizing energy consumption and support green computing in cloud data centres, The work has joint Virtual Machine and Container consolidation solution, where more than one criterion are used to decide to migrate Virtual Machine or Container to help reduce energy consumption and SLA violation. The JVCMMMD (Joint Virtual Machine Container Multi-Criteria Migration Decision) policy used in methodology finds a overloaded Virtual Machine or Container based on the Co-relation factor and then decides whether it is feasible to migrate Virtual Machine or Container to underloaded hosts. Implementation is done using Container CloudSim. Improves energy optimization by 9.9% and SLA improved by 52.8%, further suggest to implement the algorithms using heuristic algorithms.

Jialei LIU et. al. [15] in 2020, have implemented an SLA based Container consolidation which implements forecast on usage of CPU. A proactive consolidation approach is made on present and forecasted CPU utilization using history of considered PMs with the intention to optimize the power usage and also maintain the SLA. Implemented using CloudSim and results in 2.6% reduction in energy consumption compared to existing system. The work also suggests future work on improving communication between Virtual Machine's.

Tao Shi (B) et. al. [17] in 2018 have considered a multi-objective Container consolidation approach. In their approach called NSGA-II a new Container is added to Virtual Machine and existing Container are migrated to different Virtual Machine and Virtual Machines are migrated to different PM's. The methodology used in the algorithm is chromosome representation where 1 indicates non-overloaded hosts and

0 indicates overloaded hosts. Containers/Virtual Machines are migrated from 0 to 1 based on first fit policy. The implementation is analysed using Container CloudSim and results in energy consumption reduction.

Sareh Fotuhi Piraghaj et. al. [18] in 2015 have proposed a method for energy efficient Container consolidation in cloud data centres. Their work concentrates on selection Container to be migrated based on the co-correlation of the Container and the server and also the CPU utilization. The methodology uses present and future CPU utilization of the PM, if it is above a threshold value then Container with maximum CPU utilization is selected for migration The work is implemented using CloudSim. The author suggests future work to consider Virtual Machine start up delays and migration time.

Table 2 gives a sum up research work on of container live migration for energy efficiency and includes merits, future directions and objective function used.

The survey done in this section clearly has a scope in improving energy efficiency by live migration using Virtual Machine and Containers. Majority of the work surveyed in this section prefer CPU utilization as the important parameter to reduce energy optimization of the server. Some work shows the advantage of container migration over Virtual Machine migration. Some work also consider application run time to reduce energy and some others use multi criteria like present and history of CPU utilization.

Table 2 Survey of live migration in container

Sl.no	Paper	Objective function	Simulation/realtime	Merits achieved	Future directions
1	Virtual Machine and Container consolidation [16] (2020)	CPU	Container Cloudsim	9.99% reduce in energy consumption and 52.8% SLA reduction	Implement heuristic algorithms for the proposed algorithm
2	SLA [15] (2020)	CPU utilization local history of PM	CloudSim	2.6% reduction in energy consumption	Improve communication between Containers
3	Multi-objective Container Consolidation [17](2018)	CPU, migration time	Container Cloudsim	Reduce in energy consumption and no.of Container migration	NA
4	A Framework and Algorithm for Energy Efficient Container Consolidation [18] (2015)	CPU workload history	Cloudsim	Optimized energy consumption	Virtual Machine startup delays and migration time to be implemented in algorithm

3 Conclusion

In this survey made on the energy optimized Virtual Machine and Container live migration, most of the works use the parameter used to energy optimization is CPU utilization in MIPS (millions of instructions per second). Some work encourages to consider multi-parameter like memory, network bandwidth and application. Most of the work is carried using CloudSim simulator as the cost of setting up high end cloud environment is very high. There is much room for the development of energy optimization in Virtual machine and Containers. Many such works done in this survey have used CPU utilization as an objective function, Virtual Machine bandwidth used during migration, RAM used for Virtual Machine allocation can also be used as objective function for implementing the algorithms.

References

1. Clark C, Fraser K, Hand S, Hansen JG, Jul E, Limpach C, Pratt I, Warfield A (2005) Live migration of virtual machines. In: Proceedings of the 2nd conference on symposium on networked systems design & implementation, pp 273–286
2. Blackburn M, Grid G (2008) Five ways to reduce data centre server power consumption. *Green Grid* 42:12
3. Kaur T, Chana I (2015) Energy efficiency techniques in cloud computing: a survey and taxonomy. *ACM Comput Survey (CSUR)* 48(2):22
4. Sutar SG, Mali PJ, Amruta Y (2020) Resource utilization enhancement through live virtual machine migration in cloud using ant colony optimization algorithm. *Int J Speech Technol* 23:79–85
5. Paulraj GJL, Francis SAJ, Peter JD, Jebadurai IJ (2018) A combined forecast-based Virtual Machine migration in cloud data centres. *Comput Electr Eng* 69:287–300
6. Singh BP, Kumar SA, Gao X-Z, Kohli M, Katiyar S (2020) A study on energy consumption of DVFS and simple VM consolidation policies in cloud computing data centers using CloudSim toolkit. *Wireless Pers Commun* 112(2):729–741. <https://doi.org/10.1007/s11277-020-07070-2>
7. Jeba JA, Rashid ST, Whaiduzzaman AM, Roy S (2019) Towards green cloud computing an algorithmic approach for energy minimization in cloud data centers. *Int J Cloud Appl Comput* 9(1):59–81
8. Abdelsamea A, El-Moursy AA, Hemayed EE, Eldeeb H (2017) Virtual machine consolidation enhancement using hybrid regression algorithms. *Egypt Inf J* 18(3):161–170
9. Sreenivasa BL, Satyanarayana S (2020) Virtual machine's mean and minimum utilization virtual machine selection algorithm for cloud datacentre. *Int J Recent Technol Eng (IJRTE)* 8:63–67
10. Rossi FD, Xavier MG, Rose CAF, Calheiros RN, Buyya R (2016) E-eco: performance-aware energy-efficient cloud data center orchestration. *J Netw Comput Appl* 78:83–96
11. Dabbagh M, Hamdaoui B, Guizani M, Rayes A (2016) An energy-efficient virtual machine prediction and migration framework for overcommitted clouds. *IEEE Trans Cloud Comput* 6:1–13
12. Ruan X, Chen H, Tian Y, Yin S (2019) Virtual machine allocation and migration based on performance-to-power ratio in energy-efficient clouds. *Futur Gener Comput Syst* 100:380–394
13. Saadi Y, El Kafhal S (2020) Energy-efficient strategy for virtual machine consolidation in cloud environment. *Soft Comput Fusion Found Methodol Appl* 24:14845–14859

14. Zhou Q, Xu M, Gill SS, Gao C, Tian W, Xu C, Buyya R (2020) Energy efficient algorithms based on virtual machine consolidation for cloud computing: comparisons and evaluations. In: IEEE/ACM International symposium on cluster computing and the grid (CCGRID), pp 1–10
15. Liu J, Wang S, Zhou A, Xu J, Yang F (2020) SLA-driven Container consolidation with usage prediction for green cloud computing. *Front Comput Sci* 14:42–52
16. Gholipoura N, Arianyanb E, Buyya R (2020) A novel energy-aware resource management technique using joint virtual machine and container consolidation approach for green computing in cloud data centers. *Simul Model Pract Theory* 104:102127
17. Shi T, Ma H, Chen G (2018) Multi-objective container consolidation in cloud data centers. In: Mitrovic T, Xue B, Li X (eds) *AI 2018: advances in artificial intelligence*, vol 11320. *Lecture Notes in Computer Science (Lecture Notes in Artificial Intelligence)*. Springer, Cham, pp 783–795. https://doi.org/10.1007/978-3-030-03991-2_71
18. Piraghaj SF, Dastjerdi AV, Calheiros RN, Buyya R (2015) A framework and algorithm for energy efficient container consolidation in cloud data centers. In: 2015 IEEE International conference on data science and data intensive systems, pp 368–375
19. Danfoss Engineering Tomorrow. <https://www.danfoss.com/en/about-danfoss/insights-for-tomorrow/integrated-energy-systems/data-center-power-consumption/>
20. Statista. <https://www.statista.com/statistics/186992/global-derived-electricity-consumption-in-data-centers-and-telecoms/#professional>
21. Marinescu DC (2016) Cloud energy consumption. In: *Encyclopedia of cloud computing*, pp 301–314
22. Blazek M, Chong H, Loh W, Koomey JG (2004) Data centers revisited: assessment of the energy impact of retrofits and technology trends in a high density computing facility. *J Infrastruct Syst* 10(3):98–104
23. Moore JD, Chase JS, Ranganathan P, Sharma RK (2005) Making scheduling “Cool”: temperature-aware workload placement in data centers. In: *USENIX Annual technical conference, general track*, pp 61–75
24. Weiss A (2007) Computing in the clouds. *Networker* 11(4):16–25
25. Beloglazov A, Buyya R, Lee YC, Zomaya A (2011) A taxonomy and survey of energy-efficient data centers and cloud computing systems. *Adv Comput* 82:47–111
26. Toutov A, Toutova N, Vorozhtsov A, Andreev I (2021) Multicriteria optimization of virtual machine placement in cloud data centers. In: 2021 28th Conference of open innovations association (FRUCT), pp 483–487

Dynamic Shortest Path Routing Algorithm to Reduce Retransmission and Congestion Avoidance for Mobile Nodes in Wireless Sensor Network



S. Suma and Bharati Harsoor

Abstract Mobile nodes in Wireless Sensor Network (WSN) are a group of mobile sensor nodes that form a volatile topology, nodes in WSN cooperate wireless to communicate from source to destination through multihop without any infrastructure. Mobility is the key characteristics of WSN which makes nodes break links frequently and change topology. Due to link failure, the network experiences a packet loss, which could lead to congestion and degrades the overall network's quality of service (QoS) performance. It is a challenging issue for a routing protocol to adapt a proactive method for finding alternate routes with minimum delay and less energy consumption by considering congestion and link failure. To address this issue, this research paper proposes a dynamic shortest path routing by employing an Optimal Link-Sstate Routing (OLSR) scheme known as low latency optimal link-state routing (LLOLSR) to find the shortest alternate path whenever the link failure is detected between two nodes and also this scheme minimizes the path delay by integrating the congestion avoidance method and seeks to maximize the Packet Delivery Ratio (PDR).

Keywords WSN · Mobile nodes · Quality of Service (QoS) · Link quality

1 Introduction

Advancement in smart wireless computing has offered a wide range of application usage and communication between various devices for exchanging the digital data. Congestion control plays a crucial role in wireless sensor networks (WSN) to deal with network traffic, where these issues are restricted. To address this network issue, whenever there is lot of traffic, hotspot zones will appear. When the channel is overloaded at any instance, the network throughput will be measured. To eliminate

S. Suma (✉) · B. Harsoor

Department of Information Science and Engineering, Poojya Dodappa Appa College of Engineering, Affiliated to VTU Belagavi, Kalaburagi, Karnataka 585102, India
e-mail: sumas@pdaengg.com

B. Harsoor

e-mail: bharati_a@rediffmail.com

network traffic congestion, numerous congestion control techniques are proposed [1]. The mobile nodes present in WSN will communicate with each other through a wireless link. Every mobile node acts as both sender and receiver and transmit data from source to destination by cooperating with other nodes known as neighbor or relay nodes via multihop communication [2]. This type of network does not rely on any infrastructure to leverage routing services throughout the network autonomously. WSN applications are widely used in the real-time applications such as, military, wireless body area network (WBAN), Internet of thing (IoT), unmanned aerial vehicles (UAV), and natural calamities. However, this type of network has various drawbacks such as frequent topology changes, unpredictable link failure between nodes due to mobility, lack of security, high energy consumption, congestion and limited resource [3, 4]. Due to several drawbacks and load imbalance, the fast delivery of data remains as a challenging task, since nodes are tiny with limited bandwidth, energy, memory, and resources [5]. The network gets congested whenever the node disjoints occur due to the mobile nature of the node [6]. In real time, the events based on dynamic mapping and routing driven clustering face frequent connection failures, which must be addressed by providing appropriate delivery time, but there is still an overhead problem due to cluster overhead [28]. Hence, a real-time routing is established and monitored in such a way that it increases the security level, which is important for transmitting the data [29].

2 Problem Statement and Motivation

The WSN nodes operate with limited resources, battery, memory storage, and bandwidth constrain. Mobile nodes present in WSN are free to move in and around the network in any random directions with different velocities and communicate nodes within its transmission range. Due to random motion network experiences, the frequent communication link breakage between the nodes lead to network congestion and increased delay. Dynamic network topology makes the packet forwarding difficult and all nodes are required to interact with high error-prone, bandwidth-constrained, and limited queuing networks. Network congestion is not only due to redundant data but also due to the other factors such as link failure, interference, and contention. Most of the existing schemes are related to network congestion and reactive routing protocol schemes. In such schemes, it is observed that the single route path consumes more energy and it also makes the network highly congested due to heavy traffic load, which will not be uniformly distributed. However, to overcome such issues it is necessary to adopt congestion avoidance mechanism and frequently topology changes. In such scenarios, maintaining the WSN's Quality of Service (QoS) becomes a challenging issue. This leads to the selection of a routing system

for mobile nodes in order to address the aforementioned concerns and decrease packet drop to increase the network lifetime.

Related Works

In [12], the authors have proposed a hybrid QoS aware multipath routing protocol for WSN by considering different energy parameters. This routing scheme is based on multi-criteria node rank metrics, which include different parameters like residual energy and node velocity. This routing scheme intends to select a stable and energy-efficient path by using the link assessment function and high residual energy nodes available between source and destination. At regular time intervals, the topological information is flooded to all nodes in the network for developing an efficient neighboring node selection approach. During the route discovery process, multi-criteria parameters have to be updated at regular time intervals, which consumes more communication overhead and this scheme does not consider the congestion avoidance method. In [13], the author has proposed an adaptive congestion control routing scheme to bypass the route selection method to route the errors. When the source node detects congestion on the established path based on the link capacity and usage, the traffic gets distributed across an alternate path by considering the stable link availability along with the traffic split function. If congestion is not resolved by a node, it informs its neighbor node using the congestion indication bit signal to avoid the sending of packets. However, this routing scheme has more path load due to the splitting of traffic on alternative routes, which can have more communication overhead and consume more energy. The DSR protocol [14] is implemented to reduce energy consumption and load balancing over disjoint nodes but the scheme does not consider the packet retransmission and congestion. Researchers have also design an algorithm [9] based on AODV,OLSR and GPR protocols using the pause time values and by further varying the pause time, the performance can also be evaluated. The scheme does not solve the issues for an alternate route when network is loaded with high traffic and as a result the network congestion will occur. The technique proposed by [9] utilizes the traffic matrix to enhance the traffic patterns and minimize the congestion links but still the scheme is not efficient to choose an alternate route when congetion occurs in the network. In [17], the authors have implemented a congestion control scheme for performing the cross-layer adaptive transmission. The drawback of this scheme is the average end-to-end delay, which increases while the traffic rates and retransmission is not considered. In [18], the proposed CCLBARP scheme minimizes the network congestion in terms of throughput but routing overhead is higher at both packet loss and pause time. This scheme is effcient for generating a non-congestion route.

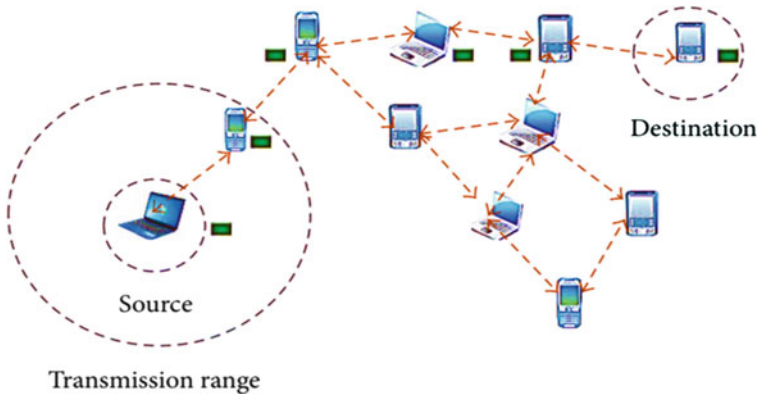


Fig. 1 WSN topology

3 Proposed Scheme

3.1 Network Model and Preliminaries

The network that consists of mobile nodes are deployed randomly in a sensing area, where each node can move in any direction and at any speed. Nodes in the network will cooperate and communicate by establishing wireless links. The network graph can be represented as $N_g = (v_n, l_n)$, where v_n represents the number of nodes and l_n represents the link between nodes.

Source node communicates to destination through multihop and routing protocol finds the best optimal path between source and destination as in Fig. 1. Due to the dynamic nature of nodes, frequent link failure and topology changes occur and new neighbors are computed by using Euclidean distance equation, which can be expressed as [19]:

$$D_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \quad (1)$$

where i and j represent the distance between two nodes.

3.2 Energy Consumption Model

The routing protocol should be able to select energy-efficient nodes to maintain a stable link lifetime and reliable data transfer [21]. Let the total energy $E_{tot}(n_i)$ represent energy consumption on the selected k path, total energy consumption is given as

$$E_{tot}(n_i) = \sum_{j=0}^{j=n_i-1} cost_{j,j+1} \tag{2}$$

n_i represents the number of nodes in k path, energy consumed for data transmission and receiver from node j to next hop is represented as $C_{j,j+1}$.

```

Algorithm for LLOLSR
Step 1: Source node wants to send data to destination node
Step 2: broadcast a HELLO message along with periodic time intervals
Step 3: Generating Link Status and neighbour nodes
if the link status is symmetric then link nodes are selected by multipoint relay (MPR) through which path is selected
end if
else
Link is lost
end if
end
if destination nodes are within one-hop then link is within the range
else
the nodes which consist two-hop count or more is not in the range
function connectivity range update
for destination do
function queue waiting time updated
for destination do
if connectivity range=0
for all nodes one and two hops MPR is selected and the packet transmission
end if
end
    
```

In Fig. 2, the proposed LLOLSR activates random mobile nodes, and the method would check for any disjoint links between nodes whenever an event takes place. Furthermore, the neighbour nodes are selected by using the Mutli Point Relay (MPR) selector. The network’s MPR selects neighbour nodes for multipoint relays. Further, the multipoint point relay nodes retransmit the control packets from neighbouring nodes that are not in the MPR (N) process for N packet control. After selecting the connectivity range, the nodes that are present within the connectivity range will be considered and the shortest distance from source to destination path will also be selected. Furthermore, while updating the connectivity range, if congestion is detected due to control packet overhead, the queue waiting time will get reduced by minimizing the retransmission. The queue waiting time iteration will occur even before transferring all the packets from source to destination without any retransmission and congestion.

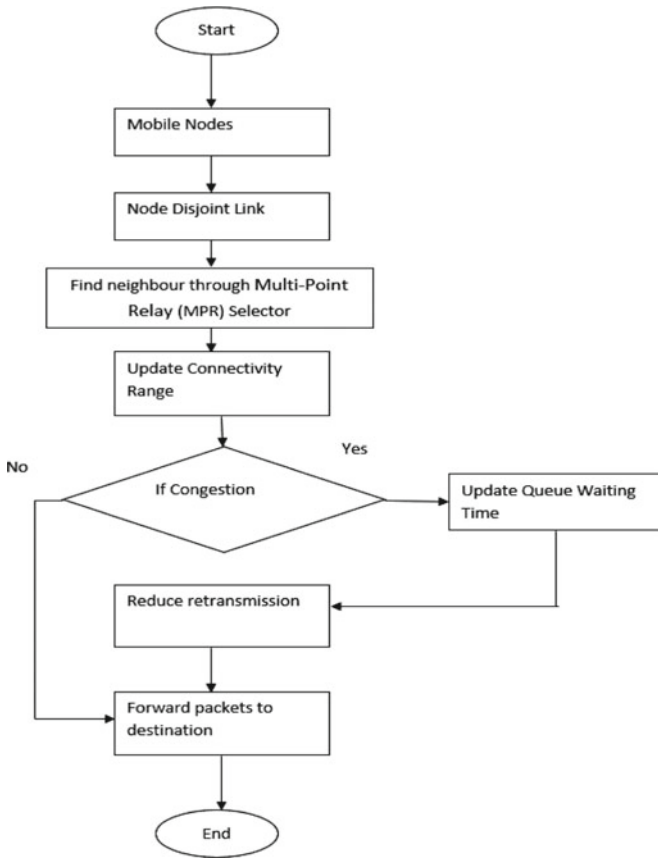


Fig. 2 Flow diagram of proposed scheme

Simulation Setup and Parameters

Performance of the proposed LL-OLSR is evaluated by using an event-driven network simulation tool [22] and further the simulation experiment is carried out by exploring different scenarios by comparing it with existing CCLBARP protocol as mentioned in Table 1 [18]. In this scenario, the mobility speed is varied to 10, 20, 30, and 40 m/s by maintaining the constant node size as 80 with a maximum of 8 connections.

3.3 Performance Metrics

i. Packet Delivery Ratio (PDR)

Packet delivery ratio(PDR) is calculated as summation of number of packets recieved to summation of number of packets sent [23].

Table 1 Simulation parameters

Parameter	Value
Nodes	20–80 nodes
Network area	1000 × 1000 m ²
Node deployment	Random
Propagation model	Two-ray-ground
Mobility model	Random waypoint
Nodes speed	10,20,30,40 and 50 m/s
MAC protocol	802.11
Queue size	50 packets
Transmission range	250 mts
Packet size	512 bytes
Initial energy	100 J
Traffic type	CBR-UDP
Protocol	LLOLSR, CCLBARP
Antenna type	Omni antenna

$$PDR = \frac{\sum \text{number of packets recieved}}{\sum \text{number of packts sent}} \tag{3}$$

Packet Delivery Ratio vs Mobility Speed: Figure 3 shows the PDR performance of the proposed LLOLSR scheme when compared to CCLBARP. It is observed that the packet delivery ratio of LLOLSR has a higher delivery ratio compared to CCLBARP due to congestion-free route selection, which aims to select nodes with better queuing occupancy.



Fig. 3 Packet delivery ratio vs mobility speed

ii. **Average End-to-End Delay (E2E-Delay)**

End-to-End delay is calculated by difference between summation of packet arrival time to packet sent by summation of inconnection [23].

$$\text{End to End Delay} = \frac{\Sigma(\text{packet appearing time} - \text{packet sent})}{\Sigma(\text{number of inteconnection})} \tag{4}$$

End-to-End Delay vs Mobility Speed: Figure 4 shows the end-to-end delay performance of the proposed LLOLSR when compared with CCLBARP. It is observed that the delay of CCLBARP is lower than LLOLSR for the nodes moving at a speed of 20 m/s, which starts increasing abruptly when the node speed is increased and faces a delay of 0.150 ms at 40 m/s. In the proposed scheme, initially the delay will be higher as the mobility speed increases LLOLSR as less delay because it checks queue waiting time which reduce the retransmission. Therefore, CCLBARP has more delay when compared to the proposed LLOLSR scheme.

iii. **Energy Consumption**

Energy consumption is calculated by summation difference of intial energy(I) to residual energy (R_{res}) [25].

$$\text{Energy Consumption} = \sum_{i=1}^n I - R_{res} \tag{5}$$

Energy Consumption vs Mobility Speed: Figure 5 shows the comparison of energy consumed by the proposed LLOLSR with CCLBARP by varying the node's mobility

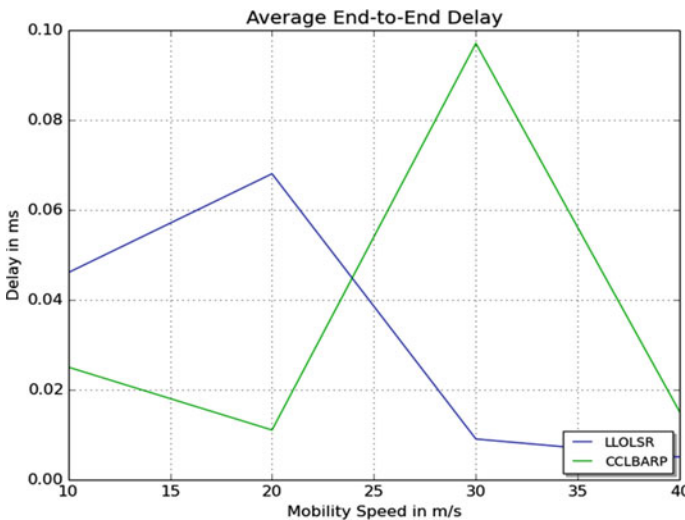


Fig. 4 End-to-end delay vs mobility speed

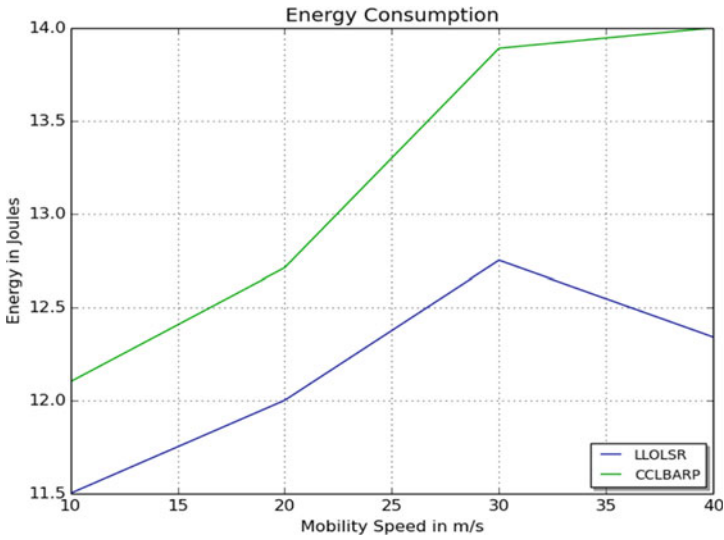


Fig. 5 Energy consumption vs mobility

speed. It is observed that the energy consumed by CCLBARP is higher and it also increases constantly. This is because the proposed scheme considers the residual energy in the node while creating a path.

4 Conclusion

Due to its several potential properties, mobile nodes in WSN technology have emerged as a unique research topic. The mobile nodes in WSN are widely used as a wireless networking technology due to its attractiveness, cost-effectiveness, and ease of usage. With its dynamic nature and frequent topology changes, several researchers have used WSN technology to regulate congestion, reliable routing, and energy usage. According to the simulation results, the proposed LL-OLSR method outperforms the existing methods in terms of PDR, end-to-end delay, and energy consumption parameters.

References

1. Nicolaou A, Temene N, Sergiou C, Georgiou C, Vassiliou V (2019) Utilizing mobile nodes for congestion control in wireless sensor networks. In: 2019 15th International conference on distributed computing in sensor systems (DCOSS), pp. 176–178

2. Sharma RK, Sharma AK, Jain V (2018) Genetic algorithm-based routing protocol for energy efficient routing in MANETs. In: Lobiyal DK, Mansotra V, Singh U (eds) Next-generation networks, vol 638. Springer, Singapore, pp 33–40. ISBN 978–981–10–6004–5.
3. Hassan MH, Mostafa SA, Budiyo A, Mustapha A, Gunasekaran SS (2018) A Hybrid algorithm for improving the quality of service in MANET. *Int J Adv Sci Eng Inf Technol* 8:1218
4. Anjum SS, Noor RM, Anisi MH (2017) Review on MANET based communication for search and rescue operations. *Wirel Pers Commun* 94:31–52
5. Hinds A, Ngulube M, Zhu S, Al-Aqrabi H (2013) A review of routing protocols for mobile ad-hoc networks (MANET). *IJNET* 3(1):1–5
6. Sharma VK, Bhadauria SS (2012) Mobile agent based congestion control using the AODV routing protocol technique for mobile ad hoc network. *Int J Wireless Mob Netw (IJWMN)* 4(2):45–52
7. Chen X, Jones HM, Jayalath ADS (2007) Congestion-aware routing protocol for mobile ad hoc networks. In: 2007 IEEE 66th vehicular technology conference, 2007. VTC-2007 fall. IEEE, pp. 21–25
8. Zaini KM, Habbal AM, Azzali F, Hassan S, Rizal M (2012) An interaction between congestion-control based transport protocols and MANET routing protocols. *J Comput Sci* 8(4):468–473
9. Abdullah AM, Ozen E, Bayramoglu H (2019) Investigating the impact of mobility models on MANET routing protocols. *IJACSA* 10:25–35
10. Lei D, Wang T, Li J (2015) Performance analysis and comparison of routing protocols in mobile ad hoc network. In: Proceedings of the 2015 fifth international conference on instrumentation and measurement, computer, communication and control (IMCCC), Qinhuangdao, China, 18–20 September 2015, pp. 1533–1536
11. Kurniawan A, Kristalina P, Hadi MZS (2020) Performance analysis of routing protocols AODV, OLSR and DSDV on MANET using NS3. In: Proceedings of the 2020 international electronics symposium (IES), Surabaya, Indonesia, 29–30 September 2020, pp. 199–206
12. Jabbar WA, Saad WK, Ismail M (2018) MEQSA-OLSRv2: a multicriteria-based hybrid multi-path protocol for energy-efficient and QoS-aware data routing in MANET-WSN convergence scenarios of IoT. *IEEE Access* 6:76546–76572. <https://doi.org/10.1109/ACCESS.2018.2882853>
13. Vadivel R, Bhaskaran VM (2017) Adaptive reliable and congestion control routing protocol for MANET. *Wireless Netw* 23:819–829. <https://doi.org/10.1007/s11276-015-1137-3>
14. Ali HA, Areed MF, Elewely DI (2018) An on-demand power and load-aware multi-path node-disjoint source routing scheme implementation using NS-2 for mobile ad-hoc networks. *Simul Model Pract Theory* 80:50–65
15. Krishnamoorthy D, Vaiyapuri P, Ayyanar A et al (2020) An effective congestion control scheme for MANET with relative traffic link matrix routing. *Arab J Sci Eng* 45:6171–6181. <https://doi.org/10.1007/s13369-020-04511-9>
16. Sharma VK, Kumar M (2017) Adaptive congestion control scheme in mobile ad-hoc networks. *Peer-to-Peer Netw Appl* 10:633–657. <https://doi.org/10.1007/s12083-016-0507-7>
17. Kumar J, Singh A, Bhadauria HS (2020) Congestion control load balancing adaptive routing protocols for random waypoint model in mobile ad-hoc networks. *J Ambient Intell Human Comput* 12:5479–5487. <https://doi.org/10.1007/s12652-020-02059-y>
19. Kumar R, Rao SV (2008) Directional greedy routing protocol (DGRP) in mobile ad-hoc networks. In: Proceedings international conference on information technology, December 2008, pp 183–188
19. Alleema NN, Kumar DS (2020) Volunteer nodes of ant colony optimization routing for minimizing delay in peer to peer MANETs. *Peer-to-Peer Netw Appl* 13:590600. <https://doi.org/10.1007/s12083-019-00772-w>
20. Basarkod PI, Manvi SS (2014) On-demand bandwidth and stability based unicast routing in mobile adhoc networks. *Int J Electron Telecommun* 60(1): 20–32. NS (2009) Network simulator-NS2. <http://www.isi.edu/nsnam/ns>

22. Sisodia DS, Singhal R, Khandal V (2017) A performance review of intra and inter-group MANET routing protocols under varying speed of nodes. *IJECE* 7:2721
23. Boushaba A, Benabbou A, Benabbou R, Zahi A, Oumsis M (2012) Optimization on OLSR protocol for reducing topology control packets. In: Proceedings of the 2012 international conference on multimedia computing and systems, Tangiers, Morocco, 10–12 May 2012, pp 539–544
24. Mafirabadza C, Khatri P (2016) Energy analysis of AODV routing protocol in MANET. In: Proceedings of the 2016 international conference on communication and signal processing (ICCSP), Melmaruvathur, Tamilnadu, India, 6–8 April 2016, pp 1125–1129
24. Suma S, Harsoor B (2020) Congestion control algorithms for traffic and resource control in wireless sensor networks. In: Satapathy SC, Raju KS, Shyamala K, Krishna DR, Favorskaya MN (eds) *Advances in decision sciences, image processing, security and computer vision*, vol 3. Learning and Analytics in Intelligent Systems. Springer, Cham, pp 750–758. https://doi.org/10.1007/978-3-030-24322-7_88
26. Suma S, Harsoor B (2019) An effective congestion control approach through route delay estimation using packet loss in wireless sensor network, 17 May 2019, 9 Pages. Available at SSRN: <https://ssrn.com/abstract=3511452> or <https://doi.org/10.2139/ssrn.3511452>
27. Suma S, Harsoor B (2022) An approach to detect black hole attack for congestion control utilizing mobile nodes in wireless sensor network. *Science direct-materials today: proceedings*. <https://doi.org/10.1016/j.matpr.2021.11.590>
27. Haoxiang W, Smys S (2019) Enhanced VANET routing protocols for dynamic mapping in real time traffic. *IRO J Sustain Wireless Syst* 1(3):139–147
28. Bhalaji N (2020) A novel hybrid routing algorithm with two fish approach in wireless sensor networks. *J Trends Comput Sci Smart Technol (TCSST)* 2(03):134–140
29. Suma S, Harsoor B (2022) Detection of malicious activity for mobile nodes to avoid congestion in wireless sensor network. In: 2022 IEEE fourth international conference on advances in electronics, computers and communications (ICAIECC), pp. 1–6. <https://doi.org/10.1109/ICAIECC54045.2022.9716673>
30. Suma S, Harsoor B (2020) Congestion control for multihop transmission in WSN using contention window. *Int J Adv Sci Technol* 29(05):13697–13703. <http://serisc.org/journals/index.php/IJAST/issue/view/274>

A Systematic Review - Attack and Security Issues in FOG Computing



C. Sabarinathan and B. Baranidharan

Abstract Fog computing is a modern era that expands the cloud computing system infrastructure by offering computer resources at the edge devices. It could be mentioned as a platform of cloud computing where data calculation, data storage and utility services are carried out, but it is decentralized. In addition, fog computing is the ability to process a huge volume of data in nearby, pre-run, fully transferable that can be implemented on multiple hardware. These properties create fog computing more appropriate for timing and location sensitivity applications. For example, Internet of Things (IoT) equipment is required to compute huge amounts of data faster. These variety-ranging functions drive applications to exacerbate different numbers of security issues related to, attacks, network security, virus, data virtualization and surveillance. This study examines the research papers available on Fog computing applications to determine common security issues such as Edge Computing, and Cloudlets and Micro-Data Centers are also included to contribute to a complete review process. Most fog applications are triggered by a preference for functionality and end-user needs, while security features are frequently overlooked or examined as an afterthought. Additionally, it describes the consequences of these safety concerns and possible solutions, as well as upcoming safety-related rules for people in charge of building, enhancing, and maintaining fog computing.

Keywords Fog computing · Data security · Privacy · Attacks · Cloud computing

C. Sabarinathan (✉)

Department of Computer Science and Engineering, SRM Institute of Science and Technology,
Vadapalani, Chennai 600026, India

e-mail: sabarinc@srmist.edu.in

B. Baranidharan

Department of Computing Technologies, SRM Institute of Science and Technology,
KattanKulathur, Chennai 603203, India

e-mail: baranidb@srmist.edu.in

1 Introduction

In the recent innovation age, information is the primary ware. The Internet of Things has increased enormously in the past decade. In 2030, 500 billion gadgets will be associated with internet reports as indicated by Cisco [1]. This rapid increase of Internet-empowered things will lead to a tremendous expansion in the pace of huge information. The fast production of vast information makes maintenance an issue of genuine reason. During these situations, cloud technology might be used as modern technology that offers several services to the end-user such as IAAS (Infrastructure as a Service), SAAS (Software as a Service), and PAAS (Platform as a Service). Cloud computing consists of a global data centre which performs the computation. These data centres compute a huge amount of IoT information and react according to the end client's requirements.

The actual distance between end-user and cloud will increase the response time transmission and latency correspondingly. IoT application needs to be followed because it generates sensitive data. Such applications are computed near the end-gadgets and the most proper arrangement for them has been named "Fog Computing". Fog Extension of cloud computing takes computation, storage, control, and interaction closer to the user [2]. Generally, fog computing is constructed using the fog layer which will be the interface between the cloud and the gadgets as presented in Fig. 1. IoT devices also have advantages and disadvantages. Several industries like transport, finance, hospitality, etc. are normal to get transferred in compatibility with the acknowledgement of IoT. These changes accompany increased security and vulnerability [3]. It has been tracked down that about 52% of the organizations can't recognize the IoT information penetration [4]. IBM and Threatcare in 2018 recorded a sum of seventeen weaknesses with four major smart city communities based on driving worldwide smart city frameworks, and from these, eight were identified as serious issues [5].

The fog layer serves as a medium for sensitive data between the cloud and IoT components [6]. Security techniques can be provided in the fog layer. Regardless of the presence of privacy-preserving and security answers in cloud computing, the equivalent strategies may not be so much required at the fog level which has given a major contrast (location and data processing) between both fog and cloud [7]. Subsequently, fog computing provides those security and protection problems. It cannot exist in all-around controlled cloud computing. Hence the best solution needs to be conveyed.

The remaining part of this paper is facilitated as follows. Section 2 is the associated with the analysis and the investigation gap examination to the extent accessible of

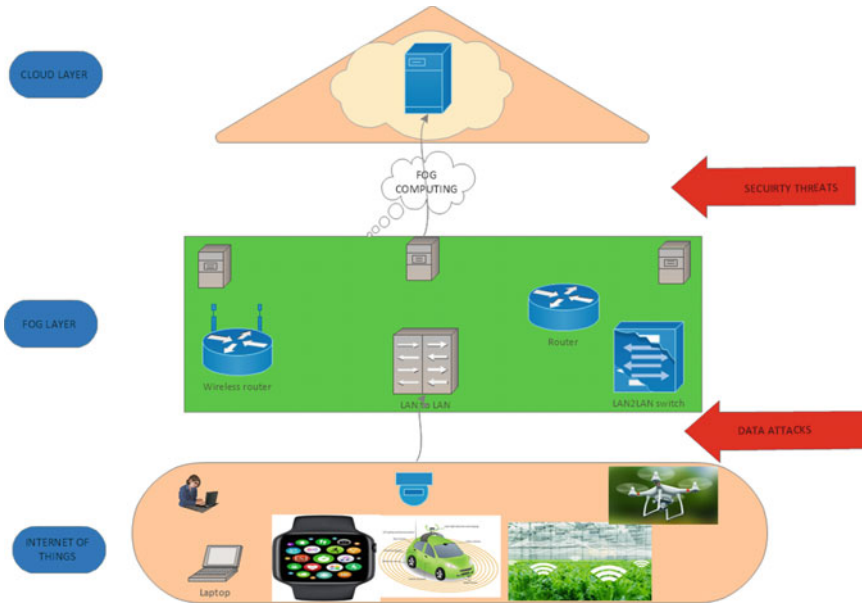


Fig. 1 Fog computing architecture

fog security-based review articles, and Sect. 3 records the survey on data attacks on fog node. The last section concludes the study with the future direction.

Research Objective

1. To understand and investigate the security study by the researchers in fog environments.
2. List out the major issue of security and privacy of fog environments according to the studies.
3. To validate the proposed security results by the investigators.

2 Fog Security Approach

2.1 Access Control

Daoud, et al., [8] proposed the security model in distributed access control for the security management along with a monitoring system and a clustering algorithm which was used for security and resource allocation, depending upon the priority. To enhance the security, the user trusts examination and monitoring to maintain the security in high level. The authors proved that their model has low latency with high security and privacy.

Kayes, et al., [9] proposed the new framework Fog-Based Context-Aware Access Control, to enable the flexi access control data from different sources. This framework reduced the computational overheads by choosing unique CAA policies and regularly maintains the user. It mapped the set of policies along with associated properties for the data access from multiple sources. This framework avoids the pitfall in security and constructs the relevant CAA solutions for privacy preservation. Along with it, it can also deal with the problems of semantic heterogeneity and data heterogeneity during the integration of data from a cloud source.

Yu, et al., [10] proposed the new protocol CPAD for secure deletion of data. The data were encrypted towards an access structure. The user having the designated attributes can have the right to decrypt the data efficiently. When a deletion request is raised, the fog device can operate, and no user can recover the key for the data and the data is unrecoverable. This protocol will protect unauthorized users from colluding attacks.

Fan, et al., [11] proposed the most efficient privacy-preserving design called PPO-MACS privacy, which was preserved by outsourcing a multi-authority access control scheme. The attributes of the user were converted into unidentified and authenticable attribute to recognize the privacy-preserving. The verification of decryption is carried out on another end to reduce the computational load on the end-user. The in-built function of single direction anonymous key agreement is used to realize the attributes of anonymous.

Li, et al., [12] proposed a new cryptographic called Secure Publicly Verifiable Revocable Large-Universe Multi-Authority Attribute-Based Encryption to obtain the fine-grained admittance control in fog computation. In the demonstration, each edge node produced the private key from several authorities, and it was differentiated by its physical location information and functions. Attributes may be represented by the strings in the universe, and they uniquely differed from each other for the fog application. This cryptographic technology supports public authentication and valid ciphertext are transmitted and saved. This security model is more efficient and flexible for public verification access control.

Sun, et al., [13] demonstrated a new design for blockchain technology with the ciphertext-strategy property-based encryption framework. Under the InterPlanetary File System, it creates a plan for secure capacity and updates for protection records (IPFS). It stores the encrypted data for ensuring security and eliminates a single point of failure from the centralized availability. This framework avoids the non-repudiation of data from both ends with the help of IPFS. The design can achieve the security for the keyword attacks.

Wang, et al., [14] conveyed a new design called the fog-based access control model. This model was implemented in the medical field. The model was used to check the validity of the user by task mapping and the schema was generated and the corresponding task was distributed. Data were classified into two types: temporary data and permanent data. The fog server acts as the access controller. This controller has a register that is used to communicate with applications. AC verifies if the incoming tasks are registered previously for privacy settings. It verifies the time,

Table 1 Hyperparameter setting for the proposed network

Tampering	The hackers stop the transmitting data over the network to corrupt and disturb the efficient fog computing
Spam	The irrelevant message flooded by the attacker
Jamming	Divert the data packets or series of denial-of-services attacks in wireless communication media by a radio frequency signal to jam large numbers of data packets
Denial-of services	Generating a high number of fake requests to the fog node to make service unavailable to authenticate the user
Eavesdropping	Attackers hack the information from computers, mobile or devices during transmission time in the network without user's knowledge
Collusion	Misguide the group of the fog node

date, and time stamp of the user. Each user is associated with a privacy value to make the access control strategy This model reduces the privacy leakage.

Ma, et al., [15] proposed an architecture using protection of privacy and hierarchy level access control. The proposed architecture utilized blockchain-based key management architecture with fog computing and cloud computing. The system was designed with different authorization methods and group access patterns. The key information is stored in the blockchain, and it's split into six categories. An encrypted access key is assigned to the new user to enhance the authorization. The lifetime of the key time is verified for the access of the objects. When the network size grows, this model's scalability is advantageous and improves system performance. The below Table 1 is based on the Hyper parameter setting for the proposed network.

Wen, et al., [16] proposed a new design called ciphertext policy Attribute-Based Encryption. Users can lose data control when data are outsourced and may arise in security issues. The group key is currently distributed using the Diffie-Hellman tree in various states. It avoids collision attacks by the valid user. The design utilized a 2PC protocol between the cloud service provider and key authority to produce the private key for the user only and this model outsourced the complex operation to the fog nodes. Users will have the proxy key and the private key to improve the security and minimize the storage overhead of the system.

Hong, et al., [17] proposed another security protecting algorithm for outsourced service from fog to cloud privacy-preserving authentication scheme by incorporating k-times unknown validation (k- TAA) and characteristic based admittance control. The expert organizations can individually choose a fine-grained admittance approach and the maximum access times for accepted clients under the proposed scheme. The benefits of this service are available to users who adhere to the access rules for a limited number of times without disclosing any personal information. The lightweight and believed charging system utilized Merkle Hash Tree (MHT), which can recognize the cloud's administration with high likelihood, without costing a lot of specialist co-op's transmission capacity and calculation.

2.2 Key Exchange Algorithm

Du, et al., [18] proposed a model that utilizes the question model to catch the design data of the reasonable fog computing processing. At that point, the author infused Laplacians to accomplish the differential protection. Hypothetical investigation demonstrates that the proposed calculations of QMA and IQMA fulfil the differential security. The explored results show that the technique can stand up to different types of protection assaults and can accomplish generally high information utility under the reason of better security savings.

Dewanta, Favian, and Masahiro Mambo [19], in consideration of the limited admission FCS in vehicular networks and the administration reservation situation at login and administration demand stage, developed a simple and safe common verification approach. The proposed plot was lightweight and efficient in the protection of private data due to the usage of a one-way hash function with exclusive OR operation and formal security investigations through the Real-Or-Random model. The BAN logic shows that the proposed strategy can fulfil and ensure the security of shared validation measures. The programming-based approval by utilizing SPAN programming dependent on AVISPA additionally confirms that the strategy has gotten against replay and man in the centre attacks. The model achieved 1.1%, 56.67 quicker calculation and reduced the quantity of message volume by 30%, and 58.21% compared along with existing scheme in the authentication.

Kumar, et al., [20] proposed a framework system that utilized a mark of the total from the fog gadgets which limits the cryptographic functions with overall functions. Also, reserve utilization was demonstrated helpful in LFCSF as the comparable questions can be dealt with by store which in the long run builds the asset usage. This calculation performed better in correlation with different calculations in terms of calculation cost and capacity cost. It additionally beat different calculations in taking care of inquiries. In normal, LFCSF performed 35% better when contrasted with different calculations.

Hussain, et al., [21] developed a context-aware method that made use of a framework for evaluating client reliability that is multi-source trustworthy and distinction-based. The approach fused the setting of an associating gadget/client for the trust assessment which used a setting-based standing model that considers the presumed hubs for the trust assessment identified with the specific situation of the interfacing hub. Trust Input and Trust Feedback Crawler framework to guarantee the trust assessment framework is fair-minded and viable.

Chen, et al., [22] proposed a plan to give solid security safeguarding by the planned testing sampling perturbation encryption method which can likewise diminish the measure of repetitive information transmissions, essentially by the information created for preparing at fog hubs. The developed improvement model for the estimation network ensured the effective reproduction of unique information with high reproduction precision. Specifically, a culmination time minimization issue was

formed for delay-delicate client demands, and an effective offloading choice calculation was created to track down the base culmination time by together enhancing the allotment of nearby CPU, outside CPU, and channel transfer speed assets.

Kashan, et al., [23] demonstrated re-encryption using a lightweight proxy for fog and IoT. Cryptosystems using hybrids dependent on lightweight encryption calculations were introduced to decrease the computation burden on fog devices and end-users. It accomplished a significant decrease in the preparing time that fog nodes need to complete the re-encryption measure. Likewise, the execution of encryption and decoding dependent on the XXTEA and ECC lightweight calculations was demonstrated. Such conspire caused the least calculation cost than standard codes, subsequently, providing the best asset obliged gadgets in a fog network.

Yu, et al., [24] proposed a property-based signcryption conspire with half breed access strategy and certain re-appropriated decoding (LH-ABSC). Additionally, most marking over-burden and confirmation burden were re-evaluated to haze hubs. The ABSC can accomplish message classification and cipher text unforgeability at the same time is more effective. While CPABE can accomplish fine-grained one-to-number information sharing while at the same time keeping the privilege of who can unscramble in information proprietor. In addition, the unique plan has a steady mark size and fulfils public check which is significant for the IoT framework. Finally, the author demonstrated CMA security, the plan for CCA secure furthermore.

3 Data Attacks on Fog Node

Saha, et al., [25] experimented on the privacy of personal data in the health-care system. Information aggregators keep away from various point cryptographic cycles and that has facilitated the circumstance requirements. Combining the query controller and functions-based admittance control components took care of the review part of the mentioned questions. The effective agreement-based methodology utilized in the system guarantees the dependability of the requester to see the EMR.

Liu, et al., [26] proposed a method that suggested a few techniques of building secure public-key encryption to conspire challenging arbitrariness attacks. It proposed an RRA-CPA secure PKE conspire with a proficient unscrambling calculation and short ciphertexts size, to acquire RRA secure PKE conspire against subjective capacity. He demonstrated that any openly deniable encryption conspire is an RRA-CPA secure public-key encryption conspire against self-assertive work standard IND-CCA PKE plot with bad-to-the-bone capacity for self-assertively corresponded contributions to get an RRA-CCA secure public-key encryption to conspire against subjective capacity. The first plot secure against discretionary capacity is freely deniable encryption conspire so it is wasteful as of now, as the realized freely deniable encryption plans are built-in light of vagary jumbling which is not pragmatic at this stage. It must be conceded that the utilization of indistinctness confusion in the initial two plans of the work makes the techniques just with hypothetical importance.

Gu, et al., [27] proposed the secure data query framework, wherein a client may quickly acquire the required information from the fog organization, and the cloud administration should test more information in the fog organization to review the last results and give to the client. Because of security prerequisites of fog processing, the system not just ensures the unwavering quality of information yet additionally viably ensures information against man-in-the-centre assault. Likewise, the tests show that the system is successful and productive.

Pacheco, et al., [28] showed how the framework is dependent on the Anomaly Behaviour Analysis Strategy, which is maintained by several artificial neural networks. The proposed strategy incorporated the utilization of a profile dependent on highlights extricated from the hub and taken care of by Artificial Neural Networks designed to precisely characterize the ordinary activities of the edge node. The approach was demonstrated to be viable in distinguishing both common and obscure assaults with high identification ranges (over 90%) and low false-positive alarms (under 3.3%), additionally having minimum overhead in terms of CPU usage, memory use, and execution time.

Samy, et al., [29] proposed attack detection frameworks based on LSTM DL that are utilized for IoT traffic grouping. The trained model was implemented in fog layer nodes to detect the attacks as in Fig. 2. It executed the identification framework on the fog node to investigate the information near the edge layer to reduce latency. It is too exhibited that the DL models can identify ordinary also. Cyber-attacks occur in various datasets. The LSTM model is better than any remaining directed DL models utilized in the examination since it has a forget gate to store the previous state data and can gain from long successions. This proposed structure defeats the issues of how to carry out the substantial DL recognition framework straightforwardly on restricted limit IoT gadgets, distinguishes a few attacks with high recognition rate and high precision rates, and instructs to screen the location framework and update it to distinguish new attacks.

Sadaf, et al., [30] introduced the Deep learning technique (Auto-IF) for intrusion detection. The approach utilized the autoencoder (AE) and Isolation Forest (IF) for the fog environments. The binary classification approach method was used to target the incoming packets in nodes and was also involved in classifying the attacks from the normal packets in a real situation. System prepared the autoencoder to learn just the typical information. Demands upon the saturation value regarding loss AE, classify the normal and attack data. The method accomplished a high precision of 95.4%. In addition, this proposed method was compared with intrusion detection.

Araujo-Filho, et al., [31] proposed FID-GAN, a unique unsupervised methodology to identify the cyber security attacks in CPSs utilizing a GAN. An identification was done according to combining and discrimination, and rebuilding lost, that demands the mapping of data samples. When compared to other works, the encoder performs the mapping to such a degree that the reconstruction loss computation is increased. It assessed both detection performance and latency detection during the time of the attack. This proposed method achieved a higher detection rate of 5.5 times speedier than IDS proposed.

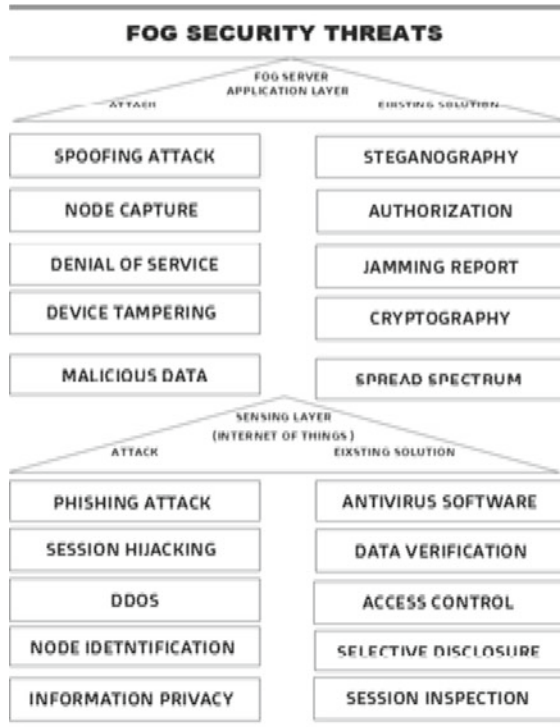


Fig. 2 Fog computing security threats

Alshehri, et al., [32] presented a novel method based on the cryptographic primitives of hashing and hypertext-policy-Attribute-Based-Encryption to decrease the amount of information at risk. Risk for penetrating by rouge fog node with keeping up the fog computing services is allowed to end-user. The algorithm solution and execution assessment method prove that the method is very efficient in reducing the total quantity of data in risk being reduced by maliciously breaking by rouge of the fog node.

Chen, et al., [33] proposed a method to reduce the complete energy utilization of calculation assignments with security, in which a fog-assistance secure three-layer registering design utilizing a descent-based energy-efficient offloading decision algorithm and computing architecture. The absolute energy utilization can be below average, by a normal of 23.1% contrasted and benchmarked PGCO arrangement.

The challenges in fog computing is shown in Table 2 and the Table 3 shows services, existing solutions, and the limitations.

Table 2 Challenges in fog computing

Method	Challenges
Access control	It identifies the user to access the resource by authentication method and task performance in the system [36]
Light-weight protocol design	Processing computing in IoT device services is performed in real-time within a short duration. Therefore, this type of protocol is intermediate between IoT devices and Fog nodes [34]
Malicious fog node	Affects the security, integrity and privacy of data. It may also reduce the performance of fog platforms [39]
Fault tolerance	Fog node works in normal mode for individual platforms, networks and sensors. If any failure occurs, the IoT device can switch to another adjacent node [37]
Trust management	The fog environment should enable two -way trust process between IoT devices and fog node [38]

Table 3 Services, existing solutions, and the limitations

Services	Existing solutions	Limitation
Network services		
Authentication	Infrastructures like public key-based and biometric-based authentication [35] Check the credential in the fog system [40] Identify the anomaly of authentication [34]	Authentication protocols are not efficient in fog systems due to the heterogeneity of IoT Node and Fog Node More Communication delays because of user mobility
Access control	Security Management framework [39] Role-based and Policy-based access control [41] Attribute-based access control [43]	Decreases the computational process due to security verification because it occupies more storage A variety of regulations and requirements are used Key management is difficult
Light-weight protocol design	Masking techniques are used in this [37] Stream cyphers and HashFunction are used	A very effective protocol needs to be built to manage the real-time services
Malicious fog node	Trust-based Mechanism Implementation of detection system [35]	Difficult to identify the fake node due to the complexity in trust management and hard to maintain the list for rogue node

(continued)

Table 3 (continued)

Services	Existing solutions	Limitation
Fault tolerance	Combining various methods and mechanisms to enable fog infrastructure, needs to operate at any point of failure	Different infrastructures are available in the same place
Data Processing		
Data integrity and identification	Symmetric and Asymmetric encryption [34] Trusted platform module	Difficulty to identify the sensitive data and duplicate data Varieties of requirements in IoT application
Data Search and protection	Symmetric and Asymmetric search based encryption Integration of hybrid key with data encryption used in one keyword search [34]	Overhead in Computational Hard to protect the valid data due to the huge no. of IoT devices
Forensics	Maintain the records for the changes in data location within the region by using the register database [41]	Capturing the log data from the fog node is challenging
Device privacy		
Data privacy	Encryption in home area network [42] Encryption based on location and attribute	Computation and processing are challenging due to the predefined key size, data loss and gaining network access in fog node
Intrusion Detection		
Intrusion detection	The soft computing techniques and frames to identify the intrusion in the social networks [44]	It increases the time utilization and security parameter within the predefined dataset
Malware detection	Machine learning techniques are used to analyse and categorize various malware [45]	The method defines only the specific dataset problem

The X-axis represents different research papers that suggested the best security solution and Y-axis represents the different methods adopted in the research papers as in Fig. 3.

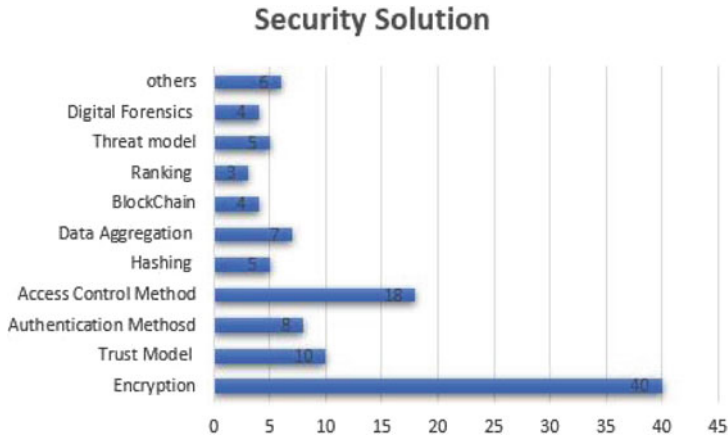


Fig. 3 Various research contribution on security solution

4 Conclusion and Future Direction

This exploration attempt depicts that fog computing discovers its application in different areas by tremendous application regions like medical care, industries, and so on. The client's sensitive information turns into a piece of the Internet. The above investigation of fog computing and systems administration shows that it is so uncovered to security dangers and malicious clients. The total amount of work carried out on fog security issues has been estimated from the analysis of most of the experiments found in the authentication. As a result, intrusion detection and network security are very low. However, there were some unresolved problems that require the researchers' and industry experts' attention. Due to the differences in features between fog computing and cloud computing, the privacy and security policy in cloud computing cannot be implemented in fog computing. Both privacy and security should be implemented in fog computing. The higher level of security and privacy implementation slows down the performance of fog nodes. Security models need to maintain both performance and security levels in fog environments. This paper discusses various levels of security and privacy issues such as, privacy location for IoT devices, fault tolerance, trust management, intrusion, access control, data protection. Further, researchers must implement new tools and technologies in fog computing environments.

References

1. Shahrestani S (2018) Internet of Things and Smart Environments. Springer international, Cham. <https://doi.org/10.1007/978-3-319-60164-9>

2. Chiang M, Zhang T (2016) Fog and IoT: an overview of research opportunities. *IEEE Internet Things J* 3(6):854–864
3. Pal S, Hitchens M, Rabehaja T, Mukhopadhyay S (2020) Security requirements for the internet of things: a systematic approach. *Sensors* 20(20):5897
4. Zalewski J (2019) IoT safety: state of the art. *IT Prof* 21(1):16–20
5. Yousuf O, Mir RN (2019) A survey on the internet of things security: state-of-art, architecture, issues and countermeasures. *Inf Comput Secur* 27(2):292–323
6. Singh SP, Nayyar A, Kumar R, Sharma A (2019) Fog computing: from architecture to edge computing and big data processing. *J Supercomput* 75(4):2070–2105
7. Choo K-K, Rongxing L, Chen L, Yi X (2018) A foggy research future: advances and future opportunities in fog computing research. *Futur Gener Comput Syst* 78:677–679
8. Daoud WB, Meddeb-Makhlouf A, Zarai F (2018) A trust-based access control scheme for e-Health cloud. In: 2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA), pp 1–7. IEEE
9. Kayes AS, Rahayu W, Watters P, Alazab M, Dillon T, Chang E (2020) Achieving security scalability and flexibility using fog-based context-aware access control. *Futur Gener Comput Syst* 107:307–323
10. Yu Y, Xue L, Li Y, Xiaojiang D, Guizani M, Yang B (2018) Assured data deletion with fine-grained access control for fog-based industrial applications. *IEEE Trans Ind Inf* 14(10):4538–4547
11. Fan K, Huiyue X, Gao L, Li H, Yang Y (2019) Efficient and privacy preserving access control scheme for fog-enabled IoT. *Futur Gener Comput Syst* 99:134–142
12. Li D, Liu J, Qianhong W, Guan Z (2019) Efficient CCA2 secure flexible and publicly-verifiable fine-grained access control in fog computing. *IEEE Access* 7:11688–11697
13. Sun J, Yao X, Wang S, Ying W (2020) Non-repudiation storage and access control scheme of insurance data based on blockchain in IPFS. *IEEE Access* 8:155145–155155
14. Wang X, Wang L, Li Y, Gai K (2018) Privacy-aware efficient fine-grained data access control in Internet of medical things based fog computing. *IEEE Access* 6:47657–47665
15. Ma M, Shi G, Li F (2019) Privacy-oriented blockchain-based distributed key management architecture for hierarchical access control in the IoT scenario. *IEEE Access* 7:34045–34059
16. Wen M, Chen S, Rongxing L, Li B, Chen S (2019) Security and efficiency enhanced revocable access control for fog-based smart grid system. *IEEE Access* 7:137968–137981
17. Hong J, Xue K, Gai N, Wei DS, Hong P (2018) Service outsourcing in F2C architecture with attribute-based anonymous access control and bounded service number. *IEEE Trans Dependable Secure Comput* 17(5):1051–1062
18. Du M, Wang K, Liu X, Guo S, Zhang Y (2017) A differential privacy-based query model for sustainable fog data centers. *IEEE Trans Sustain Comput* 4(2):145–155
19. Dewanta F, Mambo M (2019) A mutual authentication scheme for secure fog computing service handover in vehicular network environment. *IEEE Access* 7:103095–103114
20. Kumar G et al (2020) A novel framework for fog computing: lattice-based secured framework for cloud interface. *IEEE Internet Things J* 7(8):7783–7794
21. Hussain Y et al (2020) Context-aware trust and reputation model for fog-based IoT. *IEEE Access* 8:31622–31632
22. Chen S, Zhu X, Zhang H, Zhao C, Yang G, Wang K (2020) Efficient privacy preserving data collection and computation offloading for fog-assisted IoT. *IEEE Trans Sustain Comput* 5(4):526–540
23. Khashan OA (2020) Hybrid lightweight proxy re-encryption scheme for secure fog-to-things environment. *IEEE Access* 8:66878–66887
24. Yu J, Liu S, Wang S, Xiao Y, Yan B (2020) LH-ABSC: a lightweight hybrid attribute-based signcryption scheme for cloud-fog-assisted IoT. *IEEE Internet Things J* 7(9):7949–7966
25. Saha R, Kumar G, Rai MK, Thomas R, Lim SJ (2019) Privacy Ensured $\{e\}$ -healthcare for fog-enhanced IoT based applications. *IEEE Access* 7:44536–44543
26. Liu P (2020) Public-key encryption secure against related randomness attacks for improved end-to-end security of cloud/edge computing. *IEEE Access* 8:16750–16759

27. Gu K, Na W, Yin B, Jia W (2019) Secure data query framework for cloud and fog computing. *IEEE Trans Netw Serv Manag* 17(1):332–345
28. Pacheco J, Benitez VH, Felix-Herran LC, Satam P (2020) Artificial neural networks-based intrusion detection system for internet of things fog nodes. *IEEE Access* 8:73907–73918
29. Samy A, Haining Y, Zhang H (2020) Fog-based attack detection framework for internet of things using deep learning. *IEEE Access* 8:74571–74585
30. Sadaf K, Sultana J (2020) Intrusion detection based on autoencoder and isolation forest in fog computing. *IEEE Access* 8:167059–167068
31. De Araujo-Filho PF et al (2020) Intrusion detection for cyber–physical systems using generative adversarial networks in fog environment. *IEEE Internet Things J* 8(8):6247–6256
32. Alshehri M, Panda B (2020) Minimizing data breach by a malicious fog node within a fog federation. In: 2020 7th IEEE International Conference on Cyber Security and Cloud Computing (CSCloud)/2020 6th IEEE International Conference on Edge Computing and Scalable Cloud (EdgeCom), pp 36–43. IEEE
33. Chen S, You Z, Ruan X (2020) Privacy and energy co-aware data aggregation computation offloading for fog-assisted IoT networks. *IEEE Access* 8:72424–72434
34. Ni J, Zhang K, Lin X, Shen X (2017) Securing fog computing for internet of things applications: challenges and solutions. *IEEE Commun Surv Tutor* 20(1):601–628
35. Yi S, Qin Z, Li Q (2015). Security and privacy issues of fog computing: a survey. In: Xu K, Zhu H (eds) *Wireless Algorithms, Systems, and Applications*. WASA 2015. LNCS, vol 9204, pp 685–695. Springer, Cham. https://doi.org/10.1007/978-3-319-21837-3_67
36. Zhang PeiYun, Zhou MengChu, Fortino G (2018) Security and trust issues in fog computing: a survey. *Futur Gener Comput Syst* 88:16–27
37. Hu P, Dhelim S, Ning H, Qiu T (2017) Survey on fog computing: architecture, key technologies, applications and open issues. *J Netw Comput Appl* 98:27–42
38. Mukherjee M et al (2017) Security and privacy in fog computing: challenges. *IEEE Access* 5:19293–19304
39. Khan S, Parkinson S, Qin Y (2017) Fog computing security: a review of current applications and security solutions. *J Cloud Comput* 6(1):1–22
40. Moysiadis V, Sarigiannidis P, Moscholios I (2018) Towards distributed data management in fog computing. *Wirel Commun Mob Comput* 2018:1–14, 7597686, 14 p
41. Dang TD, Hoang D (2017) A data protection model for fog computing. In: 2017 Second International Conference on Fog and Mobile Edge Computing (FMEC), pp 32–38. IEEE
42. Prakash P, Darshaun KG, Yaazhlene P, Ganesh MV, Vasudha B (2017) Fog computing: issues, challenges and future directions. *Int J Electr Comput Eng* 7(6):3669
43. Kumar P, Zaidi N, Choudhury T (2016) Fog computing: common security issues and proposed countermeasures. In: 2016 International Conference System Modeling & Advancement in Research Trends (SMART), pp 311–315. IEEE
44. Sathesh A (2019) Enhanced soft computing approaches for intrusion detection schemes in social media networks. *J Soft Comput Parad (JSCP)* 1(02):69–79
45. Vivekanandam B (2021) Design an adaptive hybrid approach for genetic algorithm to detect effective malware detection in android division. *J Ubiquitous Comput Commun Technol* 3(2):135–149

Examination of Water Impurities Using IoT and Machine Learning Techniques



M. Pyingkodi, K. Thenmozhi, K. Nanthini, M. Karthikeyan, T. Kalpana,
and P. V. Deepak

Abstract Using the analogue front circuit LMP91200, the microcontroller ATMEGA328, and the GSM module, this venture attempts to construct a pH sensor that is wireless and intelligent that can keep track of the pH of a nutritional in real-time solution. The experiment demonstrates that this apparatus has a high level of performance. Level of accuracy, up to 0.01, and that it can connect to a corporate cloud services platform to perform operations such as accurate pH value acquisition and calibration. The interaction with the cloud platform via TCP/IP PROTOCOL, Android Remote measurement through APP and PC is stable. With no packet loss after 12 h of testing; because of the high uploading speed, the device networking and upload can be completed in 5 s; and, in addition, the traditional method of nutrient solution measurement has been altered. This paper discusses the concepts of the water contaminants, salts and the hydrogen ions present in the water. We have also analyzed the concepts in fresh water as well as waste water to sum up the results using machine learning technology using linear regression.

Keywords TDS—Total dissolved solids · Ph—Potential of hydrogen · EC—Electrical conductivity

M. Pyingkodi (✉) · K. Nanthini · M. Karthikeyan · T. Kalpana · P. V. Deepak
Department of Computer Applications, Kongu Engineering College, Erode, India
e-mail: pyingkodikongu@gmail.com

K. Nanthini
e-mail: nandhini@kongu.ac.in

M. Karthikeyan
e-mail: mkarthikeyan@kongu.ac.in

K. Thenmozhi
Department of Computer Science, Kristu Jayanti College, Bengaluru, India

1 Introduction

Water is used in multiple activities such as consumption, agriculture, and travel, all of which have an impact on water quality. Water is now being polluted by a variety of businesses, resources, and analysis of the water is the most important thing for globalization to address this enormous disadvantages. Water monitoring should be done on a constant basis [20]. As a result, water quality monitoring is critical, which includes PH, turbidity, temperature, BOD, and TDS, among other chemical parameters [5]. Organic, nutrient problems are the primary causes of water quality problems in surface water bodies.

A compelling water quality observing is a fundamental component for displaying an important component to keep an eye on the water to identify the presence of any biological wastes in the water that could pollute the environment cause a large or minor issue to human health as well as crop cultivation in agriculture. 70% of all industrial waste is thrown into bodies of water around the world. The contribution of home rubbish and sewage, which accounts for 80% of global water contamination, is even bigger than the dumping of dangerous chemicals.

The plan identifies the potential of a water sample, which refers to the concentration of hydrogen bonds useful in determining ph value on pH scale, with Acidity is indicated by values below 7, and alkalinity by values above 7 [9]. ph reading measures from zero to fourteen, with values below 7 indicating acidity and values below 7 indicating alkaline as well as pure water has pH water level of 7 [10]. Water acidity and basicity must be closely monitored.

This paper covers how to assess the acidity and basicity of both fresh and waste water utilising ph sensors and chemical tests in order to determine their acidity and alkali. provide the data on liquid screen display for user communicate customer over internet implementing techniques in identification of quality of water through the use of a The data from the GSM module is kept in the cloud for future [12]. On land, the ESP8266 The Wi-Fi module utilised to join people and machines.

2 Materials and Methods

2.1 Device

The Arduino ATMEGA 328 controller as shown in Fig. 1 is used in this device. 16 MHz quartz crystal, There are 14 digital input/output pins on this board. (This can be used to generate PWM outputs), six analogue ICSP headers, inputs, USB connector [6], a power outlet, and reset button are all found adjacent the board. It includes all we need to start. Support microprocessor; all you have to do is It should be connected to a computer or use an AC-to- DC converter to provide the necessary energy. A computer, another Arduino board, or other microcontrollers are all possible options can all be communicated with using the Arduino Uno.

Fig. 1 Arduino
ATMEGA328



Fig. 2 Transformer 240



The transformer acts as a step down converter, converting 240 V AC to 12 V AC as in Fig. 2. Power supply for a variety of projects and circuit boards are available. Step down 230 V AC to 12 V AC with a maximum current of 1 amp.

Transformers can be used to transform AC voltage, current, and waveform in AC circuits. In electronic equipment, the transformer is crucial. Transformer transformation and commutation almost achieve AC and DC voltage in power supply equipment.

Transformers are used to modify AC voltage levels, and step-up (or step-down) transformers are used to increase (or reduce) voltage levels. Transformers can also be employed to create galvanic isolation between circuits and to connect signal-processing stages. Transformers have become indispensable for the transmission, distribution, and use of alternating current electric power since the introduction of the first constant-potential transformer (1885).

For E-blocks, an LCD display has been built. It has a single D-type connector with 9 ways and a LCD display with 16 characters and 2 lines of alphanumeric characters. This enables the device to connect to the majority of I/O ports on E-Blocks. The LCD screen necessitates serial data, which is described in full in the user guide below.

It operates on 5 V is also required for the display. Please ensure that the voltage does 5 V is not to be exceeded. Will the device is damaged. A 5 V fixed regulated power supply or the E-blocks Multi programmer is the finest sources of 5 V. The 224 distinct characters and symbols can be displayed on the intelligent alphanumeric dot matrix displays are 16 × 2 in size. A complete list of Pages 7 and 8 contain the characters and symbols (Please keep in mind that the symbols displayed on your

Fig. 3 16 bit LCD display

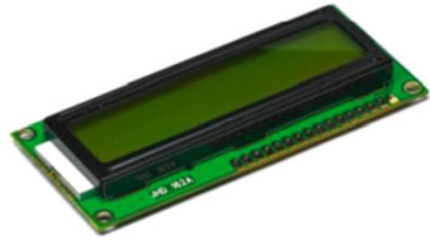


Fig. 4 GSM module



LCD as in Fig. 3 [8] may differ depending on the manufacturer.) The technical requirements for connecting the device, requires voltage of (+5 V), are listed below.

The GSM network functions as a SIM card and functions similarly to a cellphone, each with its own telephone number this modem's RS232 connection can be utilized for communication and development of embedded programmes which is an advantage of employing it? The SIM800C is a GSM/GPRS dual-band solution a single device [17]. The SIM800CS is a GSM/GPRS quad-band module that runs on the frequencies GSM850MHz and GSM900MHz provides SMS, Data, Voice, and Fax are all examples of electronic communication as shown in Fig. 4. Capability in a tiny form factor that consumes little electricity [18]. It is an SMT module with an industry-standard interface.

The circuit consists of a linear voltage regulator 7805, capacitors, and resistors, as well as a bridge rectifier composed of diodes. [7] The diodes and capacitors manage high-efficiency signal transmission in a variety of ways, from providing a constant voltage source to ensuring that output reaches the appliance without any break.

Water Pollutant Monitoring System using IoT Sensors is shown in the below Fig. 5.

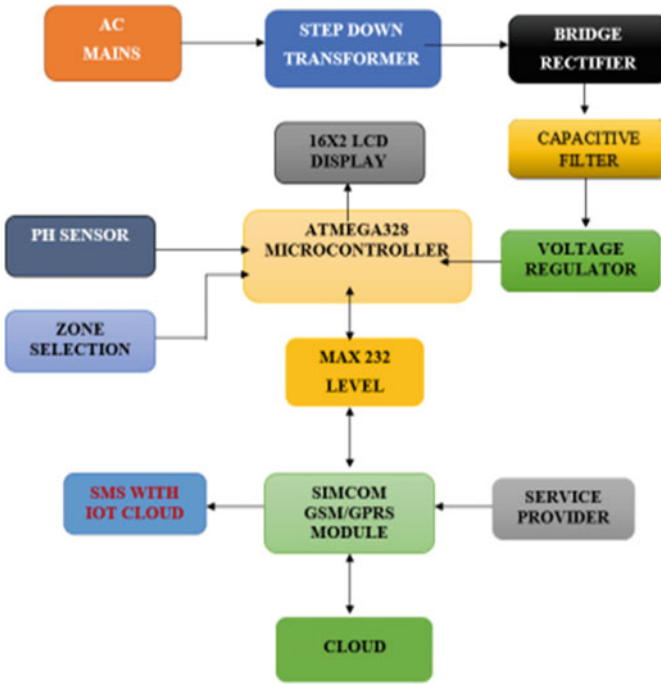
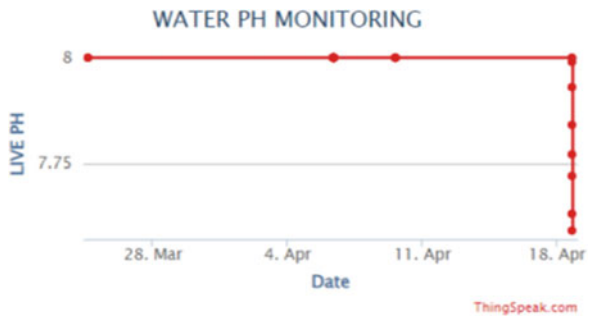


Fig. 5 Water Pollutant Monitoring System using IoT Sensors

2.2 IOT Platform

ThingSpeak is a cloud-based IoT analytics application that allows you to gather information from connected model as in Fig. 6. Visualize, analyse Data of streams that are updated in real time. ThingSpeak provides live visualisations of data supplied to it by use of your this cloud analyses online data, matlab and analysis of data as it is received ThingSpeak widely applied in the construction of IoT systems and proof-of-concept analytics.

Fig. 6 Thingspeak cloud graph of water PH monitoring



Using MQTT or a Rest API, you may submit data directly to Thing Speak from any internet- connected device. Sensor data can also be sent to Thanks to cloud-to-cloud connections things network, senet, labellum meshlium gateway, and particle.io are part of the Libelium Meshlium gateway, ThingSpeak may be used 4G/3G cellular and LoRaWAN networks.

These have the ability to store and retrieve information. Analyzed without data on the cloud having to set up web servers with ThingSpeak, and you can set up complex email that is sent in response to an occurrence notifications that are triggered depending on information gathered from your linked devices.

3 Sensors

Sensors are used as an primary source of measurements in IOT to Check accuracy of the levels of ph, TDS and EC [16].

3.1 Ph Sensor

The In water-based systems, hydrogen- ion activity are measured by Ph Sensor, which indicates the solution’s acidity or alkalinity in terms of PH [17]. The pH meter is abbreviated as a potentiometric pH metre because it shows difference in electrical potential between reference electrode and PH electrode The acidity or pH of the solution is related to the difference in electrical capacity [19]. The pH metre is utilised in a variety of settings, from laboratory testing to quality control.

The sensor is PH SENSOR measures acidity of Water. From range of 0 to 8 in accuracy of 0.01 as shown in Fig. 7.

Fig. 7 Ph sensor of range 0 to 8 Ph



3.2 *EC Sensor*

The electrical conductivity of a solution is measured by an EC Sensor as in Fig. 8 [16]. It is commonly used to monitor the amount of nutrients, salts, and contaminants in the water to monitor the amount of nutrients, salts, and contaminants in the water [4] in hydroponics, aquaculture, aquaponics, and freshwater systems.

3.3 *TDS Meter*

The suspended [6] content of all inorganic and organic compounds, whether micro-granular, analysed or molecular is measured by a TDS metres shown in Fig. 9. [2] TDS levels are frequently expressed ppm (parts per million). Water tds level are measured using a digital metre.

Fig. 8 EC meter

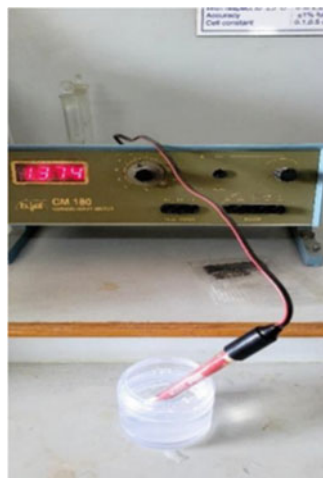


Fig. 9 TDS meter



4 Algorithm

Multiple regressions are similar to linear regression, but it includes more than one independent value, implying that we are attempting to predict a value using two or more factors. Multiple regressions, often known as MLR or multiple regressions, uses statistics that employs multiple linear regression. It can use a variety of parameters to generate a single variable. Multiple regression attempts to model the linear relationship between independent and dependent variables.

In this paper we have made an analysis of comparing the water parameters like Ph, EC and TDS. [8] Individually with each other using Linear Regression.

This algorithm uses two or more predictors to predict a dependent variable. In three issue categories, multiple regressions have extensive real-world applications: evaluating correlations between variables, producing numerical predictions, and time series. Figures 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 and 21 shows the evaluated results based on the proposed algorithm.

Fig. 10 Ph vs EC (training set)

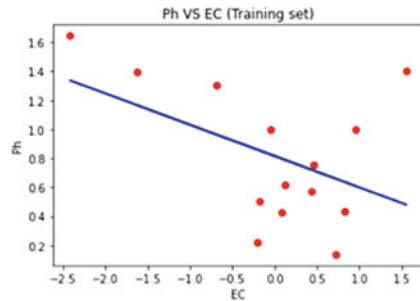


Fig. 11 Ph vs EC (test set)

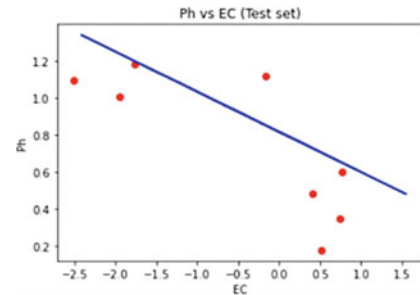


Fig. 12 Ph vs TDS (training set)

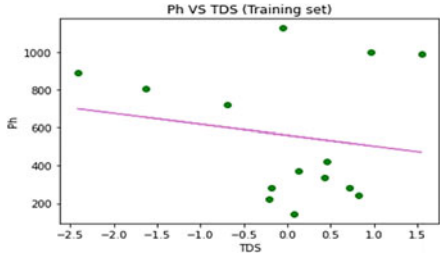


Fig. 13 Ph vs TDS (test set)

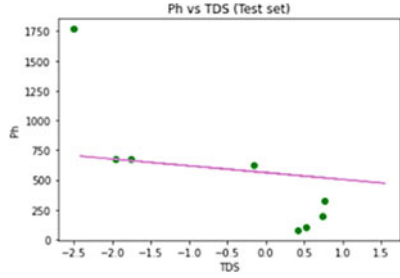


Fig. 14 TDS vs EC (training set)

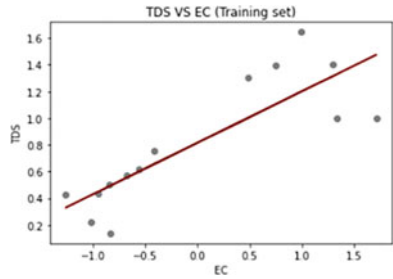


Fig. 15 TDS vs EC (test set)

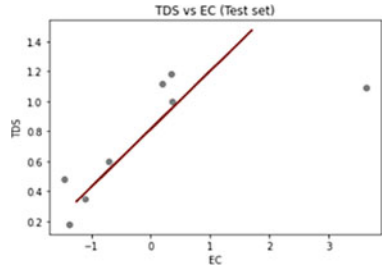


Fig. 16 TDS vs PH (training set)

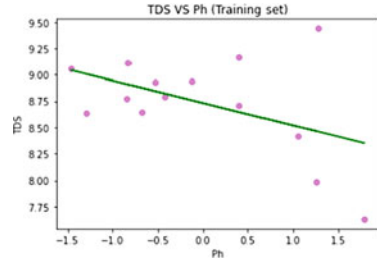


Fig. 17 TDS vs PH (test set)

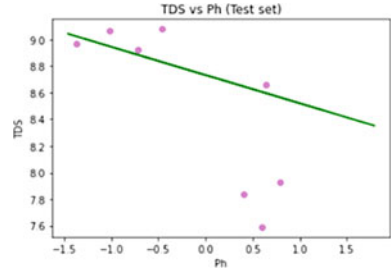


Fig. 18 EC vs PH (training set)

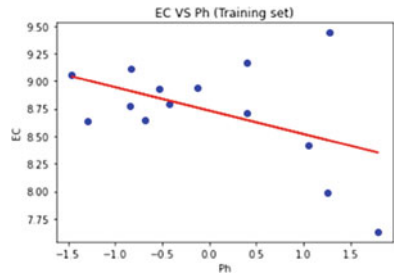


Fig. 19 EC vs PH (test set)

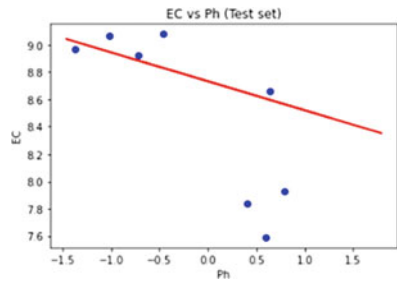


Fig. 20 EC vs TDS (training set)

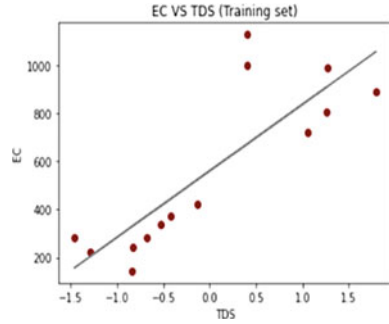
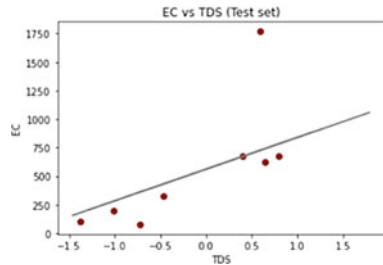


Fig. 21 EC vs TDS (test set)



5 Conclusion and Future Work

The IOT System for measuring the ph level of water is discussed in this study is identified and analyzed. The data generated from the model and the meters are created as a dataset and implemented using multi linear regression in machine learning. We have also analyzed the water parameters like Ph, TDS and EC to sum up the contents of quality of water, Salt level of water and also the ions (particles).

In This paper the future work is to say like we can implement all these concepts in all areas which can be controlled to make up the pollutants in water and find out the solution for it.

References

1. Rekha P et al (2020) Sensor based waste water monitoring for agriculture using IoT. In: 2020 6th international conference on advanced computing and communication systems (ICACCS). IEEE
2. Pappu S et al (2017) Intelligent IoT based water quality monitoring system. Int J Appl Eng Res 12(16):5447–5454
3. Martínez R et al (2020) On the use of an IoT integrated system for water quality monitoring and management in wastewatertreatment plants. Water 12(4):1096
4. Mukta M et al (2019) IoT based smart water quality monitoring system. In: 2019 IEEE 4th international conference on computer and communication systems (ICCCS). IEEE

5. Pasika S, Gandla ST (2020) Smart water quality monitoring system with cost-effective using IoT. *Heliyon* 6(7):e04096
6. Hong WJ et al (2021) Water quality monitoring with Arduino based sensors. *Environments* 8(1):6
7. Daigavane VV, Gaikwad MA (2017) Water quality monitoring system based on IoT. *Adv Wirel Mob Commun* 10(5):1107–1116
8. Putra TM et al (2021) Measurement of physical parameters of water quality in real-time based on Arduino. *J Phys Conf Ser* 1751(1)
9. Sibiyi M, Sumbwanyambe M (2020) PH sensor using Fuzzy Logic on Arduino for the monitoring and control of acidity or alkalinity in reservoir's irrigation water. In: 2020 international conference on artificial intelligence, big data, computing and datacommunication systems (icABCD). IEEE
10. Ariswati HG, Titisari D (2020) Effect of temperature on PH meter based on Arduino uno with internal calibration. *J Electron Electromed Eng Med Inform* 2(1):23–27
11. Daigavane VV, Gaikwad MA (2017) Water quality monitoring system based on IoT. *Adv Wirel Mob Commun* 10(5):1107–1116
12. Taru YK, Karwankar A (2017) Water monitoring system using Arduino with labview. In: 2017 international conference on computing methodologies and communication (ICCMC). IEEE
13. Moparthi NR, Mukesh C, Vidya Sagar P (2018) Water quality monitoring system using IoT. In: 2018 fourth international conference on advances in electrical, electronics, information, communication and bio-informatics (AEEICB). IEEE
14. Singh M, Ahmed S (2021) IoT based smart water management systems: a systematic review. *Mater Today Proc* 46:5211–5218
15. Hamid SA et al (2020) IoT based water quality monitoring system and evaluation. In: 2020 10th IEEE international conference on control system, computing and engineering (ICCSCE). IEEE
16. Kothari N et al (2021) Design and implementation of IoT sensor based drinking water quality measurement system. *Mater Today Proc*
17. Tripathy AK, Das TK, Chowdhary CL (2020) Monitoring quality of tap water in cities using IoT. In: *Emerging technologies for agriculture and environment*. Springer, Singapore, pp 107–113
18. Kanade P, Prasad JP (2021) Arduino based machine learning and IoT smart irrigation system. *Int J Soft Comput Eng (IJSCE)* 10(4):1–5
19. Johar HL et al (2021) Water quality monitoring and controlling using IoT. *J Electron Volt Appl* 2(1):20–25
20. Geetha S, Gouthami SJSW (2016) Internet of things enabled real time water quality monitoring system. *Smart Water* 2(1):1–19
21. Pyingkodi M et al (2022) IoT technologies for precision agriculture: a survey. In: 2022 6th international conference on computing methodologies and communication (ICCMC). IEEE
22. Karthikeyan M, Vijayachitra S (2021) A novel experimental study and analysis of electrocoagulation process for textile wastewater treatment using various sensors with integration of IoT monitoring system. *J New Mater Electrochem Syst* 24(2):95–102
23. Pyingkodi M, Muthukumaran M, Shanthi S, Saravanan TM (2020) Performance study of classification algorithms using the microarray breast cancer dataset. *Int J Future Gener Commun Netw* 13(2)
24. Pyingkodi M et al (2020) Hybrid bee colony and weighted ranking firefly optimization for cancer detection from gene regulatory sequences. *Int J Sci Technol Res* 9(01)
25. Pyingkodi M, Thangarajan R (2018) Informative gene selection for cancer classification with microarray data using a metaheuristic framework. *Asian Pac J Cancer Prev APJCP* 19(2):561–564. 26. <https://doi.org/10.22034/APJCP.2018.19.2.561>

26. Pyingkodi M et al (2022) Sensor based smart agriculture with IoT technologies: a review. In: 2022 international conference on computer communication and informatics (ICCCI), pp 1–7. <https://doi.org/10.1109/ICCCI54379.2022.9741001>
27. Pyingkodi M et al (2020) A novel deep learning method for identification of cancer genes from gene expression dataset. In: Mahrishi M et al (eds) Machine learning and deep learning in real-time applications. IGI Global, pp 129–144. <https://doi.org/10.4018/978-1-7998-3095-5.ch006>

Cache Coherence for Embedded Multi-core System Architectures: A Survey and Challenges



M. Thillai Rani, R. Rajkumar, K. P. Sai Pradeep, M. Jaishree,
and S. TamilSelvan

Abstract Cache coherency refers to the ability of multiprocessor system cores to share the same memory structure while maintaining their separate instruction caches. Cache coherency is used in coherence protocols to maintain data consistency between cache memory in multiprocessor systems. All cores have the same design, share same main memory (MM) and have their own cache memory. Whenever a core requests a block of data from MM for its cache, it needs a protocol to broadcast the status of blocks in MM and cores. Various hardware and software-based cache coherent mechanisms including contemporary protocols, have been thoroughly explored. This survey focuses on analyzing the different cache coherence techniques used in SoC devices. With a variety of cache coherence techniques to choose from, the best strategy is determined by a number of factors such as latency, scalability and so on.

Keywords Cache coherence · Protocols · Main memory · Coherence mechanism and SoC

M. Thillai Rani

Department of ECE, Sri Krishna College of Technology, Coimbatore, Tamilnadu, India
e-mail: thillairani.m@skct.edu.in

R. Rajkumar (✉)

Department of ECE, Vel Tech Rangarajan Dr.Sangunthala R&D Institute of Science and Technology, Avadi, Tamilnadu, India
e-mail: rajkumarramasami@gmail.com

K. P. Sai Pradeep

Department of ECE, Dr. N.G.P Institute of Technology, Coimbatore, Tamilnadu, India
e-mail: saipradeep@drngpit.ac.in

M. Jaishree

Department of ECE, Sri Ramakrishna Engineering College, Coimbatore, Tamilnadu, India
e-mail: jaishree.m@srec.ac.in

S. TamilSelvan

Department of CSE, Saveetha School of Engineering, Saveetha Institute of Medical and Technical Science, Chennai, India

1 Introduction

Cache coherence techniques have a huge influence on the performance of a centralized and distributed shared memory of multi-core systems architecture [1]. Correct execution happens when any one of the two cores perform an instruction to read a value from variable “a.” Core 2 must see the changed value if core 1 conducts a store instruction that alters variable “a,” and core 2 then performs a load instruction from that variable. As a result, the new value must be transmitted to core 2’s copy of variable “a.” This is known as the cache coherence problem shown in Fig. 1.

Consider two systems interacting in a shared-memory paradigm. They may navigate a shared address-based namespace domain, perform direct read and write operations on places inside the region, and hence transfer data through different addresses.

In a single processor context, read and write operations are primarily used for inter-processor communication. As the shared-memory model is a simple and straightforward expansion of the single processor, numerous devices need be combined for calculations to generate correct results by ensuring cache coherence [2]. This cooperation is in addition to the SoC basic memory access transaction flow. These new transactions must also be handled differently.

An action that necessitates the invalidation of particular information from cache for example shown in Fig. 2, should broadcast to all other caches in the system. When read or write exchange with no coherency have D2D communication topologies, enforcing cache coherence for the similar communications results in difficult many-to-many communication topologies. This methodology demands the collection of specific coherence responses transaction by following multi-casting of the combined coherence output. All of these actions must be supported efficiently by SoC coherence connectivity [3]. The objective of this study is to explore cache coherence protocols and their challenges for multi-core systems.

A few frameworks are normally progressive, including frameworks contained various multicore chips. Inside every cores, there is an intra-chip convention as well as a convention across the chips. Coherency issues may be fulfilled by the intra-chip convention don’t interface with the between chip convention; just when a solicitation

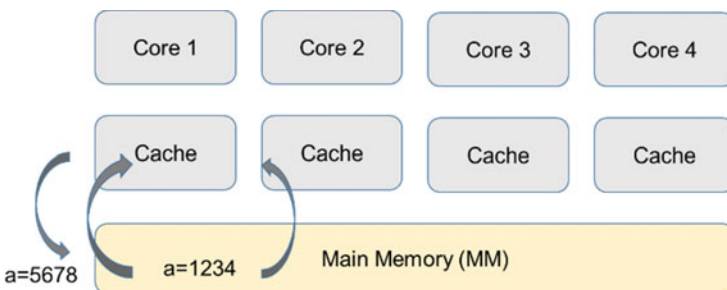
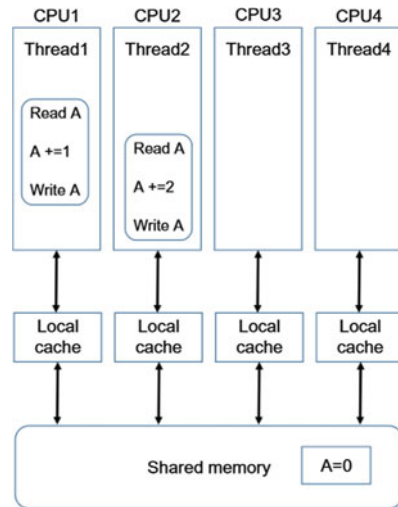


Fig. 1 Cache coherency addressed in multi-core system architecture

Fig. 2 Cache coherency addressed in shared memory architecture



can't be fulfilled by one more hub on the chip requests gets elevated between the chip convention. This decision of convention at each core is generally autonomous of the decision to next core. For instance, an intra-chip sneaking around convention can be made viable with off-chip index convention. Each chip needs a solitary registry regulator which believes that whole chip are connected with single hub in the registry convention. The off-chip registry convention could be indistinguishable from any registry conventions introduced in following section with index process normally addressed in a coarse design.

Another conceivable progressive framework is index conventions for inter-chip and intra-chip conventions. These conventions will be something very similar and unique when compared with each other. A benefit of progressive conventions for various leveled frameworks is that empowering plan of basic, possibly non-adaptable intra-chip plan for production. Once it is processed, it is useful to plan a solitary convention that scales to the biggest conceivable amount of centers which can exist in a framework. Such a convention is probably going to be needless excess in huge majority of frameworks contained a single chip.

This paper is organized as follows. Section 2 begins with overview of cache coherence mechanisms. Section 3 explains the protocols to maintain cache coherence. Section 4 investigates the challenges in coherencies followed with concluding remarks in Section 5.

2 Cache Coherence Mechanisms

The underlying interconnection system is enhanced by cache coherence, which introduces a specific group of procedures and network traffic topologies. With the intention of maintaining coherency over MM and caches, it may be necessary to exchange some information to various entities in the system. The organization of cache coherence mechanisms is shown in Fig. 3.

For hardware-based approaches, because of the instruction count is unaffected by hardware, it's tempting to use it to judge processor performance. Many a computer designer has been stymied by such oblique performance indicators [4]. The temptation for evaluating memory hierarchy performance is to focus on miss rate because it is also unaffected by hardware speed. As we'll see, the miss rate is just as deceptive as the instruction count. The average memory access time (T_{avg}) is a strong metric of memory hierarchy efficiency given by,

$$T_{avg} = T_H + T_M \cdot T_P \tag{1}$$

where, T_H , T_M and T_P are hit time, miss rate and miss penalty rate respectively.

Cache coherence mechanisms based on compilers analyze the code to identify which information are potentially dangerous for caching [5]. Snooping protocol and directory-based protocol are the two most used cache coherence techniques. Only a bus-based system can utilize the snooping protocol, which employs a number of states to identify whether there is necessity to update cache entries and control over write process to blocks. The directory-based protocol is scalable to multiple processors

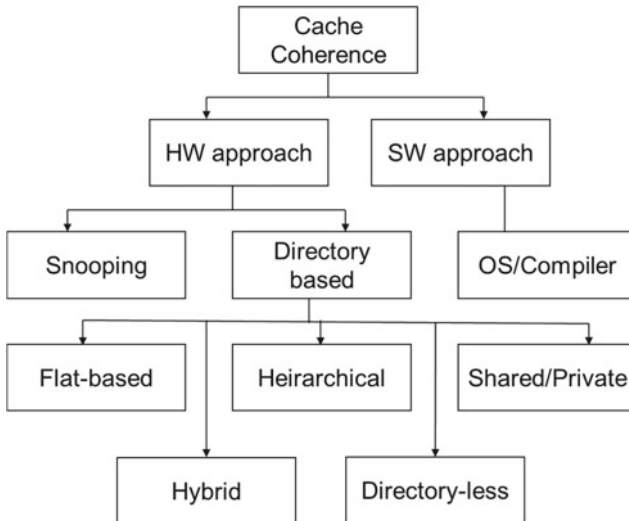


Fig. 3 Classiifcation of cache coherence

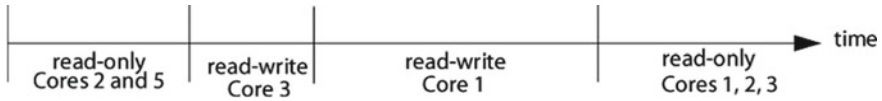


Fig. 4 Partitioning memory locations into iterations or epochs

or cores since it may be used on any network. Snooping, on the other hand, is not scalable [6].

In this architecture, a directory is utilized to keep track of which memory addresses are shared across several caches and which are reserved for one core's cache alone. Snooping, on the other hand, is not scalable. In directory-based method, a directory is utilized to keep track of which memory addresses are shared across several caches also may reserve for single cache. The directory is aware when a block needs to be updated or invalidated [5].

One method is to utilise a cache coherence technique that is based on invalidation. This method tackles the cache coherence problem by requiring that whenever a core asks to write to a cache block, the core must invalidate (delete) the block's copy from any other core's cache that has the block. The requesting core now owns the lone copy of the cache block and has complete control over its contents. When any other core tries to read the block later, it will encounter a cache miss and will have to retrieve the updated data from the core that modified it. Figure 4 demonstrates the iterative partitioning of memory locations.

Shared memory is supported in hardware by many processors and multi-core processors as well. Each of the processor cores in a shared memory system may read and write to a single shared address space. The memory consistency models specify the architectural and observable behaviour of a shared memory machine's memory system. Read (load) and write (store) operations can affect memory defined by reliability characterizations. Numerous mechanisms implement cache coherence methods to assure that multiple copies of cached data are maintained at present as part of providing a memory consistency model.

2.1 Distributed Memory Architecture

Data must be explicitly transferred between jobs in the distributed memory architecture shown in Fig. 5. Synchronous send-receive semantics can be used to accomplish task synchronization. The receiving job is paused until the data for transmission is ready. Asynchronous message passing is also possible. Send-receive semantics are employed in this technique, and the receiving process checks or is told when data is ready without blocking. This allows processing and communication to overlap, resulting in considerable speed increases.

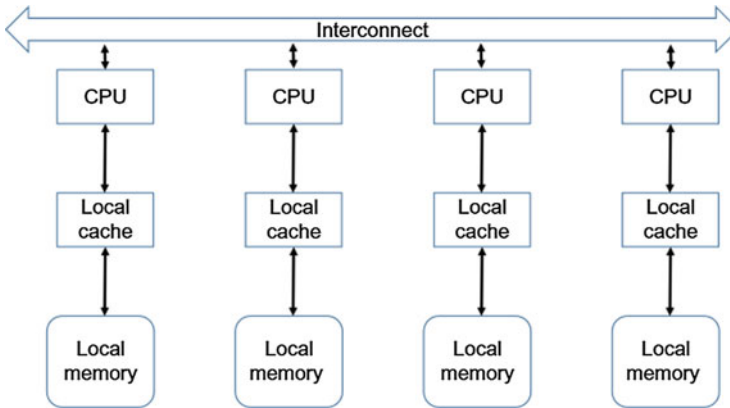


Fig. 5 Organization of cache coherence

In comparison to shared memory, the distributed memory architecture generally results in a greater communication cost (mostly in message creation and tear-down, as well as explicit data copying). As a result, message forwarding for both functions should be optimized. The fact that memory is estimated based on the number of processors is another significant benefit of the distributed paradigm. As a result, there will be an increase in the number of processors.

As a result, as the number of processors rises, so does the size of memory. Another benefit is that each CPU now has direct access to its own memory, free of interference from other cores and the expense of maintaining cache coherency. It is easier to employ commodity, off-the-shelf CPUs and networking using this method. The main downside of this design is that the programmer is now in charge of all data transmission aspects. Existing data structures based on a global memory layout may also be more challenging to translate to a distributed memory architecture.

2.2 Symmetric Multiprocessing Architecture (SMP)

Two or more identical processors are combined to particular, shared main memory in symmetric multiprocessing (SMP) systems. All I/O devices, such as UARTs and Ethernet, are accessible to SMP systems. Any processor on an SMP system may work on any job, irrespective of location where data for each operation is stored in memory. The system's tasks should not be running on two or more processors at the same time. Figure 6 gives the multicore SMP based architectures. Many ways have been proposed to improve the scalability of directories for multicore SMP. However, they often decrease directory memory cost by reducing coherence information that results in extra unneeded coherence messages and, as a result, energy wasted, performance deterioration, and lack of scalability.

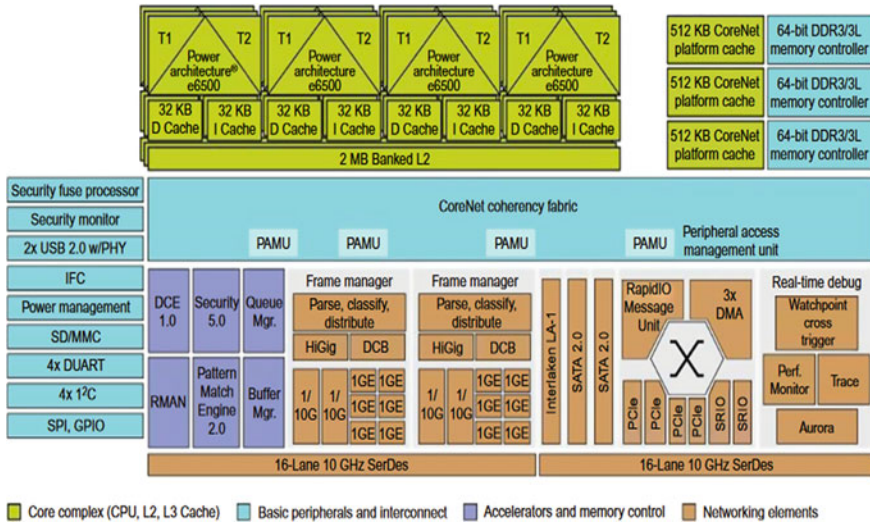


Fig. 6 Detailed structure for multicore architectures

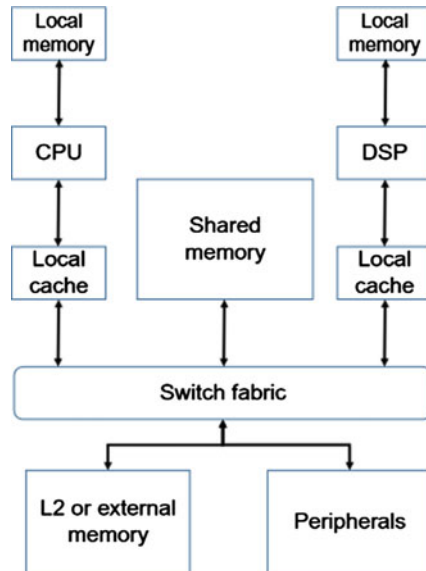
2.3 Asymmetric Multiprocessing Architecture (AMP)

All CPUs are not handled equally in an AMP systems shown in Fig. 7. First and foremost, the CPUs do not have to be identical; alternative core architectures and instruction sets can be used. Certain processors in AMP systems can be dedicated to I/O activities with certain peripherals. This sort of CPU specialisation has the potential to improve system performance.

2.4 Maintaining Cache Coherence

The preceding section’s coherence invariants give some insight into how coherence protocols function. The great majority of coherence protocols, known as “invalidate protocols,” are built with these invariants in mind. If a core reads a memory location, it transmits information to cores to determine present value of that particular memory location by ensuring that no other cores having cached read–write replicas of that same memory location. With these messages, any active read–write epoch iteration is terminated, by initialising read-only operation. When the particular core decides to write an information to any memory locations, it transmits a message to other cores to get the memory location’s present value. It does not have a definite read-only cached copy, and to ensure that no other cores have read-only or read–write cached copies of the memory location.

Fig. 7 Structure of asymmetric multiprocessing architecture



2.5 Directory Based Cache Coherence

This is a cache coherence protocol system that does not employ the broadcasting technique, therefore it must keep all of the cached information of every block in the shared data, whether it is centralized or dispersed across several processors [7]. Every memory block has a directory entry, which is the name of the structure that keeps all of the information about the shared block's various locations. Depending on the current state of the directory and transactions, the directory-based cache will take any action. The directory-based protocol must also be disseminated across the nodes. Each nodes in interconnection networks has blocks of local memory that are always associated with local cache and directory in the cache coherent non-uniform memory access. Distributing the directory-based protocol have benefits of reducing bandwidth difficulties and potential bottlenecks. Cache coherence techniques based on directories offer the ability to grow shared memory multiprocessors to huge numbers of processors. An important advantage of directory based protocols is the effective scaling over snoopy protocols. The significant characteristic of directory protocols is the ability to exploit arbitrary point-to-point interconnects.

2.6 Snooping Cache Coherence

Snooping method is used for write-invalidate and write-update protocols. Each cache listens in on bus transactions to observe what other processors are doing in memory,

and thus requires the employment of a broadcast media in the machine. One of the major advantages of snoopy protocols is the average cache-cache miss latency is very low. The bandwidth required to broadcast messages to all processors is limited by cache coherence overhead and performance of shared buses. Another issue with snoopy protocols is their inefficiency in terms of power consumption. Each cache controller has the ability to “snoop” on the network in order to detect and respond to these broadcasted notifications. Snoopy protocols are particularly suited to a bus-based multiprocessor because the shared bus provides a direct method for broadcasting [8].

3 Cache Coherence Protocols for Multi-core System Architectures

3.1 MSI Protocol

The MSI protocol detects when a cache line has been modified (M), indicating that the cache block has been changed and the data in the cache varies from data in backup repository [9]. The modified block cache is in control over updating underlying store and read data, but it will not share or send any information externally.

A block that has not been updated and is read-only in at least one cache is referred to as shared (S). Without first updating the underlying store, the cache can get the data. Invalid (I) state indicates that the block is invalid and unavailable, so it is invalidated by bus request in the present cache, requiring it to be retrieved from memory. The connection between the cache and the backup store keeps the states S, M, and I active.

3.2 MESI Protocol

MESI protocol is also termed as Illinois protocol as it was developed in University of Illinois [13–16]. This is well-known protocol that incorporates a write-back cache. Even though MESI is the extension of MSI protocol, two transitions for each write operation is processed and there is no sharing in data blocks. In the first epoch, the memory block is put into shared state, and in the second epoch, the status of data blocks are changed from modified to shared state. It includes a new Exclusive (E) state to the MSI protocol, which saves bandwidth use by writing to a shared data block.

3.3 MOSI Protocol

MOSI is a variation of MSI protocols. The additional state is called as Owned state [11] which is accurate and has the present copy of data is obtainable as the cache line is in this state. The owned state is related to the shared data memory and same as changed state since main memory may have an expired backup of the information. Only one cache may be in the owned state at a time, with all other caches holding the data in the shared state. It converts into shared state after writing by updating the MM.

3.4 MOESI Protocol

MOESI protocol [12] is classified into five states. The data update and data sharing is denoted by Owned (O) state shown in Fig. 8. This eliminates the requirement for changed data to be written back to main memory before being shared.

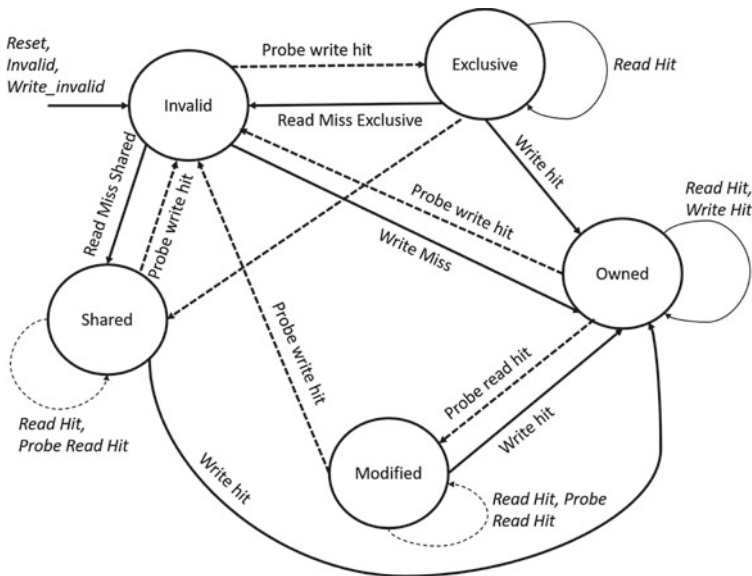


Fig. 8 MOESI protocol

3.5 Firefly Protocol

Firefly algorithm involves in three states such as Valid-Exclusive, Shared and Dirty. The cache block is valid, clear, and unique in that it only exists in one cache. The cache block is legitimate, clean, and may be found in several caches [16]. The block is the sole duplicate of the memory, and it is dirty, meaning that its value has changed since it was brought from MM. The only condition that causes a write-back is the replacement of block in the cache.

3.6 Dragon Protocol

When a write operation to a cache block is followed by many read operation by processor, update-based protocols like the Dragon protocol [17] perform well because the updated cache block is quickly accessible over every processors. It involves in four states which are as follows. The present processor was the first to fetch the cache block, and no other processor has accessed it subsequently. The cache block is very certainly present in many processor caches.

The protocol keeps states Exclusive (E) and Shared clean (Sc) distinct to prevent read–write operations on non-shared cache blocks from causing bus transactions and so slowing down the execution. In shared modified (SM), the block persists in many processor caches, with the present processor being the latest to alter it. As a result, the present processor is referred to as the block’s owner. Unlike invalidation protocols, the block only has to be updated in the processor, not the MM. When a cache block is evicted, the processor is responsible for updating the main memory. In single-threaded applications, this is a regular event. Finally, modify (M) the value from MM.

3.7 MECSIF Protocol

MECSIF is a custom established hybrid cache coherence method that employs both directory and snoopy protocols [18]. It involves in seven states including clean (C) and forward (F) states. If a copy of a data block has never been updated and at least one cache contains a copy of the data block, the processor will only respond to a remote processor request with a copy of the data block in its current state.

4 Challenges in Cache Coherency

Though the protocol discussion appears to be straightforward, the implementation is fraught with difficulties. The protocol's main flaw is that it assumes operations are atomic—that is, that an operation may be performed without any intervening actions. For example, the given protocol implies that write misses may be detected, the bus acquired, and a response received in a single atomic operation. This is not the case in reality. Similarly, read misses would not be atomic if we employed a switch, as many current multiprocessors do. Non-atomic actions increase the likelihood that the protocol may deadlock, or reach a point where it will be unable to proceed.

Any centralized resource in the system can become a bottleneck as the number of processors in a multiprocessor increases, or as each processor's memory demands increase. The bus and memory form a bottleneck in the simple case of a bus-based multiprocessor. As a result, scalability becomes a problem. Designers have employed numerous buses as well as interconnection networks, such as crossbars or tiny point-to-point networks, to boost the communication bandwidth between CPUs and memory. The memory system can be divided into many physical banks in such architectures to increase the effective memory bandwidth while maintaining a consistent memory access time. Both hardware and software is always seen as integrated and interdependent in multicore systems.

With growing complications, the resulting logical steps involved in shifting and transforming parallel hardware and software via plenty of programming models, as well as programmable processors that specify operating on multiprocessor systems-on-chips (MPSoCs) [19]. As of now, design of multicore processors is getting higher difficulties, forcing the manufacturers move ahead to block-level design, which separate the design of sub-system blocks from design of SoC platform also keeping power problems in mind. The technological aspects has enhanced the electronics hardware part to enable for verification and automated synthesis of gates from gate designs to transistors levels and eventually to register-transfer level (RTL) design [20, 21].

Abstraction, in combination with automation, translations, and validation in contradiction of the resulting lowest level of abstractions have traditionally been the response to escalating complexity, resulting in increased silicon real estate in prior decades. Both hardware and software must be viewed as integrated in multicore designs. Even the best hardware multicore architectures in industry will perform badly if only inadequate group of programmers can program to accomplish tasks. The most imaginative algorithm, on the other hand, will not run as planned if the underlying hardware design lacks sufficient computation, storage, and communication resources. The following are the common problems addressed in cache coherency mechanisms:

- 1) Detecting certain sharing patterns in order to improve coherence. We may use modified adaptability coherence mechanisms to increase system performance by exploiting application-level asymmetrical behaviours, such as application access patterns at application-level granularity.

- 2) Workload heterogeneity is taken into account while designing coherence protocols. We can change the operating system to lead coherence protocol activities and transition across protocols.
- 3) Adapting area-effective directory architecture to assure scalability of on-chip directory storage. For these goals, we can use hierarchical and tag-less design alternatives, for example.
- 4) Optimizing data placement strategy in a multi-core cache system to reduce distant cache visits by bringing private data closer to the “home core” and reducing data migration operations in the coherence protocol.

5 Conclusion

The different coherence maintenance strategies exploited in multi-cores are investigated in this survey. Cache coherence is expected to be phased away soon, according to few researchers, because it raises hardware costs by storing more state, transmitting more messages, and verifying that everything is correct.

However, we envisage that coherence will endure popular since the software cost of dealing with incoherence is always high and is recognized by few established design engineers rather than developers who should deal with it. This work aids researchers in comprehending coherence processes and investigating the implementation issues provided by the rapidly expanding number of cores. Despite significant progress in this field, it remains a very active study topic. Many research areas exist, such as protocol correctness verification, performance assessment, comparison, directory size, and protocol overhead minimization that has to be looked at in the future.

Acknowledgements The authors would like to thank reviewers for their valuable comments and suggestions.

References

1. Ros A, Acacio ME, Garcia JM (2010) A direct coherence protocol for many-core chip multi-processors. *IEEE Trans Parallel Distrib Syst* 21(12):1779–1792. <https://doi.org/10.1109/TPDS.2010.43>
2. Joshi AD, Ramasubramanian N (2015) Comparison of significant issues in multicore cache coherence. In: 2015 international conference on green computing and internet of things (ICGCIoT), pp 108–112. <https://doi.org/10.1109/ICGCIoT.2015.7380439>
3. Al-Waisi Z, Agyeman MO (2017) An overview of on-chip cache coherence protocols. In: 2017 intelligent systems conference (IntelliSys), pp 304–309. <https://doi.org/10.1109/IntelliSys.2017.8324309>
4. Jang YJ, Ro WW (2009) Evaluation of cache coherence protocols on multi-core systems with linear workloads. In: 2009 ISECS international colloquium on computing, communication, control, and management, pp 342–345. <https://doi.org/10.1109/CCCM.2009.5267596>

5. Kaushik AM, Hassan M, Patel H (2021) Designing predictable cache coherence protocols for multi-core real-time systems. *IEEE Trans Comput* 70(12):2098–2111. <https://doi.org/10.1109/TC.2020.3037747>
6. Tomasevic M, Milutinovic V (1992) A simulation study of snoopy cache coherence protocols. In: *Proceedings of the twenty-fifth Hawaii international conference on system sciences*, vol 1, pp 427–436. <https://doi.org/10.1109/HICSS.1992.183192>
7. Mittal S, Nitin (2014) A new approach to directory based solution for cache coherence problem. In: *2014 3rd international conference on eco-friendly computing and communication systems*, pp 9–13. <https://doi.org/10.1109/Eco-friendly.2014.77>
8. Bhardwaj K, Havasi M, Yao Y, Brooks DM, Lobato JMH, Wei G (2019) Determining optimal coherency interface for many-accelerator SoCs using Bayesian optimization. *IEEE Comput Archit Lett* 18(2):119–123. <https://doi.org/10.1109/LCA.2019.2910521>
9. Fuchsen R (2010) How to address certification for multi-core based IMA platforms: current status and potential solutions. In: *29th digital avionics systems conference*, pp 5.E.3-1–5.E.3-11. <https://doi.org/10.1109/DASC.2010.5655461>
10. Patel, Ghose K (2008) Energy-efficient MESI cache coherence with pro-active snoop filtering for multicore microprocessors. In: *Proceeding of the 13th international symposium on low power electronics and design (ISLPED 2008)*, pp 247–252. <https://doi.org/10.1145/1393921.1393988>
11. Yang Q, Bhuyan LN, Liu B (1989) Analysis and comparison of cache coherence protocols for a packet-switched multiprocessor. *IEEE Trans Comput* 38(8):1143–1153. <https://doi.org/10.1109/12.30868>
12. Li S, Guo D (2017) Cache coherence scheme for HCS-based CMP and its system reliability analysis. *IEEE Access* 5:7205–7215. <https://doi.org/10.1109/ACCESS.2017.2701406>
13. Martin MMK, Hill MD, Wood D (2003) Token coherence: decoupling performance and correctness. In: *30th annual international symposium on computer architecture*, 2003. *Proceedings*, San Diego, CA, USA, pp 182–193
14. Sun S, An H, Chen J (2014) Cache coherence method for improving multi-threaded applications on multicore systems. In: *2014 6th international conference on multimedia, computer graphics and broadcasting*, Haikou, pp 47–50
15. Ahmed RE, Dhodhi MK (2011) Directory-based cache coherence protocol for power-aware chip multiprocessors. In: *2011 24th Canadian conference on electrical and computer engineering (CCECE)*, Niagara Falls, ON, pp 1036–1039
16. Kaur DP, Sulochana V (2018) Design and implementation of cache coherence protocol for high-speed multiprocessor system. In: *2018 2nd IEEE international conference on power electronics, intelligent control and energy systems (ICPEICES)*, Delhi, India, pp 1097–1102
17. Lametti S (2010) Cache coherence techniques, A Technical report
18. Li J et al (2011) A new kind of hybrid cache coherence protocol for multiprocessor with D-cache. In: *2011 international conference on future computer science and education*, pp 641–645. <https://doi.org/10.1109/ICFCSE.2011.160>
19. Babu P, Parthasarathy E (2021) Reconfigurable FPGA architectures: a survey and applications. *J Inst Eng Ser (B)* 143–156
20. Durai PM (2019) Enhanced network performance and mobility management of IoT multi networks. *J Trends Comput Sci Smart Technol (TCSST)* 1(02):95–105
21. Krishnaraj N, Smys S (2019) A review of multi homing and its associated research areas along with internet of things (IOT). *IRO J Sustain Wirel Syst* 1(1):69–76

Comparison of Supervised Machine Learning Algorithms for Predicting Employee Performance on Real Time Dataset



Devanshu Joshi, Garima Sharma, Ankita Nainwal, and Vikas Tripathi

Abstract Various important factors are there for a developer that are efficient to provide profitability to an organization. It is important to know the progress of a developer as an individual. The organization may help these developers to become the most productive version of themselves by various methods. These methods will tend to provide an economic value and improve the performance of an employee as well as the organization. A survey has been made to generate the reviews which directly or indirectly influence the performance criteria such as job satisfaction, performance, productivity, etc. More than 1200 people participated in the survey and based on that Machine Learning has been applied. Finally, after performing several learning algorithms XGBoost performed well and provided a training and testing accuracy of 92.6% and 93% respectively.

Keywords Workplace · Involvement · Job · Productivity · Efficiency · Retention · Analysis

1 Introduction

The established ways of organizing and functioning are being drastically altered by global economic trends. The workplace has become more psychologically demanding and complicated, flexible working arrangements are becoming more prominent, cooperation has virtually become the trend, and the workforce composition has become significantly more varied than it was previously. Workplace changes, as well as the emergence of the digital economy, have reignited academic curiosity and shifted the study focus from jobs to work characteristics. [1] In the quest for a deeper understanding of today's employment, employees' work behavior, performance, and a broader spectrum of work characteristics have been identified.

It's important to boost software engineer's productivity. Developers who have finished their work are free to do other things, by definition. Adding new features

D. Joshi (✉) · G. Sharma · A. Nainwal · V. Tripathi
Graphic Era Deemed to be University, Dehradun, Uttarakhand, India
e-mail: devanshu0500@gmail.com

or verifying and validating fresh data. The established techniques of organizing and functioning are dramatically shifting as a result of global economic trends [2, 3]. Occupations have gotten more intellectually stimulating, tough, and complex; flexible jobs are gaining popularity; alliances have become almost the norm, and jobs are significantly more diverse than before. Significant variations in the job are required, as well as the growth of wealth, have reignited scholar's curiosity and widened interest from job-oriented to work-oriented, as well as from a piece of work to working in a group. In the quest for a better idea of modern jobs, a person's behavior, and achievements are broader characteristics of a job [4-6].

A developer's schedule can be altered by many variables, including the task that is been completed, conference, co-worker intervention, infrastructure, and office environment [6, 7]. Several characteristics induce project shifts, which can disintegrate the work and negatively affects the developer's performance, task progression, and output quality [8, 9]. As an outcome, both scholars and practitioners have been interested to learn more about how they operate and their efforts may be measured to improve performance and capability. Job involvement differs from employee satisfaction as it involves a lot of work engagement (dedication) with a high level of authentication; job happiness, on the other part, is often a more passive sort of employee fulfillment. Work engagement varies from work-oriented flow in that it refers to a lengthy performance event, whereas flow generally refers to a maximum experience lasting no more than one hour. Finally, job engagement varies from motivation in that it encompasses both cognitive (absorption) and affective (vigor) aspects (dedication). It's hardly unexpected that workplace incivility is a better predictor of employment success than many previous models.

Despite the fact that work design is the major part of performance, it is still in the research phase as a predecessor of the company's behavior. This is especially true in jobs that require a lot of mental effort. The multidimensionality of work features, for example, has not been adequately stressed [10], as well as a diverse range of knowledge work relations (such as job expertise, solving problems, and information processing) have not yet been fully examined [11]. Enhancing the breadth of work traits across task attributes and delving further into their interactions and multidimensionality may help us better comprehend the information work environment. Despite having a pretty thorough understanding of the current job pattern theory, their models should be practically confirmed in a variety of situations before being validated. The research demonstrates how specific job traits are linked to task and context performance. The results give a broader perspective of design and valuable information both in terms of HRM philosophy and use by taking into account a large variety of job variables and two unique work outcome metrics. Furthermore, because little is known about how work is designed in evolutionary economies, this empirical study shed light on the nature of occupations and their associated work traits in different environments.

This paper is divided into 4 sections. Section 2 consists of literature review of various papers which establishes the relation between employee productivity to the growth of a company. In Sect. 3, the methodology has been discussed where we will see the overall architecture of the model, and further in its sub section, we will see

different steps which are proposed in the architecture. Section 4 consists of results where the performance matrices of the model is described. Lastly, Sect. 5 summarizes the paper and discusses a realistic future perspective.

2 Literature Review

Bok et al. [12] suggested that few companies evaluate the impact of the 5 criteria (human, organizational, technology, application, and project traits) when implementing Function Point Analysis (FPA), and as a result, their productivity assessment programs are unsuccessful. As a result, the primary goal of this study was to draw conclusions from an experiment-based study in order to encourage researchers and practitioners to investigate and improve the success rates of FPA productivity assessment programs. A thorough understanding of FPA, according to this case study, is essential for success of FPA productivity program. Such understanding should extend over fundamental FPA definitions. Thus, engaging in user groups like the IFPUG and associated conferences may help companies stay on top of the newest advances. More significantly, having a strong understanding of FPA will aid in calibrating an FPPI that satisfies the measurement goals. In the measuring process, it is critical to consider organizational and human elements such as political interest, staff expertise, counting standards, and user viewpoints.

Another paper by Ramirez et al. [13] shows that employee involvement appears to be directly connected to company's achievements and results. Employee withholding is better in companies with occupied workers as a result of low turnover and purpose to quit, as well as to improve profitability, productivity, growth, and customer satisfaction. Companies with extricate staff, on the other hand, misuse resources and lack in expertise, receive not much commitment from their staff, have more absenteeism, have less customer pivot, lower productivity, and poor work culture and less profit. Most studies focus solely on the value of employee involvement and its positive effects on company results, omitting to give a cost-efficient analysis for more involvement. Engagement decisions, like any other management decision, should be weighed in terms of both its benefits and expense, with no preference given to any other in order to avoid biasing decision-makers. As a result, there is a need to investigate the financial aspects of involvement decisions. The surprising reality is that the findings of today's study may be utilized as a cornerstone for the construction of a full structure [14]. Furthermore, organizations and consultancy are assigned with most of the work connected to the "employee engagement" notion. As a result, academia must focus on this new construct and should emerge with a precise definition and dimensions that can be used to measure engagement of an employee, thereby manifesting the concept's value.

Arnold Bakker [15] research concludes that employees that are engaged are physically, intellectually, and emotionally attached to their jobs. They have a lot of energy, are committed to achieving their job-related objectives, and are frequently entirely

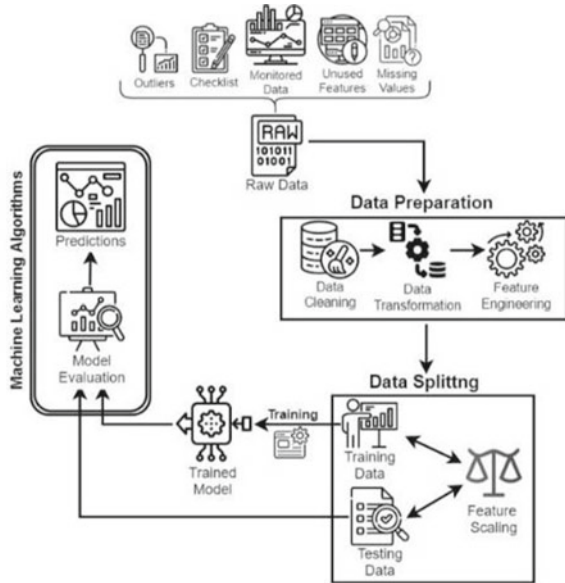
involved in their task. Work resources and individual resources predict job engagement, which leads to improved job performance. As a result, job engagement is a significant indication of individual and organizational well-being. HR managers may do a number of things to help their workers become more engaged at work. The criterion assessment of engagement and its drivers among all individuals, such as utilizing the work architecture provided in this study, is a vital initiation point for any active policy. It may be identified whether individual people, groups, job roles, or departments scoreless, median, or high on work engagement and its predecessor based on this evaluation, so we can identify where to focus interventions most effectively. Individuals, teams, and the company as a whole should be the target of interventions geared at leveraging the positive potential of work engagement.

Morgeson et al. [16] focused on the substance and structure of the occupations that people do, as well as the larger setting in which they work. This expanded focus on work design allows us to not only capture the breadth of study undertaken under the umbrella of job design, but also to increase our scope to include research that goes beyond what has traditionally been examined in the area. Because of space limits, we'll concentrate on studies from the I-O and organizational behaviour literatures, but readers should be aware that work design concerns have been explored by a variety of disciplines (e.g., industrial engineering, operations management, ergonomics). Further Murphy Hill et al. [17] designed survey for three companies to determine their rank (based on various factor) and drew the inferences about their productivity factors. Developers, managers, and executives may utilize their productivity factor rating to prioritize what is otherwise a plethora of investment possibilities in order to focus their efforts. In brief, earlier study has recognized a number of strategies for companies to boost software developer productivity, and their study has identified a mechanism to prioritize them.

3 Methodology

For the generation of our dataset, a survey has been taken based on various parameters which can influence the performance of the employee. There were more than 1200 employees who participated in the survey, and each employee was asked 26 questions, based on those questions they gave themselves a rating on a scale of 1 to 5. The interesting thing is that the rating is between 2 and 4, so no one gave themselves 1 nor 5. There were a few questions that were based on those parameters which can influence their performance. The questions were designed carefully, based on various researches, and were chosen such that it directly or indirectly impacts the overall performance of a person. Out of 32,374 data points, quite a few were missing, but they can be fixed by the measure of centrality. Some of the questions were Employee educational background, Marital Status, Employee Department, Business travel frequency, Distance from home, Employee environmental satisfaction, overtime, etc. Some of the attributes are object type, such as employee job role, in which there are different categories such as Sales Executive, Manager, Developer, Sales

Fig. 1 Proposed model architecture



Representative, Human Resources, Senior Developer. And other numerical attributes are either based on rating (which are between 1 to 5) and others are normal numerical data with their respective SI unit, for example, distance from home is in km. For training of the model, 75% of the total data points were taken and the rest 8,039 data points were taken for testing. The model is implemented considering the above 26 independent parameters and taking performance rating as a dependent parameter.

Figure 1 depicts the overall model architecture; the collection of data is explained in previous section. The implementation follows the above workflow. For data preparation, three things were done, 1. Data Cleaning, 2. Data Transformation, 3. Feature Engineering. Further data splitting is done to split the data for training and testing. After splitting, on the training data classical classification ML algorithms such as Logistic Regression (LR), K-Nearest Neighbour, Support Vector Machine (SVM), Decision Tree (DT), and some bagging and boosting techniques like Random Forest, XGBoost classifier, are applied. Further model is evaluated using various other performance matrices such as precision, accuracy, f1 score, recall.

3.1 Data Pre-processing

The work of changing data from one type to another useable data into the intended one, i.e., making it more useful and instructional. This whole operation could be automated. With the use of Machine Learning techniques, mathematical modelling, and statistical expertise. Graphs, videos, photos, and a variety of other formats can be

produced as a result of this entire process and machine’s requirements. The sample data consist of data acquired from a wide range of sources, which is merged logically to make a dataset. Dataset formats differ based on the use case. For example, a business dataset will be substantially different from a medical dataset. A medical dataset may contain data about healthcare, whereas a corporate dataset will contain critical industrial and business data.

As Python is the most widely used and recommended library among Data Scientists all over the world. Specific data pre-treatment activities can be performed using the specified Python libraries like NumPy, pandas, matplotlib. For example, if we want to import a CSV file, we can import it with the help of pandas by `pandas.read_csv()`. After loading the dataset, we have to check for any missing values, outlier detection, the shape of the data. This is done, as it influences the model while training. Outliers are sections of the dataset that deviate from the model’s predictions and it is something that doesn’t add to any learning model. Outliers have a negative impact on the model and must be deleted. To characterize data, descriptive statistics uses tools such as visualizations, central tendency measurement, and to estimate dispersion. It provides a meaningful summary of the information, allowing us to draw judgments see Fig. 2 which shows different descriptive statistical analyses of two of the attributes of the dataset.

The process of transforming information into a visual representation, such as a map or graph, to make data easier to grasp and extract important information is known as data visualization. The main goal of data visualization is to make it easier to see patterns, trends, and outliers in large data sets. The phrases data systems, statistical graphics, and information visualization are sometimes used interchangeably. With the help of data visualization, many things can be observed such as 1. the capacity to retain information quickly, develop new ideas, and make a better resolution; 2. a better view of which steps have to be done next to improve the business; 3. improved ability to hold the viewer’s attention by presenting facts which they can understand; 4. a basic technique of delivering information which increases the opportunity for

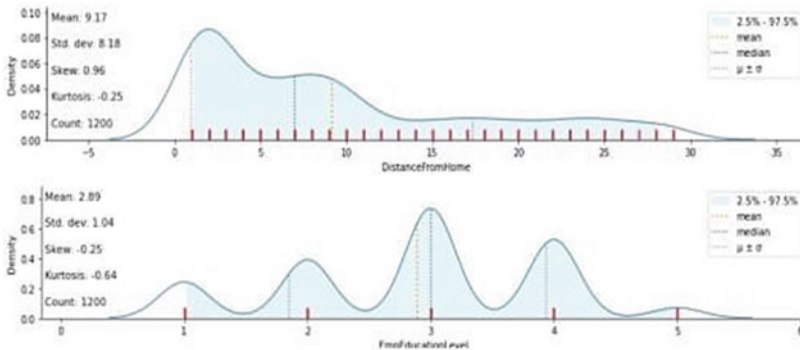


Fig. 2 Statistical analysis of distance from home and employee education label

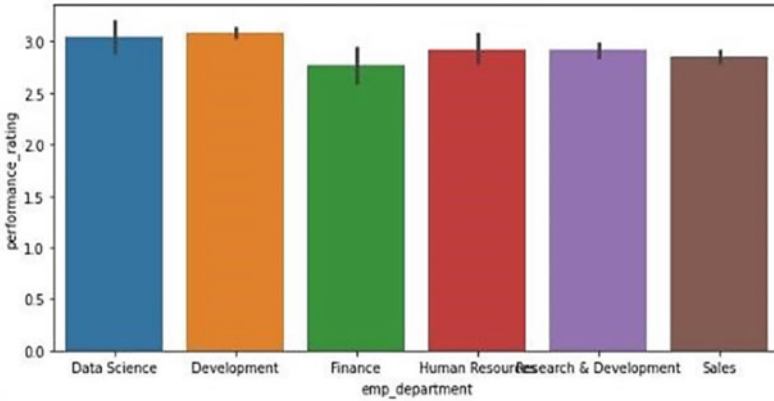


Fig. 3 Data visualization of different departments

all people involved to exchange ideas; 5. As data is more accessible and understandable, data scientists are no longer necessary; and 6. enhanced capacity to respond rapidly on insights and, as a result, achieve achievement with more speed and fewer errors. Figure 3. gives the spreading of data grouped by employee department and performance rating.

3.2 Model Deployment

As the performance rating is not a binary attribute, it is good to deploy some model rather than statistical analysis. We can do statistical analysis but it will give the statistical inference between two columns. While on the other hand if we deploy any algorithm, it will consider all the features at the same time and give the result in the context of employee rating.

For these datasets following algorithms have been deployed, i.e., Logistic Regression, SVM, Decision Tree, K-Nearest Neighbour, Random Forest, XG Boost.

3.3 Logistic Regression

The term “logistic regression” refers to an expansion of the term “linear regression”. Instead of modeling a proportion between the independent variable (X_1) and the probability of the result see Fig. 4, which would enable projected probabilities beyond the range of 0–1, the outcome is assumed to have a straightline relationship with the logit (which is the natural log of odds). The regression coefficients indicate the intercept b_0 and slope b_1 of this line.

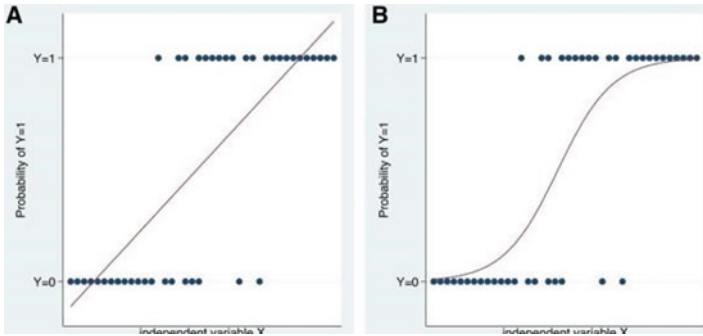


Fig. 4 Probability of an outcome

$$\frac{p}{\ln(1 - p)} = b_0 + b_1 X_1 \tag{1}$$

After solving this equation, the probability (P) has a sigmoidal relationship with the independent variable, and the predicted probabilities are now suitably between 0 and 1. Like linear regression, logistic regression can simply be expanded to cover several independent variables.

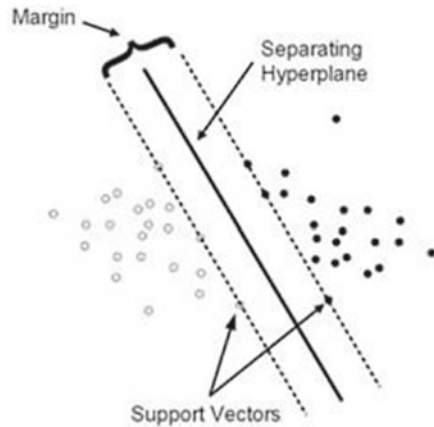
3.4 Support Vector Machine

The prominent state-of-art ML technique is the Support Vector Machine. Its primary function is classification. SVM is used to draw lines between classes and is based on the concept of calculating margins. The margins are designed such that there is an as little gap between the margin and the classes as feasible, which reduces the classification error see Fig. 5. Which is an example of an SVM operation.

3.5 Decision Tree

Decision trees are trees that group qualities by ordering them according to their values. The decision tree is mostly used for classification. Nodes and branches make up each tree. Each branch indicates a value that the node can take, and each node represents qualities in a group that needs to be categorized [18]. With the help of decision rules or decision trees, a decision tree algorithm is primarily used to construct a training/classification/regression model in the form of a tree structure (root, branch, and leaf), which is based on (inferred from) previous data to classify/predict class or target variables of future/new data. Decision trees can be used to analyze both numerical and categorical data.

Fig. 5 Working of SVM



3.6 *K-Nearest Neighbour*

The learner learns a specific pattern in instance-based learning. It tries to fit the freshly provided data into the same pattern. As a result, the term “instance-based” was coined. It’s a lazy learner who waits for test data to come before acting on it in conjunction with training data. The magnitude of the data raises the difficulty of the learning process. The k-nearest neighbor algorithm is an example of such an algorithm. The training data (which is labeled) is given to the learner in k nearest neighbor (or KNN). When the learner is given the test data, it compares the two sets of information. The training set yields k most connected data points. The majority of k is used, and the test data is assigned to a new class.

3.7 *Random Forest*

In both classification and regression problems, the subset of predictor variables used for dividing an internal node is determined by stated splitting criteria, which is considered an optimization problem. Entropy is the frequent dividing criteria in classification concerns since it is a practical implementation of Shannon et al. (2001) [26] source coding theorem that establishes the lower constraint on the length of a random variable’s bit representation. Each internal node of the decision tree is given with the following entropy equation:

$$E = -\sum_{i=1}^n p_i \log(p_i) \tag{2}$$

where n is the number of distinct classes and pi denotes the prior probability of each one. Overfitting is a drawback of decision trees, in which the model adheres too

closely to the eccentricities of the test dataset and performs poorly on a new dataset that is the test data. Overfitting decision trees will result in low general prediction accuracy, also known as generalization accuracy. Bagging or bootstrap aggregating is used to increase the accuracy and stability of a machine learning model. It may be used for regression as well as classification. Bagging also helps to minimize overfitting by reducing variance.

3.8 *XG Boost*

XGBoost is a decision-tree-based ensemble Machine Learning technique that uses a gradient boosting framework to produce a scalable, distributed gradient-boosted decision tree. Artificial neural networks surpass all known algorithms or frameworks in prediction challenges involving unstructured data (pictures, text, etc.). Decision tree-based algorithms, on the other hand, are presently rated best-in-class for small-to-medium structured/tabular data.

4 Results and Discussion

The result will be beneficial for any company or any developer/employee if they want to know their performance in the future. The questions asked in the survey deal with issues faced by the employee day-to-day. This result is showing which algorithm is performing well in their type of datasets. The training set is 75% of the total data points. The shape of the training dataset is (900,9) and the shape of the test dataset is (300,9). The above six machine learning algorithm is deployed on this dataset. The first model is Logistic Regression which gives the training accuracy of 82% and testing accuracy of 83%. Since the testing accuracy is higher than the training accuracy, it means that the model slightly over-fit itself. Some of the other algorithms are there which overfits themselves like a decision tree. The SVM gives the training accuracy of 84% and testing accuracy of 84%. As the accuracy is almost the same as LR but there is a significant change in other performance metrics. For the random forest training accuracy of 92.6% and testing accuracy of 93%. For XGBoost the training accuracy and the testing accuracy are almost the same as random forest. The performance matrices are given in Table 1. The model is performing multiclass predictions, in which the performance rating is taken as the dependent parameter and this parameter consists of three values i.e., 2.0, 3.0, 4.0. The model is trying to drive the relation between other independent parameters and the performance rating. With the help of these models, we can evaluate the performance of the employee/developer in the future, given these parameters.

Table 1. Performance matrices

	Accuracy		Recall			Precision			F1-Score		
	Training	Testing	0	1	2	0	1	2	0	1	2
Logistic Regression	0.82	0.83	0.46	0.93	0.74	0.64	0.86	0.77	0.54	0.89	0.75
Support Vector Machine	0.84	0.84	0.78	0.87	0.74	0.76	0.92	0.53	0.77	0.90	0.62
Decision Tree	0.90	0.91	0.83	0.95	0.74	0.83	0.94	0.80	0.83	0.94	0.77
Random Forest	0.92	0.93	0.91	0.95	0.74	0.91	0.95	0.77	0.91	0.95	0.75
XG Boost	0.92	0.93	0.91	0.95	0.78	0.91	0.95	0.78	0.91	0.95	0.78
K Nearest Neighbour	0.84	0.84	0.59	0.94	0.52	0.73	0.87	0.78	0.65	0.90	0.62

5 Conclusion

In this paper, we explore various parameters which can influence the performance of the employee. The model is implemented on 26 such parameters which directly or indirectly influence the performance, resulting in an overall decrease in the yield of the both company and the developers. A total of 1,200 members participated in the survey, based on their inputs, the further model is deployed. There are six classification models deployed, including ensemble-based algorithms and instance-based algorithms. For training of the model 75% of the total data points were taken and for testing remaining 25% were used. All the algorithms perform well after some data pre-processing and hyperparameter tuning. XGBoost performs well among all the algorithm, and give an accuracy of 92.6% on training data. For, the future work we have to gather more employee data so that the model can perform more robustly. We should perform statistical analysis for more incite between the attributes.

References

1. Hernaus T (2011) Business trends and tendencies in organization design and work design practice: identifying cause and effect relationships. *Bus Syst Res Int J Soc Adv Innov Res Econ* 2(1):4–16. Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016).
2. Chin WW (2010) How to write up and report PLS analyses. In: *Handbook of partial least squares*. Springer, Heidelberg, pp 655–690
3. Humphrey SE, Nahrgang JD, Morgeson FP (2007) Integrating motivational, social, and contextual work design features: a meta-analytic summary and theoretical extension of the work design literature. *J Appl Psychol* 92(5):1332

4. Boehm B, Abts C, Chulani S (2000) Software development cost estimation approaches—a survey. *Ann Softw Eng* 10(1):177–205
5. Perry DE, Staudenmayer NA, Votta LG (1994) People, organizations, and process improvement. *IEEE Softw* 11(4):36–45
6. Parker SK, Wall TD (2001) Work design: learning from the past and mapping a new terrain. In: *Handbook of industrial, work and organizational psychology*, vol 1, p 90109
7. Czarnecki K, Østerbye K, Völter M (2002) Generative programming. In: *European conference on object-oriented programming*. Springer, Heidelberg, pp 15–29
8. Morgeson FP, Humphrey SE (2006) The Work Design Questionnaire (WDQ): developing and validating a comprehensive measure for assessing job design and the nature of work. *J Appl Psychol* 91(6):1321
9. Patro CS (2013) The impact of employee engagement on organization's productivity. In: *2nd international conference on managing human resources at the workplace*, pp 13–14
10. Meyer AN, Fritz T, Murphy GC, Zimmermann T (2014) Software developers' perceptions of productivity. In: *Proceedings of the 22nd ACM SIGSOFT international symposium on foundations of software engineering*, p 1929
11. Palvalin M, Vuolle M, Jääskeläinen A, Laihonen H, Lönnqvist A (2015) SmartWoW—constructing a tool for knowledge work performance analysis. *Int J Prod Perform Manage*
12. Bok HS, Raman KS (2000) Software engineering productivity measurement using function points: a case study. *J Inf Technol* 15(1):7990
13. Ramírez YW, Nembhard DA (2004) Measuring knowledge worker productivity: a taxonomy. *J Intellect Cap*
14. Markos S, Sridevi MS (2010) Employee engagement: the key to improving performance. *Int J Bus Manage* 5(12):89
15. Bakker AB (2011) An evidence-based model of work engagement. *Curr Dir Psychol Sci* 20(4):265–269
16. Morgeson FP, Garza AS (2013) From individual work characteristics to work design configurations. In: *Work design symposium: comprehensive work design analysis—insights from around the globe*, Münster, Germany
17. Murphy-Hill E, Jaspán C, Sadowski C, Shepherd D, Phillips M, Winter C, Knight A, Smith E, Jorde M (2019) What predicts software developers' productivity? *IEEE Trans Softw Eng*
18. Sungheetha A, Sharma R (2020) A comparative machine learning study on IT sector edge nearer to working from home (WFH) contract category for improving productivity. *J Artif Intell* 2(04):217–225
19. Hakanen JJ, Perhoniemi R, Toppinen-Tanner S (2008) Positive gain spirals at work: From job resources to work engagement, personal initiative and work-unit innovativeness. *J Vocat Behav*

A Resilient and Efficient Protocol for Strengthening the Internet of Things Network Performance



Salma Rattal, Isabelle Lajoie, Omar Sefraoui, Kamal Ghoumid, Réda Yahiaoui, and El Miloud Ar-Reyouchi 

Abstract Reliable and efficient communication with a low rate of errors and packet loss is widely requested for the Internet of things (IoT) devices. This paper offers a resilient and efficient protocol (REP) to manage transmission status, recover and fault in IoT wireless networks. The network coding (NC) implementation is used to portray a realistic IoT scenario in which one sink node (SN) interacts with numerous monitored IoT device nodes over a wireless link between two separate networks. The proposed approach is compared and evaluated against two of the most extensively used error detection and correction code (EDCC) techniques available in the literature, namely the forward error correction (FEC) and the automated repeat request (ARQ). This analysis determined that the data throughput (kbps) and forwarding delay (ms) in terms of data length (bytes) and modulation rate (kbps), using REP, outperform the FEC and ARQ by more than 50% and 42%, respectively.

Keywords Internet of things · Error detection correction codes · Network coding

1 Introduction

The IoT system has quickly become the dominant big data contribution and distribution platform. However, the IoT today offers a best-effort service without any guarantee of quality. Packet loss, fluctuations in throughput, and delays will plague big data due to congestion and the diverse architecture of the IoT.

NC [1] is a fundamental combined approach that can enhance the IoT network in various contexts. As proven in [2], it promotes throughput, lowers forwarding delay,

S. Rattal · O. Sefraoui · K. Ghoumid
National School of Applied Sciences (ENSAO), Mohammed First University, Oujda, Morocco

I. Lajoie · R. Yahiaoui
NanoMedicine Lab, Therapeutic, Imagery, Franche-Comte University, Besançon, France

E. M. Ar-Reyouchi (✉)
Telecommunication Computer Science, Abdelmalek Essaadi University, Tétouan, Morocco
e-mail: e.arreyouchi@m.ieice.org

and offers other benefits to the network by improving network latency measurements [3], streamlining wireless communication [4] IoT devices, and providing other benefits to the network. The random linear NC (RLNC) [5, 6] is still considered currently as an effective NC procedure that enables network IoT nodes to generate coded packets in an arbitrary mixture of input origin data across a finite field. It is a new technology that can potentially improve the performance of many modern wireless communication systems. RLNC can speed up the Internet, enhance throughput over the IoT, shorten the sending time, and improve the reliability of cloud computing data centers. RLNC is also able to decode information and reduce the retransmission repetition that is required at endpoint nodes. A novel medical communication method for the efficiency of healthcare is described in the papers [7] and [8], using RLNC and exploring narrowband IoT (NB-IoT) technology for critical wireless mesh networks [9] applications. Both papers [10, 11] have proved that the round-trip time, overall network capacity, and delivery delay are much better than the state-of-the-art schemes, reducing the transmissions repetition and improving the round-trip time end-to-end probing cycle. The most frequently used methods for fixing and restoring lost packets in a wireless network may be divided into two categories. The authors analyze the EDCC scheme's performance FEC and ARQ for data transmission dependability [12] and [13]. A very minimal number of permitted retransmissions is taken into consideration. However, they are not extinguished and do not determine their performance.

The authors of [14] completely examine EDCCs for digital video broadcasting (DVB) to enhance information security. They make a significant contribution by elucidating the distinctions between the various EDCCs used in the three versions of DVB. These codes can promote flexibility of protocols and favor energy efficiency in the Internet of things, as in [15, 16], with reliable and accurate data on the current situation.

The authors of [17] analyze the link between packet length, modulation rate, packet forwarding time, and bit rate. Various modulation rates result from different limitations on transmitted packets. The transmission of data packets happens more quickly when the modulation rate increases; however, this comes at the receiver's sensitivity and the coverage area. In the IoT, various parameters, such as the packet delay [18] and the probing round [19], must be optimized. In addition, the reliability of communication systems for the IoT [20] is typically higher with lower modulation speeds.

We strongly recommend a network-based coding system as the best method for restoring packets through a single resend which increases data throughput (kbps) and forwarding delay (ms) in terms of data length (bytes) and modulation rate (kbps). In this context, each packet is delayed with a new format for the other network before the transmission. The performance of both separate networks depends on several parameters. We can cite the forwarding time, packet size, and modulation rate. This study demonstrates how NC can minimize the retransmission requests and replace or regain lost packets, improving IoT applications.

Network security is essential for securing data and preventing unwanted access to systems. The RLNC can enforce security regulations by preventing assaults on the

planned healthcare network. This method may verify that all devices on a network are secure from viruses and malware. A lightweight authentication-based secure methodology [21] is also available to improve the model's security. The IoT devices are interconnected via a mesh network in Routing Protocol, as employed in our case. Although it is created with encryption security to safeguard messages, it is prone to selfish conduct and internal assaults. Some studies have been conducted to overcome this issue. The authors of [22] suggest an innovative trustworthiness approach based on metrics for incorporating trust evaluation, hence improving the resilience of the security system. The primary goal of this research is to significantly and quickly cut down on the number of retransmissions needed to fix possible errors, which will make the device more efficient.

This article's introduction is the first of six sections. Section 2 goes through the EDCCs. Section 3 describes the problem and suggests a solution based on the system model. Section 4 describes the procedure. The findings are presented in Sect. 5 of the paper. The conclusions are established in the last Sect. 6.

2 Review of EDCC and the Proposal

Consider the two primary methods for recovering error packets in FEC and ARQ networks. The FEC approach allows IoT devices [23] to provide an excellent framework for guaranteeing a reliable wireless connection. FEC is commonly employed in broadcasting and Rayleigh fading channels. It also discovers and repairs problems without needing a reverse channel. It is the most common and is often used extensively in IoT and wirelessly telecommunication [24]. The FEC may recover lost bits when reversible or irreversible connections are allowed.

But again, FEC is not just a solution needed for the erroneous control scheme. As explained in [25], it is frequently used in conjunction with another approach, such as ARQ. Relatively insignificant errors are corrected without retransmission, whereas significant errors require retransmission. Between the two most important FEC classes, both convolutional codes and block codes are used. Because FEC and ACK cut down on network capacity, they do not work well with many people.

For each packet, the ARQ protocol uses a verification acknowledgment (ACK) [13] and a negative-acknowledgment (NAK) [12] to retransmit dropped and missing packets in a cyclic fashion. This approach is included in the suggested protocol to identify and correct faults. It is not a packet combination but rather a collection of fragments from the original packets. Before the native message is sent, it is essential to have a solid understanding that the FEC is predicated on the idea of incorporating managed duplication into the delivery message. Rather than retransmitting the data, the redundancy is used to recover any lost, dropped, or missing packets at the destination.

In [26], the authors propose a novel approach that incorporates word interleaving, FEC, and ARQ to lessen the likelihood of errors and data loss occurring in apps that use wireless Internet. However, in [27], the authors take a more in-depth look at IoT,

highlighting various protocols and concerns. The two most important parameters for improving IoT network performance are data throughput and forwarding delay. Several retransmissions are required to send the data when reception conditions deteriorate properly. They have a substantial impact on latency and throughput. NC has shown again that it is more than double the performance [28]. The proposed REP allows IoT devices to deliver reliable service for important applications like SCADA [29], sustainable energy diagnostics, oil, and gas pipelines, and other smart critical IoT uses. The REP is also applied to simplify an overview of IoT device network performance, more precisely, the correct receipt of the data packet by the IoT devices. The specifications for all IoT devices on both networks, networks 1 and 2, are the same. As in this paper, network performance can be more objective and systematic if important parameters are measured.

Furthermore, they provide the proposed protocols with a fast and easy idea of improving the IoT device communication reliability, including the repetition of transmission and retransmission, data throughput, forwarding delay, data length, modulation type, and payload bitrate.

3 The Statement of the Problem and the Solution

This part discusses the challenges of designing a viable packet loss repair solution regarding the IoT.

3.1 *Issue Statement of Correcting Codes*

FEC dramatically minimizes the erroneous packets. In theory, FEC should make it easier for IoT devices to discover and remediate errors. without having to send them again, making data transmission more efficient. It does, however, need a long EDCC because it transmits the code to the IoT system even if it is not correct, making things more complicated in low-error situations. FEC has also had a limited coding rate, which means it cannot utilize the channel efficiently; it cannot eliminate or minimize jitter except for out-of-order packets, which are frequent in the Network connection. The FEC protocol needs a more considerable bandwidth cost than other methods.

ARQ demands employing a reversing channel to broadcast ACKs and NAKs; consequently, it leads to poor signal conditions, which in turn result in slow transfer rates and drive delay changes caused by retransmitted data. Additionally, since a preset algorithm must process the duplicated information in the transmission, it exerts a larger computing strain on the receiving device. The data throughput or the average latency while transmitting a single packet between two IoT nodes may need to be decreased to make room for improvement.

We carry out an analysis of how well the suggested methodology works and the effectiveness of REP compared to the most current and cutting-edge protocols. These

protocols include the three most common forms of packet-loss recovery algorithms, namely ARQ and FEC.

3.2 *Proposed Remedy*

RNC [30] solves the problem of regaining lost and damaged packets, which reduces the number of transmissions. It has the ability to bring about a major improvement in application performance. Nonetheless, in reality, and especially in actual application, the REP approach is not as simple and traditional as it first seems. It is practicable and lucrative to investigate the IoT protocol intended to be based on RNC. Its whole performance becomes more powerful than RNC's utilization of its functional potential. The RNC technique can bring many useful and workable improvements to wireless networks. These improvements can include a decrease in the number of times a packet needs to be resent, the ability to find and fix missing or lost packets, a decrease in latency, and a stronger network topology. Thanks to the recommended protocol, the information delivered by the packet to the destination are protected from transmission errors. The message may add more data if a packet has a defined length. This procedure guarantees that the packet is always the right size.

3.3 *Presentation of System Model*

Assume a monitoring architecture of the system in which IoT devices are wirelessly connected and communicated by the *SN*, as depicted in Fig. 1. The scenario suggests that the *SN* wants to send packets from network 1 to network 2, as seen in Fig. 1. The *SN* also acts as a bridge between networks 1 and 2, linking the two. We are operating under the assumption that the *SN* forwards packets to the second network and that each IoT device on the second network can only take one packet at a time.

This monitoring prototype system delivers correct information about renewable energy's essential circumstances [31, 32]. This system model contains an Internet-connected computer that can collect and analyze data and communicate with all the IoT devices and the sink node. It is necessary for two networks separated by obstacles or cannot communicate with each other for various reasons.

For example, two groups of solar panels or two hospitals have their wireless network. The monitoring stations in areas 1 and 2 are divided into two wireless networks, each annotated with n IoT devices such as $N_{1,1}, N_{1,2}, \dots, N_{1,n}$, and $N_{2,1}, N_{2,2}, \dots, N_{2,n}$ respectively, as shown in Fig. 1.

The nodes $N_{i,j}$, $i = 1$ and $j = 1, 2, 3, \dots, n$ send their information directly to the *SN*.

In summary, the research addresses data collection by the *SN*, which uses RNC to transfer data to IoT devices $N_{2,1}, N_{2,2}, \dots, N_{2,n}$ which are presented in wireless network 2. Assume each node in network 1 sends one packet, which is

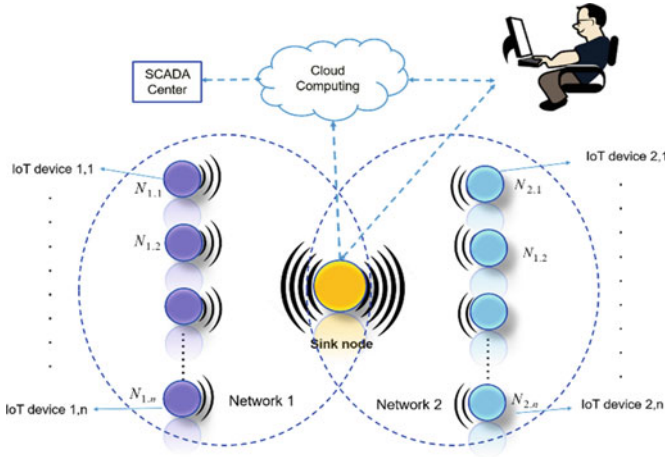


Fig. 1 The system model has two separate networks in which the IoT devices communicate through one sink node

$p_{1,1}, p_{1,2}, \dots, p_{1,n}$. A SN can appropriately gather the whole of all transmission packets from $N_{1,1}, N_{1,2}, \dots, N_{1,n}$ in a single step. IoT SN has a coverage area incorporating both sets of nodes $(N_{1,1}, N_{1,2}, \dots, N_{1,n})$ and $(N_{2,1}, N_{2,2}, \dots, N_{2,n})$.

To evaluate the success of the proposed protocol in error-free transmission, keep in mind that a protocol’s reliability ($P_{prot(er=0)}$) is the ratio of suitable packets transmitted (P_a) to the data packets transferred n : $P_{prot(er=0)} = P_a/n$.

Consider the probability of wrong packet transmission, denoted by the notation $(1 - P_e)$, and corresponds to the probability that a packet is successfully delivered. For n packets involved in the transmission, the chances of all n of those packets being accurately sent are as follows:

$$P_{prob} = (1 - P_e)^n \tag{1}$$

The authors in [1] compare the effectiveness of the method in the event of a failure to the absence of an error:

$$P_{prot(er \neq 0)} = P_{prot(er=0)} \times (1 - P_e)^n \tag{2}$$

The envisaged model could be considered a communication reference used in different contexts where nodes can communicate with each other without error. When the nodes number is very numerous and condensed, the protocol must be improved and requires an update to eliminate noise in the channel.

4 Protocol Description

4.1 A Specific Protocol Presentation

For the rest of the description of the proposed protocol, we will continue with the topology depicted in Fig. 1.

The SN intends to send the n packets that originate from $N_{1,1}, N_{1,2}, \dots, N_{1,n}$, denoted, $p_{1,1}, p_{1,2}, \dots, p_{1,n}$ to the destinations, indicated by $N_{2,1}, N_{2,2}, \dots, N_{2,n}$. The $N_{2,1}, N_{2,2}, \dots, N_{2,n}$ are located inside the SN's field of coverage and will continue to collect the packets $p_{1,1}, p_{1,2}, \dots, p_{1,n}$ if they are in the field of coverage.

The packet $p_{1,1}$, is distributed by the SN, however, just one IoT device node $N_{2,1}$ accepts it. The SN transmits the packet with the address $p_{1,2}$, however, the IoT device node with the identifier $N_{2,2}$ is the only one that can successfully collect it.

This operation is repeated until the transmission of all n packets has been completed their transmission. Finally, The SN transmits the last packet $p_{1,n}$, and $N_{2,n}$ satisfactorily receives it.

As a result, every IoT device gets its corresponding packet. However, the inappropriate propagation conditions often request additional transmissions to deliver the data effectively. This phenomenon, in turn, leads to a significant increase in latency as well as an important rise in throughput. Common ARQ methods [33] use a store-and-forward methodology consisting of one acknowledgment for each packet, rendering traditional approaches unsuitable for failed packet restoration and energy conservation from sink node to moving IoT device. Furthermore, the retransmission process can detect and recover missing packets.

According to the hypothetical situation shown in Fig. 1, the SN retransmits packets. In the circumstance that these retransmissions are received successfully, $N_{2,1}$ is going to get the packet that was lost ~~$p_{1,1}, p_{1,2}, p_{1,3}, p_{1,4}, \dots, p_{1,n}$~~ . IoT $N_{2,2}$ will receive the missing packet $p_{1,1}, p_{1,2}, p_{1,3}, p_{1,4}, \dots, p_{1,n}$. IoT $N_{2,3}$ is going to get the packet that went missing ~~$p_{1,1}, p_{1,2}, p_{1,3}, p_{1,4}, \dots, p_{1,n}$~~ . And so on until $N_{2,n}$ the missing packet ~~$p_{1,1}, p_{1,2}, p_{1,3}, p_{1,4}, \dots, p_{1,n}$~~ will be received at the end of the process.

Therefore, regardless of whether or not these $(n - 1)$ retransmissions are successful, we need a maximum of $(n - 1) \times n$ packet forwarding and n transmissions for all IoT $N_{2,j}, j = 1, 2, 3, \dots, n$ in order to successfully receive n packets. This means that we need $n \times n$ transmissions and retransmissions.

In the case of forwarding error checking, the duplication enables the $N_{2,j}, j = 1, 2, 3, \dots, n$ in on a particular number of errors that could occur anywhere inside the message packets and make the necessary corrections to these issues without having to resort to retransmission procedures. FEC'' comes from the fact that the code will be sent to the receiver regardless of whether or not it is valid.

The EDCC and the process of fixing mistake patterns must take significant time and effort. Because of this, FEC can lower packet loss while simultaneously raising delay and bandwidth.

Rather than retransmit multiple packets $p_{1.1}, p_{1.2}, p_{1.3}, \dots, p_{1.n}$ in the proposed protocol, the SN retransmits one coded packet (a packet combination format), which takes the following general format $p_{1.1} \oplus p_{1.2} \oplus p_{1.3}, \dots, \oplus p_{1.n}$.

IoT $N_{2.1}$ can decode packets $p_{1.1}, p_{1.2}, p_{1.3}, \dots, p_{1.n}$ upon the receipt of this packet, by utilizing the a binary operation XOR [34], in between the packets $p_{1.j}$, $j = 1, 2, 3, \dots, n$, that it has already been collected via the encrypted packet. Thus, to get the n original packets of the IoT devices, the $N_{2.1}$ decodes $p_{1.j}$, $j \neq 1$, because it previously received the coded packet, it applies the XOR operator between the packets $p_{1.2}$ that it obtains by:

$$p_{1.j} = p_{1.1} \oplus (p_{1.1} \oplus p_{1.2} \oplus p_{1.3}, \dots, \oplus p_{1.n}), \text{ where } j \neq 1 \quad (3)$$

Equally, the node $N_{2.2}$ decodes $p_{1.j}$, $j \neq 1$ since:

$$p_{1.j} = p_{1.2} \oplus (p_{1.1} \oplus p_{1.3}, \dots, \oplus p_{1.n}), \text{ where } j \neq 2 \quad (4)$$

To combine formerly received packets, we employ the XOR operation. $p_{1.2}$ with the coded packet. And next, $N_{2.n}$ decodes $p_{1.j}$, $j \neq n$ by XORing the packets $p_{1.n}$ that it's already arrived and been correctly obtained over through the coded packet.

$$p_{1.j} = p_{1.n} \oplus (p_{1.1} \oplus p_{1.3}, \dots, \oplus p_{1.n}), \text{ where } j \neq n \quad (5)$$

As a consequence, we are able to recover any lost packets using only $n + 1$ complete transmissions. This example demonstrates the advantages of employing NC in a single-hop architecture.

Communications were minimized from $(n \times n)$ to $(n + 1)$. As a result, many IoT devices' repetition of transmissions and retransmissions is considerably decreased. Using the abovementioned technique, we can considerably increase several possible network benefits, including attaining network capacity, lowering latency, and boosting network dynamics resilience. Suppose we want to send n packets without errors, and each packet will be retransmitted appropriately.

4.2 Simulation Parameters of System Model

The implemented model presented in Fig. 1 divides the SN from the IoT receivers and transmitters using variable Kbit/s links (from 62.83 to 180.89 Kbps). The packet size changes again (data rises from 0 to 1000 bytes), and each link can incur a propagation and processing delay of 20 microseconds sequentially.

Assume that the SN starts direct forwarding after receiving the packet's ending bit and that the queues are vacant.

Table 1 contains a listing of the many parameters that were utilized throughout the simulation.

Table 1 Interpretation analytical of the simulation parameters

Specifications	Values
Size of data (bytes)	We have a numerical range ranging from 0 to 1000
modulation rates (kbps)	62.83 kbps to 180.89
FEC technique	(On = 3/4)/Off
ARQ mechanism	On/Off
Processing delay (ms)	20

As with IoT wireless transmissions, it is important to remember that packet loss is possible and to presume that retransmission is initiated whenever a message is not correctly sent. This process is carried out repeatedly until the transmission is completed successfully. In this particular scenario, we can also presume that the succeeding transmissions are independent of one another.

In most cases, the direct paths that are exploited between the IoT devices of the first network and the IoT devices of the second network (via the SN) (also known as line-of-sight, or LOS) are available; as a result, the Rice distribution is an excellent choice. Matlab is the perfect tool, and this article uses it to optimize a proposed method and define the REP objective function.

5 Results

According to Subsect. 4.1, the transmission repetition inside the wireless network can be drastically decreased; as a result, lost packets are recovered with the fewest retransmissions. In Fig. 2, the results without EDCC schemes show that data throughput improves as data length grows. Without EDCC methods, data throughput increases considerably from 40 to 140 kbps for data lengths up to 400 bytes and then slightly from 141 to 158 kbps for data lengths up to 1000 bytes. Figure 2 also compares the proposed REP to two types of EDCCs, FEC, and ARQ.

It can be seen that data throughput sans correction outperforms ARQ and FEC efficiency. Furthermore, the ARQ method has a higher data throughput than the FEC technique. As a result, FEC and ARQ dramatically minimize the chance of unrecoverable transmission errors in the data flow. In addition, Fig. 2 also shows a significant improvement in the recommended REP. Table 2 presents the data throughput findings for various data lengths with and without EDCCs compared to the proposed REP.

The proposed REP protocol’s ARQ and FEC techniques are contrasted in Table 2. According to Table 2 and Fig. 2, the REP protocol collects data at a rate significantly higher than the ARQ and FEC methods. Additionally, we see that, as compared to the ARQ and FEC protocols, data throughput gradually decreases without EDCC. The simulation results thus show that the proposed protocol outperforms both ARQ and FEC.

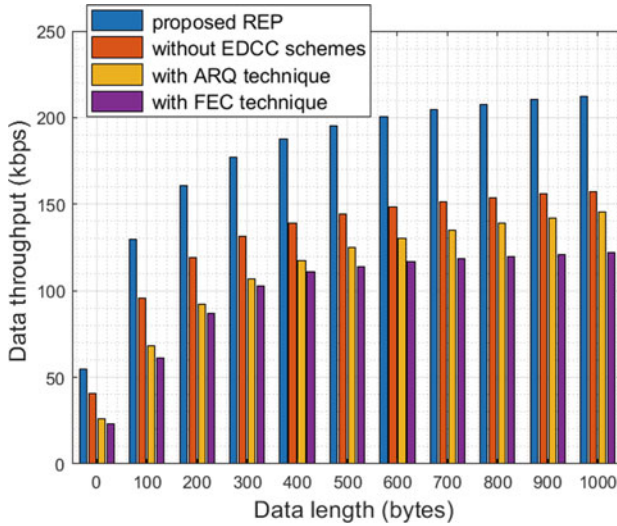


Fig. 2 The effect of data length on data throughput utilizing various error correction techniques

Table 2 Data throughput for different data lengths

Data length	100	500	1000
Proposed REP	129.60	195.17	212.40
without EDCCs	96.00	144.57	157.34
ARQ technique	68.41	124.74	142.47
FEC technique	61.04	114.13	122.14

Together with the proposed method, the most common methods are contrasted in Fig. 3, which plots forward latency versus data length.

The comparisons presented in Fig. 3, demonstrate that the forwarding delay grows when the data length of IoT device nodes rises when error-correcting codes are used. The data length affects the rate at which error corrector codes are generated in IoT devices, leading that rate to grow.

Table 3 summarizes the forwarding delay (ms) vs. data length results for different EDCCs using a data length of 1000.

In Fig. 3, the MATLAB simulation results are compared to the study results for FEC and ARQ protocols. The proposed REP protocol’s forwarding delay performance has significantly improved. For data length is 1000 bytes, the improved REP derived from Fig. 3 is at least 2.01 and 1.72 times faster than FEC and ARQ, respectively.

Figure 4 shows that using different modulation rates in wireless networks, with or without EDCC schemes, significantly affects data throughput.

The results show that the AQR increases the data throughput and reduces the time it takes to send data more than FEC. In addition, the proposed REP has more advantages

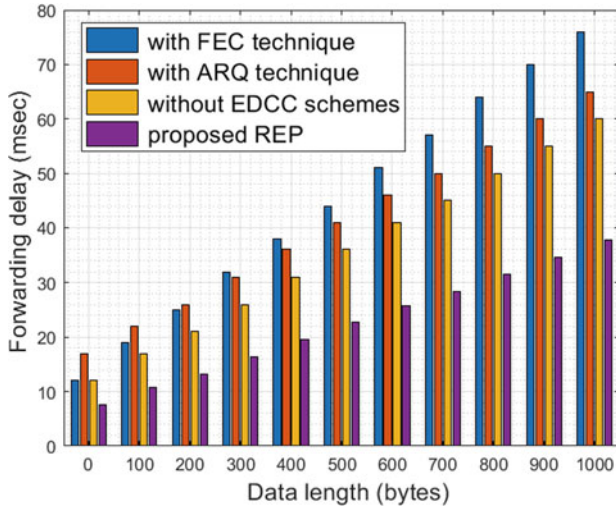
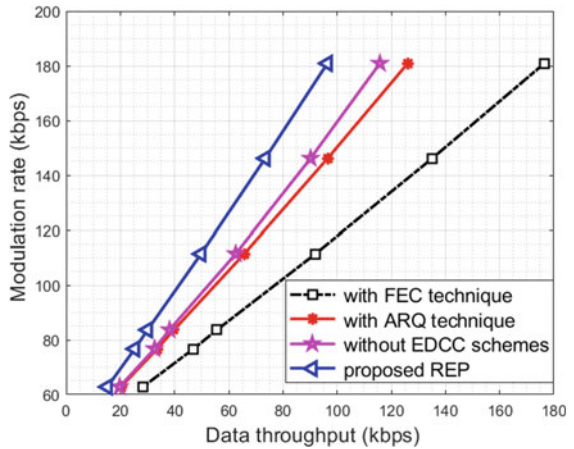


Fig. 3 Comparison between principals error correction techniques with the proposed algorithm REP forwarding delay (ms) versus data length (bytes)

Table 3 Comparison between the REP and different EDCCs

Data length	1000 Bytes		
Techniques	FEC	ARQ	REP
Forwarding delay (ms)	76.01	65.02	37.80

Fig. 4 Modulation rate (kbps) vs. Data throughput (kbps) is compared between the leading EDCCs and the proposed REP



of improving the throughput than the existing EDCCs. Moreover, according to Fig. 4, it should be mentioned that the results show that the EDCC schemes promote a further decrease in data throughput.

In addition, without any error correction, it makes them better but does not make them more resistant to link failures with errors in wireless network communication problems. Also, without any error correction, it makes them better but doesn't make them more resistant to link failures with errors in wireless network communication problems.

Figure 5 shows the evolution of the forwarding delay during the repartitioning process for the FEC and ARQ techniques compared to REP when using different modulation rates.

For the modulation rate (83.67/180.89 kbps), the comparison of forwarding delay (ms) between REP and the most critical EDCCs are summarized in Table 4.

Table 4 shows the percentage improvement in REP generated from Fig. 5 is greater than 52.8% and 31.4%, respectively, for modulation rate = 83.67 kbps, and it can forward 50.6% to 21.65% more quickly, respectively, for modulation rate = 146.17 kbps. According to the simulation results, the REP appears to be a powerful and efficient protocol for improving the performance of two wireless networks connected by a cooperative sink node. Consequently, this statement is supported by Table 4,

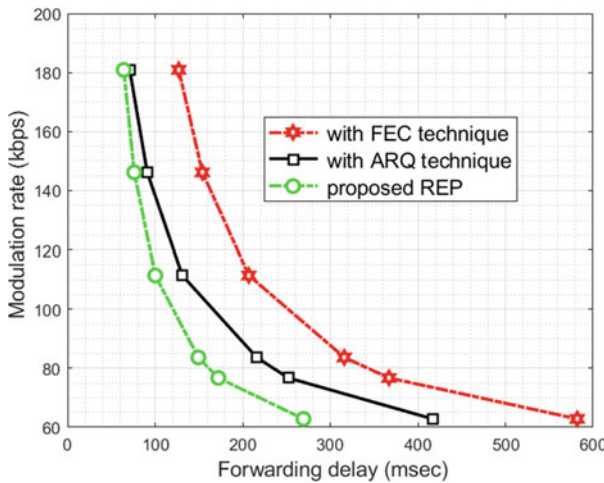


Fig. 5 Modulation rate (kbps) vs. Forwarding delay (msec): a comparison between the main EDCCs and the proposed REP

Table 4 Comparison simulation results (Modulation rate = 83.67 and 146.17 kbps)

Rate of modulation	83.67 kbps			146.17 kbps		
EDCCs techniques	FEC	ARQ	REP	FEC	ARQ	REP
Forwarding delay (ms)	316	217	149	154	97	76

which shows that REP, obtained from Fig. 5, has a better percentage improvement than 52.8% and 31.34%, respectively, for modulation rates of 83.67 and 146.17 kbps, compared to FEC and ARQ.

Based on the simulation findings, it appears that the proposed REP is an advanced and useful protocol for detecting and recovering errors while improving the data throughput and the delay in forwarding. When we employ a radio modem as an SN that relays data from various medical IoT devices, we can demonstrate that the observed findings fully correspond with those predicted by the simulation.

6 Conclusion

This article introduces a REP to detect and recover lost packets. The proposed protocol can also be used effectively to improve wireless network IoT device performance, increasing the data throughput and decreasing the forwarding delay more than most existing error correction techniques. The proposed REP provides significant improvements; it is quicker because it requires less retransmission, saves bandwidth and energy, increases data throughput, and reduces latency. The findings show that the suggested protocol outperforms the ARQ and FEC approaches when employing NC. According to the research findings, the proposed REP can considerably enhance data transfer quality, recover lost packet data, and sharply reduce the number of transmissions and retransmissions. It can also transmit packets between IoT devices on two distinct wireless networks, indicating high efficiency in IoT networks. For future work, it will be important to use this protocol to improve communication in remote sensor networks for healthcare.

References

1. Ahlswede R, Cai N, Li S-YR, Yeung RW (2000) Network information flow. *IEEE Trans Inf Theory* IT-46(4):1204–1216
2. Ho T, Koetter R, Médard M, Karger DR, Effros M (2003) The benefits of coding over routing in a randomized setting. In: *Proceedings of IEEE I. Symposium on information theory*
3. Ar-reyouchi EM, Hammouti M, Maslouhi I, Ghomid K (2017) The Internet of things: network delay improvement using network coding. In: *ICC 2017 proceedings of the second international conference on internet of things, data, and cloud computing*, Cambridge, United Kingdom. ACM
4. Hammouti M, Ar-reyouchi EM, Ghomid K, Lichioui A (2016) Clustering analysis of wireless sensor network based on network coding with low density parity check. *Int J Adv Comput Sci Appl (IJACSA)* 7(3):137–143
5. Ho T, Koetter R, Médard M, Karger DR, Effros M, Shi J, Leong B (2006) A random linear network coding approach to multicast. *IEEE Trans Inf Theory* 52(10):4413–4430
6. Barekatin B, Khezrimotlagh D, Maarof MA, Quintana AA, Cabrera AT (2016) GAZELLE: an enhanced random network coding based framework for efficient P2P live video streaming over hybrid WMNs. *Wirel Pers Commun* 95(3):2485–2505

7. Ar-Reyouchi EM, Ghoumid K, Ar-Reyouchi D, Rattal S, Yahiaoui R, Elmazria O (2021) Protocol wireless medical sensor networks in IoT for the efficiency of healthcare. *IEEE Internet Things J*. <https://doi.org/10.1109/JIOT.2021.3125886>
8. Ar-Reyouchi EM, Ghoumid K, Ar-Reyouchi D, Rattal S, Yahiaoui R, Elmazria O (2021) An accelerated end-to-end probing protocol for narrowband IoT medical devices. *IEEE Access* 9:34131–34141. <https://doi.org/10.1109/ACCESS.2021.3061257>
9. Rattal S, Ar-Reyouchi EM (2019) An effective practical method for narrowband wireless mesh networks performance. *SN Appl Sci* 1:1532. <https://doi.org/10.1007/s42452-019-1595-9>
10. Vaze R, Iyer S (2019) Capacity of cellular wireless networks. *IEEE Trans Wireless Commun* 18(3):1490–1503. <https://doi.org/10.1109/TWC.2018.2890666>
11. Ar-Reyouchi EM, Lamrani Y, Benchaib I, Ghoumid K, Rattal S (2020) The total network capacity of wireless mesh networks for IoT applications. *Int J Interact Mob Technol (IJIM)* 14(8):61–75
12. Lin S, Costello DJ (1983) Error control coding: fundamentals and applications, chap 15. Prentice-Hall, Upper Saddle River
13. Kotuliaková K, Šimlaščíková D, Polec J (2011) Analysis of ARQ schemes. *Telecommun Syst* 52(3):1677–1682
14. Ar-Reyouchi EM, Rattal S, Ghoumid K (2022) A survey on error-correcting codes for digital video broadcasting. *SN Comput Sci* 3:105. <https://doi.org/10.1007/s42979-021-00994-x>
15. Faheem M, Butt RA, Raza B, et al (2019) Energy efficient and reliable data gathering using internet of software-defined mobile sinks for WSNs-based smart grid applications. *Comput Stand Interf* 6666:2–18
16. Macher G, Diwold K, Veledar O, Armengaud E, Römer K (2019) The quest for infrastructures and engineering methods enabling highly dynamic autonomous systems. In: Walker A, O'Connor R, Messnarz R (eds) *Systems, software and services process improvement, EuroSPI 2019. Communications in computer and information science*, vol 1060. Springer, Cham. https://doi.org/10.1007/978-3-030-28005-5_2
17. Chatei Y, Ghoumid K, Ar-reyouchi EM (2017) Narrowband channel spacing frequencies metric in one-hop wireless mesh networks. In: 2nd international conference on communication and electronics systems (ICCES), Coimbatore, India, October 2017
18. Ghasempour A, Moon TK (2016) Optimizing the number of collectors in machine-to-machine advanced metering infrastructure architecture for Internet of Things-based smart grid. In: *IEEE green technologies conference*, Kansas City, pp 51–55
19. Ar-Reyouchi EM, Maslouhi I, Ghoumid K (2020) A new fast polling algorithm in wireless mesh network for narrowband Internet of Things. *Telecommun Syst* 74:405–410. <https://doi.org/10.1007/s11235-020-00671-z>
20. Al-Sarawi S, Anbar M, Kamal Alieyan K, Alzubaidi M (2017) Internet of things (IoT) communication protocols: review. In: 8th international conference on information technology (ICIT), pp 685–690
21. Karrupusamy P (2021) Advanced metering infrastructure with secure chord lookup protocol for IoT Systems. *J Electric Eng Autom* 2(3):112–117
22. Smys S, Vijesh Joe C (2021) Metric routing protocol for detecting untrustworthy nodes for packet transmission. *J Inf Technol* 3(02):67–76
23. Alabady SA, Salleh FM, Al-Turjman F (2018) LCPC error correction code for IoT applications. *Sustain Cities Soc* 42:663–673
24. Omoniwa B, Hussain R, Javed MA, Bouk SH, Malik SA (2019) Fog/edge computing-based IoT (FECIoT): architecture, applications, and research issues. *IEEE Internet Things J* 6(3):4118–4149
25. Bada AB (2017) Automatic repeat request (ARQ) protocols. *Int J Eng Sci (IJES)* 6:64–66
26. Chen D, Rong B, Shayan N, Bennani M, Cabral J, Kadoch M, Elhakeem AK (2004) Interleaved FEC/ARQ coding for QoS multicast over the Internet. *Can J Electr Comput Eng* 29(3):159–166. 25
27. Verma N, Singh S, Prasad D (2022) A review on existing IoT architecture and communication protocols used in healthcare monitoring system. *J Inst Eng India Ser B* 103:245–257

28. Vieira Luiz Filipe M, Mario G, Archan M (2013) Fundamental limits on end-to-end throughput of network coding in multi-rate and multicast wireless networks. *Comput Netw* 57(17):3267–3275
29. Rezaei A, Keshavarzi P, Moravej Z (2017) Key management issue in SCADA networks: a review. *J Eng Sci Technol* 20(1):354–363
30. Ar-Reyouchi EM, Lichioui A, Rattal S (2019) A group cooperative coding model for dense wireless networks. (*IJACSA*) *Int J Adv Comput Sci Appl* 10(7):367–373
31. Hammouti M, Ar-reyouchi EM, Ghoumid K (2016) Power quality command and control systems in wireless renewable energy networks. In: *IEEE explored renewable and sustainable energy conference (IRSEC)*. International IEEE Xplore Digital Library
32. Li J, Zhou Y, Liu Y, Lamont L (2012) Performance analysis of multichannel radio link control in MIMO systems. In: *Simplot-Ryl D, Dias de Amorim M, Giordano S, Helmy A (eds) Ad hoc networks, ADHOCNETS 2011*. Lecture notes of the institute for computer sciences, social informatics and telecommunications engineering, vol 89. Springer, Heidelberg
33. Maslouhi I, Ar-reyouchi EM, Ghoumid K, Baibai K (2018) Analysis of end-to-end packet delay for internet of things in wireless communications. *Int J Adv Comput Sci Appl (IJACSA)* 9(9):338–343
34. Lee KH, Kim JH, Cho S (2014) RLNC in practical wireless networks. In: *Cai Z, Wang C, Cheng S, Wang H, Gao H (eds) Wireless algorithms, systems, and applications, WASA 2014*. Lecture notes in computer science, vol 8491. Springer, Cham

Early Identification of Crop Disease Using Deep Convolution Neural Networks



J. Vakula Rani and Aishwarya Jakka

Abstract In India, the country's economy is mainly dependent on agriculture, and it gets highly affected if there is not enough agricultural produce or the quality of the agrarian products is poor. The image-based automatic diagnostic tool can ensure rapid and cost-effective detection of such problems. Therefore, the research aims to demonstrate various Deep Convolution Neural Network (DCNN) models and thereby apply them for the early identification of crop disease and help farmers get the predicted insights through these models. This will help the farmers detect the potential risks and find a solution to mitigate them early, reducing their loss. We have implemented DCNN models VGG16, VGG19, InceptionV3, and ResNet50 to identify the plant diseases mainly caused by pests, pathogens, nutrient deficiencies, etc. These models trained and test accuracies are compared with a Convolution Neural Network (CNN). The experimental results show an outstanding test accuracy of 97.2% for VGG19 and InceptionV3, whereas CNN stands at 94.7% accuracy.

Keywords Agriculture · Deep convolution neural network (DCNN) · InceptionV3 · ResNet50 · VGG16 · VGG19

1 Introduction

Agribusiness is the primary source of income and significantly impacts the global economy. Sustainable agriculture is the key to ensuring environmental, social, and economic concerns in agriculture. Ensuring food safety and eliminating food shortages are critical for the increasing population [1]. Based on the estimates, food production globally should be increased by 60–110% to feed around 9–10 billion

J. Vakula Rani (✉)

Department of MCA, CMR Institute of Technology, Bengaluru, India
e-mail: vakula.r@cmrit.ac.in

A. Jakka

University of Pittsburgh, Pittsburgh, USA
e-mail: aishwaryajakka@pitt.edu

people by 2050. Hence there should be a fundamental shift to the sustainable agriculture model from the agricultural yield model. Crop production depends on numerous climatic factors such as weather conditions, irrigation, soil condition, cultivation, temperature, and rainfall. The past information on crop yield is essential in supply chain operations. Modern technology made precision agriculture possible, allowing farmers to gather data from various sources and make better crop production decisions [2]. The image-based automatic diagnostic tools assist farmers in making more appropriate and comprehensive decisions to decrease risk and financial loss. Recent technologies to detect crop disease early can reduce the need for direct human involvement in plant protection. For early disease identification, several NN approaches have been used [2]. The use of Deep Learning techniques in farming will help cultivators by enabling them to manage their crops more efficiently. The deep learning techniques can be applied at different phases of crop management, planting the seeds, growth and development of the crop, and crop harvest, storage, and distribution. Disease detection is one of the major threats to crop yield and affects the growth of the plant. Plant diseases have a significant impact on crop yield. Thus, identifying conditions early and taking appropriate measures is essential [13]. An automated system is necessary for the timely identification and proper detection of plant diseases. There are many supervised and unsupervised algorithms used for this purpose. Deep learning (DL) methods have recently boomed in agriculture and provide more accurate and precise predictions than traditional methods. For detecting and classifying plant disease, color analysis and thresholding [3] of digital image processing techniques and artificial neural networks (ANNs) techniques are most used. The hybrid approach with different image preprocessing methods is used for feature extraction, and convolution neural networks (CNNs) are used in farm management such as fruit counting, prediction of crop yield, and crop disease detection [4, 10].

The contribution of this work is a comprehensive analysis and investigation of how deep convolution neural networks perform in the early diagnosis of plant disease detection and developed a method for choosing and understanding pre-trained CNN models for plant image analysis. Cotton plant disease dataset is used and trained Google deep learning pre-trained models VGG16, VGG19, ResNet50, and InceptionV3. The performance of these models is compared with a customized convolution neural network. This study provides novel results by visualizing Google deep learning pre-trained models applied for plant image analysis.

The sections of this paper are planned as follows: The literature on disease detection using deep learning models is reviewed in Sect. 2, and the methods for detecting cotton plant disease are described in Sect. 3. Finally, Sects. 4 and 5 illustrate the experimental findings, conclusions, and future studies.

2 Related Work

Plant disease identification using deep learning and computer vision has gotten much attention in the past decade. Since the physical visual inspection to identify the plant disease is a very tedious, expensive, and time-consuming process. Automatic crop disease detection is one of the significant problems, and various machine learning approaches have been proposed with high accuracy and reduced costs and subjectivity. Some commonly used ML algorithms to detect diseases are SVM, KNN, DT, RF, and Ensemble techniques and are effective under specific setups [14]. The authors worked on the cascading of two classifiers for the disease detection of cotton plants. The first classifier segments the leaf from the background, and the second classifier was used for hue and luminance from the HSV color space. KNN classifier was used for the study and got an accuracy of 82.5% [15]. The change of conditions results in a reduction in performance significantly. Deep Learning algorithms have proved high possibilities and performance in agricultural problems and supply reasonable solutions. Authors worked on CNN transfer learning architectures AlexNet and GoogLeNet and trained the models [6]. The Plant Village image dataset is used, which consists of 54,306 images from 38 varieties of plants. The images of the plants were resized into 256×256 pixels, the prediction model was built, and the model optimization was performed. Authors have worked on rice plant disease detection using Deep CNN [7]. They prepared the dataset with 500 images captured from the rice crop fields. The authors used a convolution Neural Network, Back Propagation, Support Vector Machine, and Particle Swarm Optimization methods. The results were compared with different pooling with different filter sizes. Authors [8] employed 1.8 million photos to train a deep neural network and then applied a fine-tuning technique to transfer learned recognition abilities from available domains to the unique plant identification problem. Authors [9] used pre-trained models VGG16 and RESNET50 to identify cotton leaf disease, cotton stage, and weed in cotton. The performance of RESNET50 was compared to that of VGG 16, and it was found that RESNET50 outperformed VGG 16. To detect banana diseases and pests, the authors [11] used transfer learning to train the model and used three different CNN architectures. Eighteen thousand photos were obtained from various areas of the banana trees. According to their experimental study, most algorithms achieved greater than 90% accuracy.

3 Methodology

A general overview of the proposed system for early disease identification is presented in a block diagram and shown in Fig. 1. This framework can be utilized as a subsystem in AI-Based decision support networks, and the predictions are communicated to the farmers for early identification of potential risks in crop production.

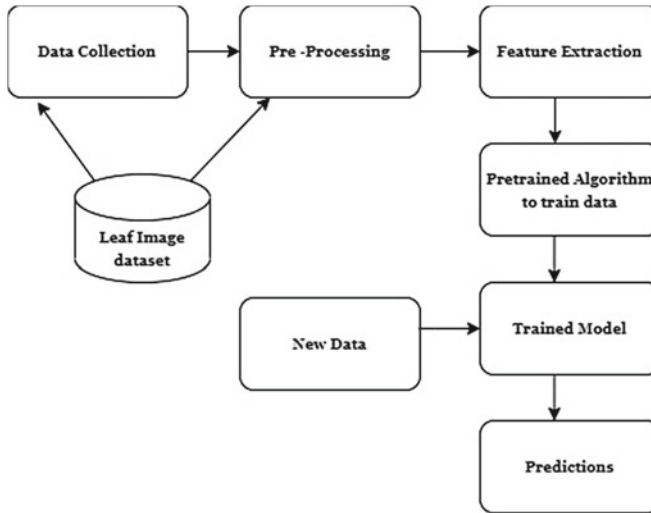


Fig. 1 Proposed framework

Initially, the system collects images from different sources and stores them in the database. Image pre-processing is critical because various types and formats of temporal and spatial data are required to be handled. It enhances the image quality and the necessary image features. Next, perform image segmentation to segment the image into its fundamental objects. Then, extract features to highlight the diseased region. The change in leaf color, texture, and morphology are the essential attributes to indicate a diseased leaf. Then, train the data using pre-trained models for classification. Now, the model is ready for prediction on test data. Finally, the model is ready for real-time data predictions after satisfactory test results.

3.1 Convolution Neural Network Models

Google pre-trained models VGG16, VGG19, ResNet50, Inception V3, and CNN models are used as training models for plant disease classification. CNN architecture comprises an input layer, a few convolution layers, a pooling layer, a fully connected layer, and a final output layer with SoftMax activation. The image data is fed into the convolution layer, which extracts features, and the pooling layer extracts feature values. The convolution and pooling layers can be added as layers to acquire more detailed features, depending on the complexity of the images. A fully connected layer combines the outputs of preceding layers into one vector applied as an input for the next succeeding layer. The final output layer performs the classification of the plant disease. Deep Convolution Neural Network models VGG16, VGG19, ResNet.50,

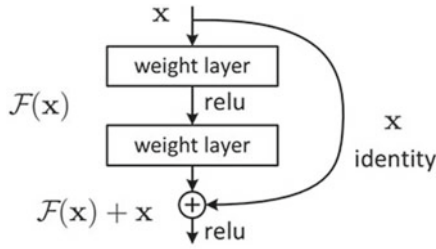


Fig. 6 Skip connections

connection’s path instead of waiting for the gradient to propagate back one layer at a time.

This results in the gradient skipping some layers and allows the gradient to reach the beginning nodes faster. Figure 6 depicts the path of the skip connection in ResNet.

3.5 Inceptionv3

Inception-v3 is a GoogLeNet extended network that uses transfer learning to produce good classification performance in several biomedical applications. The inception-v3 model constructs many different-sized convolutional filters into a new filter. The number of parameters that must be trained is reduced as a result of this design, which in turn reduces computational complexity. Inception-v3 architecture is depicted in Fig. 7.

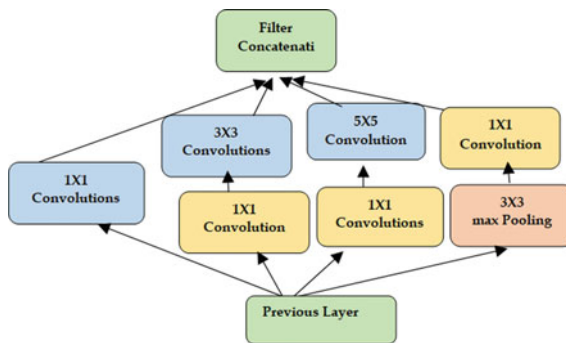


Fig. 7 InceptionV3 architecture

4 Experiments and Results

The cotton plant disease dataset contains images of the cotton plant and leaf, which are healthy and diseased, categorized into four labeled classes downloaded from Kaggle. The images are divided into four categories: plant healthy, leaf healthy, plant infected, and leaf infected. There are 1813 images in total, divided into four groups. The image dataset is divided into 90% train data and 10% test data at random. Given that the dataset is large, a higher ratio of the train to test data would improve accuracy. As a result, about 1706 photographs were used to train the model, with the remaining 106 images being utilized to assess the model’s performance. The image dimensions of the dataset are 256×256 pixels, with a resolution of 96 dpi and a bit depth of 24. Random samples of diseased plant and leaf images are shown in Figs. 8(a) and (b), while healthy plant and leaf images are shown in Figs. 8(c) and (d).

The filter transforms the input image into a feature map. The sample output of the first hidden layer of the 5th, 15th, 24th, 30th, and 31st are shown, i.e., Fig. 9. Figure 9(a) shows the original image. The features learned by the 5th, 15th, 24th, 30th, and 31st filters are shown in Figs. 9(b) and (f). In the first Conv2D layer, 32 filter images with the same padding, one stride, and the Rectified Linear Units (ReLU) activation function. Except for the last convolution layer, the Convolution layer is pipelined with the Max Pooling layer. There is some amount of data loss in learning by the neurons. Some filters are predominant in learning independently, and some are very weak in learning. The output of the first layer is given as the input to the second Conv2D layer. In second Conv2D layer has 64 filters images are passed through Max Pooling. The interpretation of the hidden layers is abstract, and humans cannot understand, but the neural network understands. There is a Conv2D and a Max Pooling with 256 filters in the last layer. Additionally, there is a dropout

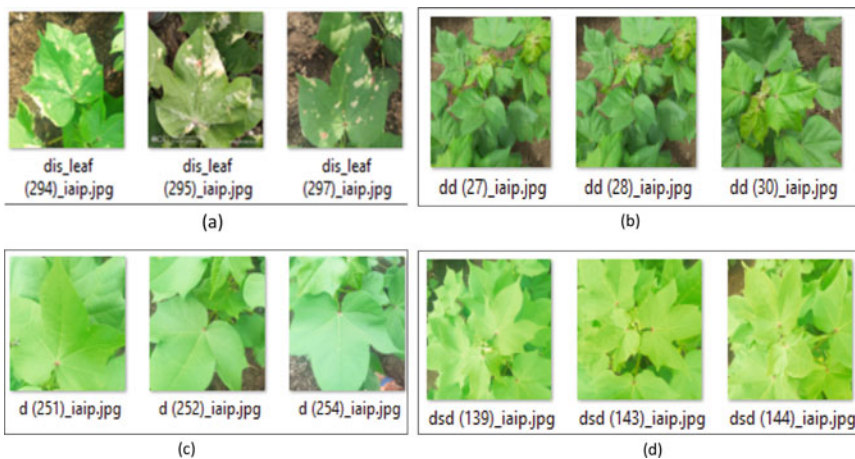


Fig. 8 Samples **a** Infected leaf; **b** Infected plant; **c** Healthy leaf; **d** Healthy plant

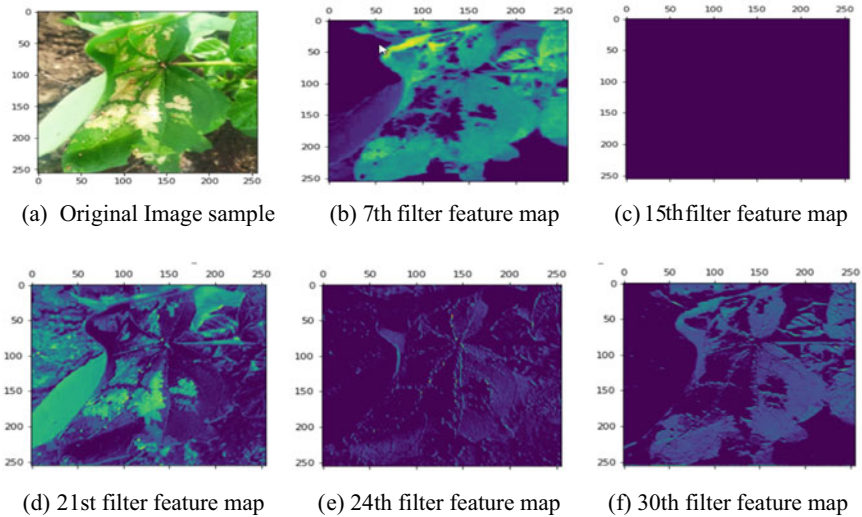


Fig. 9 Output visualization of a first hidden layer feature map

regularization layer with 25% of the dropout ratio of the neuron to avoid overfitting. There is significant data loss in the last layer. The image is broken, and there are 256 feature maps. These feature maps are to recognize images. There are Conv2D Dense layer in the output layer and has four filters or neurons to map output classes. The result is the accumulated value of all neurons contributed to output. Then, the images are saved in the HDF5 file format to handle extensive heterogeneous complex data.

5 Results and Discussion

CNN model is built with three convolutions 2D and Max Pooling Layers and Flatten with three Convolution 2D layers and dropout layer with 25, 10, and 25% output dropouts. The performance of these models is analyzed. The final layer has four units, which correspond to 24 classes, and feeds into the SoftMax layer to calculate the output probability. The experiment was conducted on the cotton plant-leaf image dataset, containing 1813 images belonging to four classes. The image dataset is randomly split into 90% train data and 10% test data, and the models were run with 20 epochs. Table 1 compares different deep learning models- VGG16, VGG19, ResNet50, Inception V3, and CNN models. VGG16 correctly identified healthy and diseased classes with train and test accuracy of 99.88 and 95.3%.

Figure 10(a) shows a comparison graph of train accuracy for these deep learning models. The figure shows that the maximum train accuracy for the model VGG16 was 99.88%. Figure 10(b) shows a comparison graph of these models' test accuracy.

Table 1 Test and train accuracy of DL models

Epochs	VGG16		VGG19		ResNet50		InceptionV3		CNN	
	Trian Acc	Test Acc	Trian Acc	Test Acc	Trian Acc	Test Acc	Trian Acc	Test Acc	Trian Acc	Test Acc
1	90.98	85.85	74.75	84.91	63.74	54.72	91.80	90.57	48.92	48.11
2	95.31	87.74	89.87	90.57	74.58	66.04	93.61	88.68	60.28	44.34
3	96.37	93.40	92.62	86.79	71.00	61.32	95.08	90.57	70.01	42.45
4	97.48	94.34	94.55	93.40	72.58	66.04	94.49	85.85	78.62	77.36
5	97.36	95.28	94.02	93.40	74.11	59.43	95.31	88.68	82.02	70.75
6	98.13	95.28	96.90	93.40	78.03	65.09	95.78	92.45	86.64	77.36
7	98.54	94.34	97.31	93.40	70.18	66.04	95.02	93.40	87.81	83.02
8	98.65	91.51	96.19	94.34	76.86	67.92	93.67	95.28	88.87	79.25
9	99.12	95.28	96.78	85.85	75.28	66.04	95.02	91.51	89.92	75.47
10	99.12	95.28	98.07	94.34	78.85	56.60	95.49	86.79	91.80	79.25
11	99.18	93.40	97.60	93.40	76.33	61.32	96.66	93.40	91.74	81.13
12	99.36	93.40	98.59	97.17	78.44	62.26	94.84	90.57	93.61	80.19
13	99.41	91.51	99.30	95.28	77.09	69.81	94.14	97.17	93.56	86.79
14	99.24	93.40	99.30	95.28	79.09	71.70	95.96	93.40	92.62	84.91
15	99.12	92.45	99.24	97.17	73.11	59.43	95.55	94.34	92.44	83.02
16	99.59	94.34	99.00	95.28	78.68	61.32	96.49	92.45	94.26	94.74
17	99.47	95.28	98.36	92.45	79.85	67.92	94.79	91.51	93.26	82.08
18	99.88	92.45	99.00	95.28	74.22	71.70	96.37	95.28	93.20	83.02
19	99.82	95.28	99.41	92.45	78.03	66.98	96.90	90.57	94.49	82.08
20	99.82	93.40	99.36	91.51	81.08	61.32	97.36	95.28	94.61	87.74
Max Acc	99.88	95.3	99.41	97.2	81.1	71.7	97.36	97.2	94.6	94.74

VGG19 and InceptionV3 had the highest test accuracy of 97.2%, while the CNN model had the lowest at 94.74%.

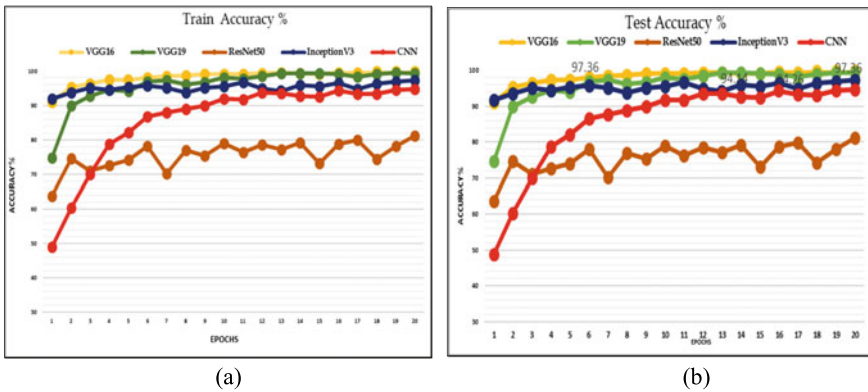


Fig. 10 a Train accuracy b Test accuracy

6 Conclusion

This paper focuses on the early identification of crop disease using DCNN algorithms for the prediction and recommendations in agricultural production. Here, early disease detection in crop management is demonstrated and how deep learning breakthroughs might benefit farming. It assists farmers in making precise judgments by guiding them through each stage of the decision-making process and giving the likelihood of various outcomes from alternative options. Also, it helps to enhance the agricultural output, efficiency, and harvest time monitoring in a crop. CNN, VGG19, and InceptionV3 had the highest test accuracy of 94.6, 97.2, and 94.74%, respectively. Instead of training the deep learning model from scratch, one can use pre-trained models that have already been trained, which helps to save resources and time. This research work is extended further by experimenting with the farming application within the agriculture ecosystem.

References

1. Wolfert S, Ge L, Jeroen Bogaardt CM (2017) Big data in smart farming – a review. *Agric Syst* 153:69–80
2. Golhani K, Balasundram SK, Vadamalai G, Pradhan B (2018) A review of neural networks in plant disease detection using hyperspectral data. *Inf Process Agric J* 5(3):354–371. ISSN 2214-3173
3. Wani H, Ashtankar N (2017) An appropriate model predicting pest/diseases of crops using machine learning algorithms. In: *IEEE 4th international conference on advanced computing and communication systems (ICACCS)*, pp 1–4
4. Huang GB, Zhou H, Ding X, Zhang R (2011) Extreme learning machine for regression and multiclass classification. *IEEE Trans Syst* 513–529
5. Sagar A, Jacob D (2020) On using transfer learning for plant disease detection. <https://doi.org/10.13140/RG.2.2.12224.15360/1>
6. Liu J, Wang X (2021) Plant diseases and pests detection based on deep learning: a review. *Plant Methods J* 17. <https://doi.org/10.1186/s13007-021-00722-9>. Article number 22
7. Tantalaki N, Souravlas S, Roumeliotis M (2019) Data-driven decision making in precision agriculture: the rise of big data in agricultural systems. *J Agric Food Inf* 20(4):344–380
8. Reyes K, Caicedo JC, Camargo JE (2015) Fine-tuning deep convolutional networks for plant recognition. *CLEF (Working Notes)* 1391:467–475
9. Banu T, Mani VRs (2020) Cotton crop monitoring system using CNN. *Xi'an Jianzhu Keji DaxueXuebao/J Xi'an Univ Archit Technol* 12(2). <https://doi.org/10.37896/JXAT12.03/529>
10. Toda Y, Okura F (2019) How convolutional neural networks diagnose plant disease. *Plant Phenom J*. <https://doi.org/10.34133/2019/9237136>. Article ID 9237136, 14 pages
11. Selvaraj MG, Vergara A, Ruiz H, Safari N, Elayabalan S et al (2019) AI-powered banana diseases and pest detection. *Plant Methods* 15(1):92
12. Khatoon S, Hasan MM, Asif A, Alshhari M, Yap (2021) Image-based on automatic diagnostic system for tomato plants using deep learning. *Comput Mater Continua Tech Sci Press* 67(1):595–12
13. Vakula Rani J, Jakka A, Kanuru H (2022) Disease detection in crop management using ensemble machine learning. In: Shakyia S, Balas VE, Kamolphiwong S, Du KL (eds) *Sentimental analysis and deep learning. Advances in intelligent systems and computing*, vol 1408. Springer, Singapore. https://doi.org/10.1007/978-981-16-5157-1_70

14. Vakula Rani J, Aishwarya J, Hamsini K (2022) Crop management using machine learning. In: Das AK, Nayak J, Naik B, Dutta S, Pelusi D (eds) Computational intelligence in pattern recognition, advances in intelligent systems and computing, vol 1349. Springer, Singapore. https://doi.org/10.1007/978-981-16-2543-5_49
15. Parikh A, Raval MS, Parmar C, Chaudhary S (2016) Disease detection and severity estimation in cotton plant from unconstrained images. In: 2016 IEEE international conference on data science and advanced analytics (DSAA), pp 594–601. <https://doi.org/10.1109/DSAA.2016.81>

An Error Dependent Enhancement Method for Images Captured in Dense Fog



Yucel Cimtay and Gokce Nur Yilmaz

Abstract Fog and haze naturally or artificially appearing in the environment, limit human visibility. As a way of improving the visibility, digital images are captured and many different image enhancement methods are applied to remove the fog and haze effects. One of the fundamental methods is the Dark Channel Prior (DCP) method. DCP can remove fog and haze on a single image by modelling the psychical diminishing structure of fog. In this study, the spectral signature of the DCP method was investigated by using the transmission maps produced by the results of the DCP method on the Spectral Hazy Image Database (SHIA) dataset, which consists of hyperspectral images taken in the visible and near infrared band range. In this study, it was observed that the transmission response of different regions in the image to the increase in fog density was different. By using this distinctiveness on two foggy images taken from the scene in two different high fog level, this study achieves to reveal the silhouette of the scene which is totally not visible to human eye.

Keywords Distance of visibility · Atmospheric effects · Image enhancement · Hyperspectral

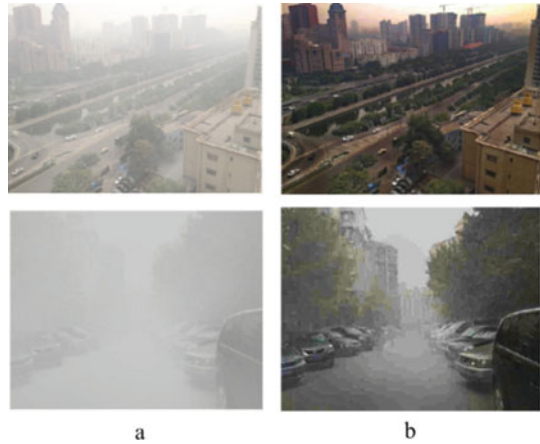
1 Introduction

The presence of haze, which is caused by microscopic water droplets or solid particles hanging in the air, causes a slew of problems in everyday living. The air can no longer be considered an isotropic medium since the transmitted light scatters. The scene picture captured by the camera or seen by human eyes has been severely deteriorated. The scattering gets more problematic as the distance from the target or the quantity of suspended particles rises. As a result, the distant target's features

Y. Cimtay (✉) · G. Nur Yilmaz
Computer Engineering, TED University, Ankara, Turkey
e-mail: yucel.cimtay@tedu.edu.tr

G. Nur Yilmaz
e-mail: gokce.yilmaz@tedu.edu.tr

Fig. 1 Hazy images and the resulted restored images. **a** Hazy image **b** Dehazed image



are lost more severely, and the contrast of the acquired image is also lowered. It is frequently necessary to remove haze from the gathered image in order to make it simpler for the observer to identify the object.

Studies in the literature within the scope of improving images containing fog and haze are based on traditional contrast enhancement methods, filtering, use and analysis of atmospheric scattering model, use of artificial neural networks and application of deep learning models [1]. Though most of the studies in the literature have focused on the enhancement of static images [2–4], enhancement methods on video have been also applied. In the defogging process, as the distance between the imaging device and the scene increases, the fog/haze thickness increases and the transmission of light decreases [1]. Similarly, when the fog/haze density is high and varies locally, the complexity of the fog removal process increases. In Fig. 1, clear images obtained from images with different densities of fog are given. As can be seen from the figure, as the fog density increases, the estimation error increases and the resulting image cannot be sufficiently improved.

2 Related Work

Many different methods are applied in the literature for the purpose of reducing fog and haze on images. Nonetheless, these methods can generally be grouped into three categories: contrast enhancement [5–7], restoration [8–11] and fusion-based [12–14] methods.

Contrast enhancement approaches can improve the visual quality of hazy images to some extent, but they cannot effectively remove the fog. Subcategories of image enhancement models are histogram enhancement [15–17], which can be applied locally and/or globally, frequency conversion methods: wavelet transform and homomorphic filtering, and Retinex method: single and multi-scale Retinex [18].

In the study performed in [19], objects were first segmented and then histogram equalization method was applied to these regions. In the study conducted in [20], the contrast enhancement method was applied to stereo images. Restoration-based methods focus on recovering lost information by modeling the image degradation model and applying inverse filtering.

Atmospheric Light Scattering (ALS) model is one of the most widely used models in image fog removal process. The most important factor which determines the success of the methods which use DCP as basis is the accurate estimation of the transmission rate and atmospheric light. The Dark Channel Prior (DCP) Method [21] is one of the most widely used methods which uses the dark channel, per pixel.

The study conducted in [22], employs Empirical Mode Decomposition method to enhance hazy images. Each image channel is decomposed into intrinsic mode functions and enhanced by multiplying the intrinsic mode functions of each channel with different weight values and summing. In [23], as a result of selecting the optimal color channel and applying non-linear transformation, foggy images are defogged.

Recent approaches in the case of defogging are mostly based on artificial intelligence approaches which use deep learning models [24, 25]. The study in [25], develops a deep architecture by using Convolutional Neural Network (CNN) and a new unit called “bidirectional rectified linear unit” was added to the neural network. Better results were obtained in this study compared to previous studies. The work in [26] used the end-to-end encoder-decoder CNN architecture to obtain the dehazed images. Besides the methods which apply defogging on single images, there are also methods in the literature that apply defogging on video data. For instance, the study in [27] trains the video data over CNN, based on the temporal similarity between consecutive video frames. Therefore it assumes that the atmospheric effects would also be similar. In addition, video defogging was performed in [28], on the basis of DCP, by utilizing the spatio-temporal similarity between video frames.

It is known that the accuracy of the transmission estimation of the DCP method in regions with very bright areas or objects is lower than in other regions. Accordingly, saturation problem occurs over the bright regions or objects [29]. In order to solve this problem, many studies have been carried out in the literature and better results have been obtained [30, 31].

In this study, a transmission signature based on haze density was created by using different intensity foggy images taken at a fixed wavelength in the Spectral Hazy Image Database (SHIA) [32] dataset. Basically, it is benefited the feature of different color tones in the image differ from each other in the transmission maps calculated by DCP. In fact, by considering and taking account the recovery performance difference of the DCP algorithm on different colored regions, enhanced clear images are handled from high density foggy images.

3 Dataset

SHIA dataset [32] which is used in this study is created as a spectral image hazy dataset. It consists of two real indoor scenes, each with 10 levels of fog, and their corresponding haze-free (ground-truth) images. Images are taken at every 10 nm from 450 to 1000 nm in the visible and near infrared band regions. The image resolutions are (1312×1082) and the total number of images is 1540. The hazy images and the haze-free images were taken under the same lighting conditions. Figure 2 shows 720 nm image with all fog levels.

Images of the 560 nm wavelength from the SHIA dataset are shown in Fig. 3. As one can observe from the images, density of the fog is increased along the levels of fog. As the amount of fog is increased in the room, the visibility of the scene is being decreased for human eye. Another set of images are given in Fig. 4 for 690 nm. The behavior is similar, as the fog intensity increases, the contrast of the images decrease.

4 Proposed Work

In this study, as the first step, the visibility of hyperspectral images taken at 10 nm intervals in the SHIA dataset was investigated. By visual inspection, the best visibility was obtained between 680–720 nm wavelengths. Example images belong to Level-4 fog intensity are given in Figs. 5 and 6. It can be observed that 720 nm is one of best visible bands. Therefore, in this study, the image of the 720 nm band, which is in the red light band of the electromagnetic spectrum, was studied in order to provide the best data in terms of visibility with the naked eye.

Transmission at a single point in physics is calculated as shown in Eq. 1. In this equation, t , $\beta(\lambda)$ and d represent transmission term, wavelength-based Rayleigh scattering coefficient and the distance between the point to the camera respectively. As depth increases, transmission decreases. Similarly, as $\beta(\lambda)$ increases, transmission value decreases.

$$t = e^{-\beta(\lambda)d} \quad (1)$$

Since DCP method estimates the transmission from a single image, there is no prior depth information available in this method. It basically estimates ALS transmission and ambient light. The ALS model is shown in Fig. 7 and its mathematic modelling is given in Eqs. 2–4 [1]. Here, $I(x, \lambda)$ represents the hazy image, $D(x, \lambda)$ represents the light passing through the haze after reflecting from the scene, and $A(x, \lambda)$ represents the atmospheric light reflected from the fog particles. The sensor integrates the incoming light and a hazy image is formed. In Eq. 3, $t(x, \lambda)$ is the transmission map of the hazy scene, $R(x, \lambda)$ is the light reflected from the scene and L_∞ is the atmospheric light. The transmission term is expressed as $e^{-\beta(\lambda)d(x)}$

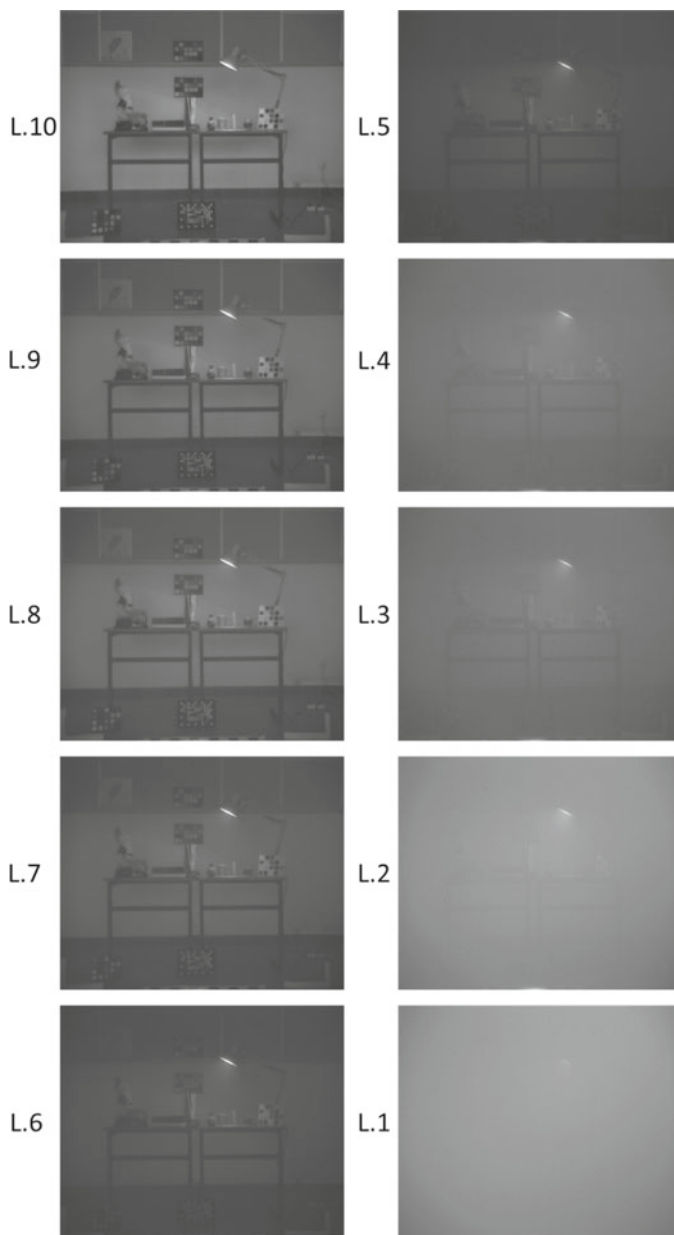


Fig. 2 SHIA dataset. Hazy images taken under 10 different level (level10-level1) of fog at 720 nm

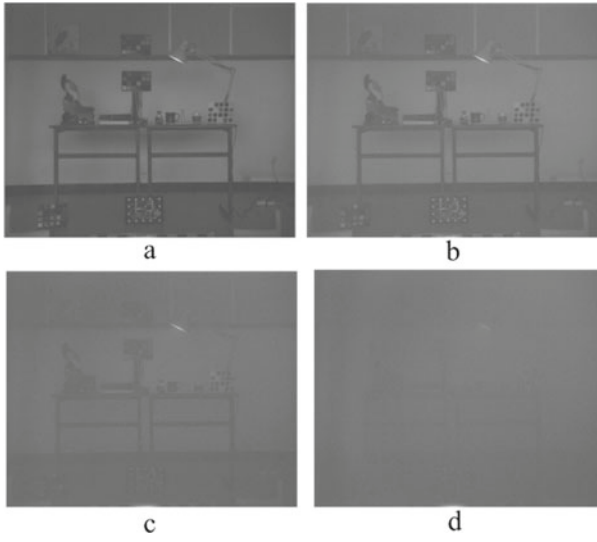


Fig. 3 SHIA dataset. Hazy images taken under different level of fog at 560 nm. **a** Clear image **b** Low level fog **c** Medium level fog **d** Dense level fog

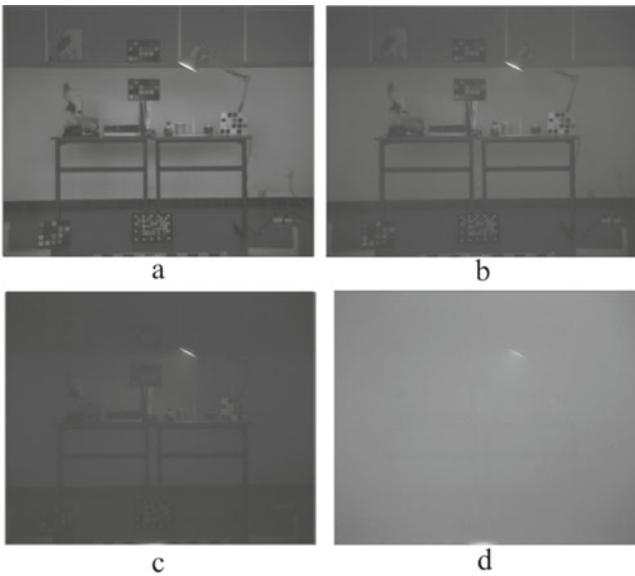


Fig. 4 SHIA dataset. Hazy images taken under different level of fog at 690 nm. **a** Clear image **b** Low level fog **c** Medium level fog **d** Dense level fog

Fig. 5 Images captured at level 4 fog. **a** 460 nm **b** 580 nm **c** 720 nm **d** 810 nm

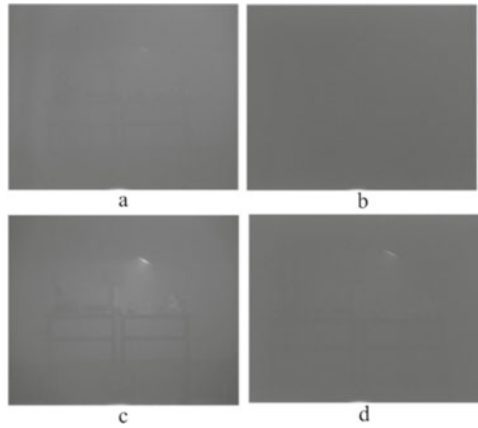
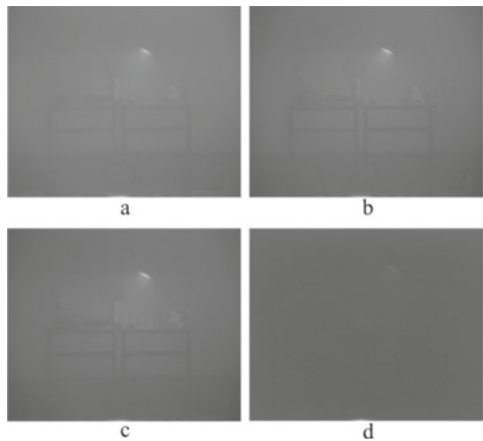


Fig. 6 Images captured at level 4 fog. **a** 700 nm **b** 710 nm **c** 720 nm **d** 730 nm



where $\beta(\lambda)$ is wavelength (λ) dependent Rayleigh scattering coefficient [33] and the d is distance between point and camera.

$$I(x, \lambda) = D(x, \lambda) + A(x, \lambda) \tag{2}$$

$$I(x, \lambda) = t(x, \lambda)R(x, \lambda) + L_\infty(1 - t(x, \lambda)) \tag{3}$$

$$I(x, \lambda) = e^{-\beta(\lambda)d(x)} R(x, \lambda) + L_\infty(1 - e^{-\beta(\lambda)d(x)}) \tag{4}$$

In the method developed in this study, which is described in Fig. 8, the transmission maps which are estimated by DCP method from the images in two different fog density levels were used. By taking the natural logarithm of Eq. 1, Eq. 5 is obtained.

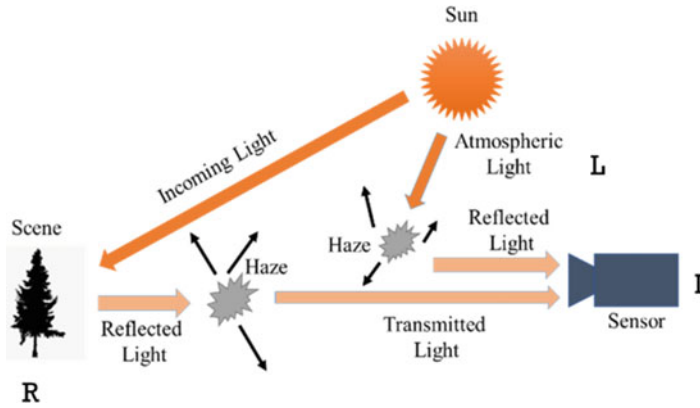


Fig. 7 Atmospheric light scattering model

In this case, different transmission maps are estimated for each fog density level (Eq. 6). When this equation is arranged, Eqs. 7 and 8 are obtained, respectively. As a result, the ratio of the logarithm of their transmission maps are calculated from the images obtained at two different fog densities by the DCP method. This value is a measure of the response of the transmission in the environment to the change in fog density level.

$$\ln(t) = -\beta(\lambda, \phi)d \tag{5}$$

$$\ln(t_i) = -\beta_i(\lambda, \phi)d \tag{6}$$

$$\frac{\ln(t_i)}{\ln(t_j)} = \frac{-\beta_i(\lambda, \phi)d}{-\beta_j(\lambda, \phi)d} \tag{7}$$

$$\frac{\ln(t_i)}{\ln(t_j)} = \frac{-\beta_i(\lambda, \phi)}{-\beta_j(\lambda, \phi)} \tag{8}$$

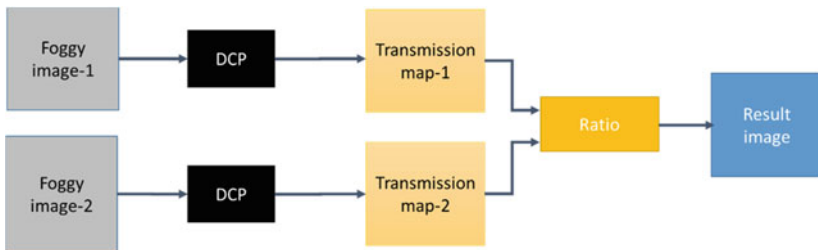


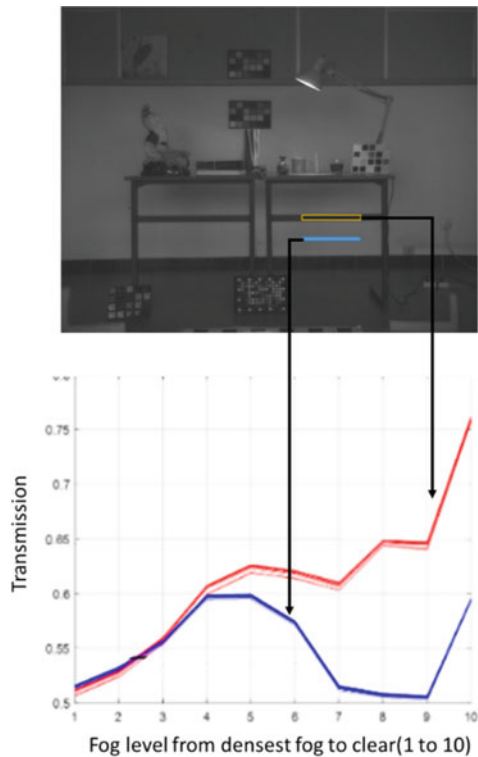
Fig. 8 Proposed method block diagram

SHIA dataset includes hyperspectral images taken at 10 different fog densities. Defogging is performed on all 720 nm images along 10 fog levels by using DCP method. The transmission maps which are obtained as a result of the DCP method are recorded. In Fig. 9, the estimated transmission spectra for those pixels corresponding to different image regions is shown. Note that the picture of the scene is the ground truth to show the pixel regions clearly. Normally, transmission spectra is estimated by using the images belong to 720 nm along the fog levels from 1 to 10.

As it can be seen from the figure, as the fog density changes, the transmission behaviors of the relatively bright and less bright regions differ. Considering the transmission behavior of the bright region in particular, the transmission value predicted by the DCP method decreases, although the fog density decreases. However, the transmission value should have tended to increase. The similar behavior is valid for some fog levels (6, 7) in the less bright region, but the spectra has lesser error compared to brighter region. This point shows one of the missing aspects of the DCP method, and this has been emphasized by several studies in the literature and different methods have been put forward for its solution [29–31, 34].

In this study, this drawback of DCP method, has been transformed into an advantage by looking at the problem from a different perspective. The hypothesis put forward in this paper is to obtain a new image by getting use of transmission error

Fig. 9 Transmission spectra of different regions with different color tones along fog levels 1–10



produced by DCP method as this error rate changes according to the pixel brightness value. Therefore, since the transmission value produced in regions that differ in contrast (especially between the background and objects in the image) will be different, these regions can be revealed in the newly created image. It is of course possible to use the image obtained directly as a result of DCP in existing foggy images. However, when the fog density becomes too high (for example, Level1 and Level2 for SHIA), DCP and many other fog removal methods fall short of quality.

5 Result Analysis

In Fig. 10, the result of foggy image and DCP method in 720 nm image for Level 2 is given. As it can be observed from the figure, it is not possible to clear the fog from the image meaningfully when it includes dense fog.

The transmission maps (a and b) obtained by DCP method on 720 nm Level-1 and Level-2 images are given in Fig. 11. The result of the proposed method and the one with histogram equalization applied to this image are shown in c and d respectively.

To obtain the result in the case of less fog density, the same process was performed on Level2 and Level3 images. Figure 12a shows the restored image which is the result of DCP on Level-3 image, Fig. 12b is the image obtained after applying histogram equalization on the result of DCP, Fig. 12c and d are the results of the proposed method and the image obtained after applying histogram equalization on it. As it can be seen from the figure, the DCP result is improved as the fog density decreases. However, the image obtained as a result of the method proposed in this study improves more and reveals the details in the scene much more clearly.

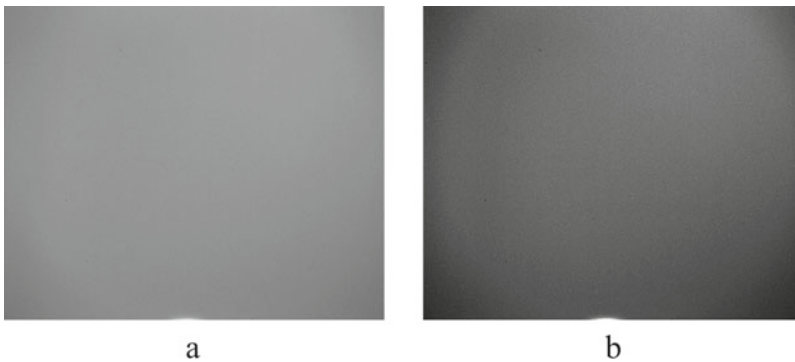


Fig. 10 Restored images by DCP at 720 nm. **a** Foggy image **b** DCP result

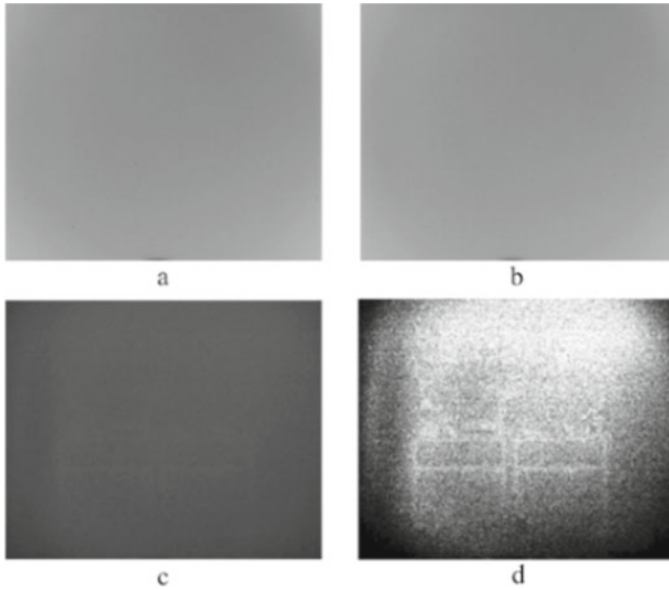


Fig. 11 DCP and proposed method results. **a** DCP result on Level1 foggy image **b** DCP result on Level2 foggy image **c** Result of proposed method **d** Histogram enhanced result of the proposed method

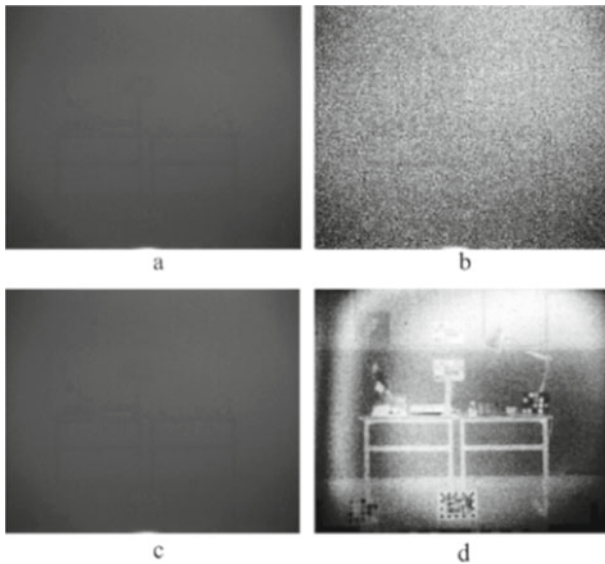


Fig. 12 Results obtained from Level2 and Level3 Images. **a** DCP result on Level3 Image **b** DCP with histogram equalization **c** Proposed method **d** Proposed method with histogram equalization

6 Conclusion

In this study, an error-dependent method is proposed for defogging of images. It has been utilized that the DCP method performs with different amount of error in the regions with different brightness. This difference becomes more apparent in the images containing medium and low levels of fog. However, as the fog density decreases, traditional methods already produce promising results. For this reason, in this study, we focused especially on the densest fog situations.

As it can be seen from the results, when there is such a large amount of fog so that the scene cannot be seen with naked eye as for Level-1 and Level-2 images, the silhouette of the scene has been made prominent. Likewise, a much better visual result was obtained in Level-2 and Level-3 images compared to the traditional DCP method. Since the fog density changes dynamically in the real world scenes and proposed method get use of the little fog density difference between frames, it can be used successfully for real world foggy videos.

SHIA dataset includes foggy images along the wavelength in visible and NIR bands of electromagnetic spectrum. However it is observed that the best visibility is handled in the red band region. Therefore, according to the results of proposed method, a filter can be used with traditional cameras to obtain images in the red band between 680–720 nm. By using the proposed method with that kind of imagery, it will be possible to produce the silhouette of the scene in dense fog conditions.

SHIA dataset presents a very useful data to be used with defogging methods and applications. However it is the only public hyperspectral hazy image dataset captured along different fog levels in the literature. Therefore the authors plan to create a similar dataset in the future for further studies.

References

1. Cimtay Y (2021) Smart and real-time image dehazing on mobile devices. *J Real Time Image Proc* 18:2063–2072. <https://doi.org/10.1007/s11554-021-01085-z>
2. He K, Sun J, Tang X (2011) Single image haze removal using dark channel prior. *Pattern Anal Mach Intell IEEE Trans* 33(12):2341–2353
3. Gibson KB, Nguyen TQ (2011) On the effectiveness of the Dark Channel Prior for single image dehazing by approximating with minimum volume ellipsoids. In: *IEEE international conference on acoustics, speech and signal processings (ICASSP)*, pp 1253–1256
4. Zhang Q, Li X (2015) Fast image dehazing using guided filter. In: *IEEE 16th international conference on communication technology (ICCT)*, pp 182–185
5. Al-Sammarai, MF (2015) Contrast enhancement of roads images with foggy scenes based on histogram equalization. In: *Proceedings of the 10th international conference on computer science & education*, pp 95–101
6. Kim JH, Sim JY, Kim CS (2011) Single image dehazing based on contrast enhancement. In: *Proceedings of the IEEE international conference acoustics, speech and signal processing*, pp 1273–1276
7. Cai WT, Liu YX, Li MC, Cheng L, Zhang CX (2011) A selfadaptive homomorphic filter method for removing thin cloud. In: *Proceedings of the 19th international conference geoinformatics*, pp 1–4

8. Tan K, Oakley JP (2001) Physics-based approach to color image enhancement in poor visibility conditions. *J Opt Soc Am* 18(10):2460–2467
9. Gibson KB, Belongie SJ, Nguyen TQ (2013) Example based depth from fog. In: Proceedings of the 20th IEEE international conference on image processing, pp 728–732
10. Fang S, Xia XS, Xing H, Chen CW (2014) Image dehazing using polarization effects of objects and airlight. *Opt Express* 22(16):19523–19537
11. Galdran A, Vazquez-Corral J, Pardo D, Bertalmio M (2015) Enhanced variational image dehazing. *SIAM J Imaging Sci* 8(3):1519–2154
12. Son J, Kwon H, Shim T, Kim Y, Ahu S, Sohng K (2015) Fusion method of visible and infrared images in foggy environment. In: Proceedings of the international conference on image processing, computer vision, and pattern recognition, pp 433–437
13. Guo CG, Li C, Guo J, Loy CC, Hou J, Kwong S, Cong R (2020) Zero-reference deep curve estimation for low-light image enhancement. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 1780–1789
14. Li S, Zhou Q (2021) Single image dehazing based on fusion of sky region segmentation. *J Phys Conf Ser* 1971
15. Simi VR, Edla DR, Joseph J, Kuppili V (2020) Parameter free fuzzy histogram equalisation with illumination preserving characteristics dedicated for contrast enhancement of magnetic resonance images. *Appl Soft Comput* 93
16. Joseph J, Periyasamy R (2018) A fully customized enhancement scheme for controlling brightness error and contrast in magnetic resonance images. *Biomed Signal Process Control* 39:271–283
17. Joseph J, Sivaraman J, Periyasamy R, Simi VR (2017) An objective method to identify optimum clip-limit and histogram specification of contrast limited adaptive histogram equalization for MR images. *Biocybernet Biomed Eng* 37(3):489–497
18. Hao W, He M, Ge H, Wang C, Qing-Wei G (2011) Retinex-like method for image enhancement in poor visibility conditions. *Procedia Eng* 15:2798–2803
19. Öztürk N, Öztürk S (2021) Bölütleme Tabanlı Yeni Görüntü İyileştirme Yöntemi. *Avrupa Bilim ve Teknoloji Dergisi, Ejosat Özel Sayı 2021 (RDCONF)*, pp 975–981. <https://doi.org/10.31590/ejosat.1041197>
20. Gövem B, Sayinta M, Somçağ E, Dönmez F (2013) Depth based 3D sharpness and contrast enhancement application on stereo images. In: 2013 21st signal processing and communications applications conference (SIU), pp 1–4. <https://doi.org/10.1109/SIU.2013.6531555>
21. Kaiming H, Jian S, Xiaoou T (2011) Single image haze removal using dark channel prior. *IEEE Trans Pattern Anal Mach Intell*
22. Çelebi AT, Güllü MK, Ertürk S (2013) Enhancement of fog degraded images using empirical mode decomposition. In: 2013 21st signal processing and communications applications conference (SIU), pp 1–4. <https://doi.org/10.1109/SIU.2013.6531404>
23. Thanh LT, Thanh DNH, Hien NN, Erkan U, Prasath VBS (2021) Single image dehazing with optimal color channels and nonlinear transformation. In: 2020 IEEE eighth international conference on communications and electronics (ICCE), pp 421–426. <https://doi.org/10.1109/ICCE48956.2021.9352087>
24. Haouassi S, Di W (2020) Image dehazing based on (CMTnet) cascaded multi-scale convolutional neural networks and efficient light estimation algorithm. *Appl Sci* 10:1190
25. Cai B, Xu X, Jia K, Qing C, Tao D (2016) DehazeNet: an end-to-end system for single image haze removal. *IEEE Trans Image Process* 25(11):5187–5198
26. Rashid H, Zafar N, Javed Iqbal M, Dawood H, Dawood H (2019) Single image dehazing using CNN. *Procedia Comput Sci* 147:124–130
27. Ren W et al (2019) Deep video dehazing with semantic segmentation. *IEEE Trans Image Process* 28(4):1895–1908. <https://doi.org/10.1109/TIP.2018.2876178>
28. Dong T, Zhao G, Wu J, Ye Y, Shen Y (2019) Efficient traffic video dehazing using adaptive dark channel prior and spatial-temporal correlations. *Sensors* 19(7):1593. <https://doi.org/10.3390/s19071593>

29. Farid MS, Fang Z, Wu Q, Huang D, Guan D (2021) An improved DCP-based image defogging algorithm combined with adaptive fusion strategy. In: *Mathematical problems in engineering*
30. Han P, Yan W, Wang D, Qin Y, Xu Z (2021) Single image dehazing method via sky-regions segmentation and dark channel prior. In: *2021 4th international conference on intelligent autonomous systems (ICoIAS)*, pp 60–64. <https://doi.org/10.1109/ICoIAS53694.2021.00019>
31. Pal T (2022) A robust method for dehazing of single image with sky region detection and segmentation. *Int J Image Graph*
32. El Khoury J, Thomas J-B, Mansouri A (2020) A spectral hazy image database. In: *International conference on image and signal processing*. Springer, Cham
33. <https://www.alanzucconi.com/2017/10/10/atmospheric-scattering-3>
34. Arulmozhi N, Chitra S (2022) Control approaches through time weighted error and gain margin tuning for unstable systems. *J Innov Image Process* 4(1):34–42

3D Video QoE Based Adaptation Framework for Future Communication Networks



Gokce Nur Yilmaz and Yucel Cimtay

Abstract A thorough investigation of the characteristics of 3D videos and the circumstances that support them in content-related contexts is necessary to help advance 3D video adaption systems. In order to create an advanced 3D video Quality of Experience (QoE) based adaptation framework for smart service management of future communication networks, various elements and situations related to content can be employed as milestones. Given this knowledge, the spatial resolution of a color + depth map 3D video representation is taken into account in this study as a factor to suggest a 3D video QoE based adaption framework. In order to construct this framework, the content-related contexts—namely, the motion and structure of a color video and the relative depth and aerial perspective of a depth map are taken into consideration. Under the condition that certain requirements are met and the 3D video QoE is maintained at an ideal level, the performance assessment results obtained using the suggested framework demonstrate its efficacy for selecting the best spatial resolutions for the color + depth map videos.

Keywords 3D video QoE · Video adaptation · Depth perception · Scalable video coding · Video quality

1 Introduction

The constant advancement of 3D video technologies has increased interest in a wide range of 3D applications [1]. While these advancements depend on 3D video coding, transmission, etc., 3D video adaption technology, which can greatly aid in smart service management of future communication networks, has not yet attracted the necessary attention.

Using various terminals and heterogeneous networks, users can engage with a 3D video anywhere. Mobile phones that support 3DTVs are a couple of examples of various terminals. Examples of constraints on the heterogeneous networks include

G. N. Yilmaz (✉) · Y. Cimtay
Department of Computer Engineering, TED University, Ankara, Turkey
e-mail: gokce.yilmaz@tedu.edu.tr

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
P. P. Joby et al. (eds.), *IoT Based Control Networks and Intelligent Systems*,
Lecture Notes in Networks and Systems 528,
https://doi.org/10.1007/978-981-19-5845-8_54

757

congestion and bandwidth restrictions. The many user preferences and ambient factors can further amplify these limits. Creating sophisticated 3D video adaptation systems is a requirement for smart management infrastructures of future networks in order to address these limits in the most transparent manner feasible and enhance users' Quality of Experience (QoE) [1].

An adaptation system depends on adaptation modules and adaptation-decision-taking. To choose the best parameters for adaptation operations while taking a set of restrictions into account, an adaptation decision taking module is employed. As a result, the idea of adapting decision taking might be referred to as "the brain" of a 3D video adaptation operation. A video form is transformed into another throughout the adaptation process using an adaptation module [2, 3].

In the literature, various techniques for adapting decision-making processes have been utilized. One of the most popular and effective decision-making techniques in the literature is optimization-based adaptation. The utility-based adaptation technique is taken into account by the optimization-based adaptation decision taking methods. This approach uses the best adaptation parameters to maximize or minimize the quantitative values of target functions, known as Utility Functions (UFs). This data is used in this study to create a 3D video QoE based adaptation framework using a UF-based adaptation decision taking procedure.

A 3D video adaptation system's primary goal is to maximize the QoE of viewers in terms of perceived quality and depth perception [2-4]. The parameters that improve the 3D video QoE should be found in order to reach the overall goal of the 3D video adaptation systems. By utilizing a novel UF-based adaptation decision taking method, these criteria can be utilized to construct the suggested adaptation structure.

A 3D video has several dimensions. Different 3D video representation formats and characteristics define this kind. This study takes into account color + depth map-based 3D video representation because it is one of the most effective representation forms [3]. Additionally, given that a 3D video QoE consists of two HVS-related components, color video and depth perception, these two components are combined to provide the 3D video QoE in this study. The spatial resolution is used as the basic factor in the suggested framework since it is believed to be one of the important characteristics defining the HVS-related 3D video QoE perception. While forming a 3D video, a color video and depth map of a 3D video supplied through a communication channel may have similar or different spatial resolution combinations. Due to the communication channel's bit rate limitations, a color video, for example, can have a resolution of High Definition (HD), but a depth map can have a resolution of Standard Definition (SD). These spatial resolution combinations should be altered while creating a 3D video representation while simultaneously attempting to maintain the 3D video QoE at the highest possible level. In light of this knowledge, the suggested framework is based on finding the best spatial resolution combinations while adhering to a number of requirements (such as bit rate) and maintaining the best possible 3D video quality of experience.

In a UF, it is also necessary for settings connected to the content to support the spatial resolution important factor in determining the best adaptation parameters in the suggested adaptation framework. Therefore, the proposed framework uses

content-related contexts that are thought to be fairly effective in the color video and depth perception aspects of a 3D video QoE from the perspective of spatial resolution. The suggested framework makes use of the motion and structure of a color film as well as the relative depth and aerial perspective of a depth map video as content-related contexts.

The Scalable Video Coding (SVC) concept [5] is taken into consideration when constructing the suggested framework due to the spatial scalability support it provides, even if there are other video coding standards that can be used to encode color video and depth map representation forms of 3D videos.

There are five sections in this paper. Section 2 provides a description of the discussion around the color + depth 3D video representation form. The proposed framework is introduced in Sect. 3. Section 4 discusses the framework's findings. The paper is concluded with Sect. 5, which also discusses further research.

2 Color + Depth Map 3D Video

Each pixel in the depth map, which sets the distance of the associated pixel from the viewer, has an accompanying pixel in the color image, as can be seen from the color + depth map representation form of a 3D movie shown in Fig. 1.

The utilization of left and right video pairs is the foundation of one of the stereoscopic video representations mentioned in the literature. Through this depiction, stereoscopy can be created in the simplest and most cheap manner [6, 7]. A different type of stereoscopic video representation that employs the Depth-Image-Based Rendering (DIBR) method is called color + depth stereoscopic video representation. A color and depth map can be used to create a representation of the left and right perspectives of a 3D video using the DIBR method [8, 9].

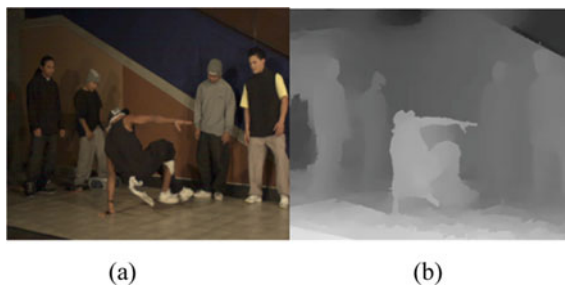


Fig. 1 Breakdance video: **a** color image **b** depth map

3 Proposed Framework

A color + depth map of a 3D movie can contain a variety of spatial resolution combinations, including Full High Definition (Full HD), HD, Standard Definition (SD), and Common Intermediate Format (CIF). The foundation of the proposed framework is built on Fig. 2, which shows these combinations (CIF). For example, as shown by the image. In both cases, a depth map and color video can be in Full HD, HD, SD, or CIF quality.

The architecture of the suggested framework for a multimedia communication chain is displayed in Fig. 3. As shown in the picture, the framework is separated into two modules: adaptation decision taking and adaptation. More details on these modules are given in the following subsections.

3.1 Adaptation Decision Taking Module

The adaptation decision taking module in the proposed framework is the core of this framework. It includes a UF used for determining the optimum adaptation parameters. The proposed UF (i.e., UF_{CD}) integrates two relatively independent UFs associated with the video quality and depth perception of the 3D video QoE. Thus, the general definition of the UF_{CD} is:

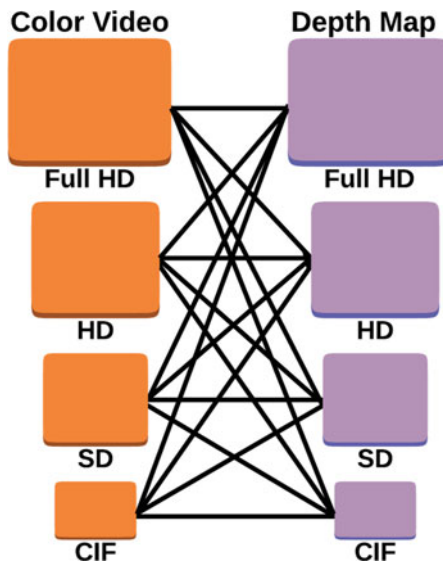


Fig. 2 The color + depth map spatial resolution combinations

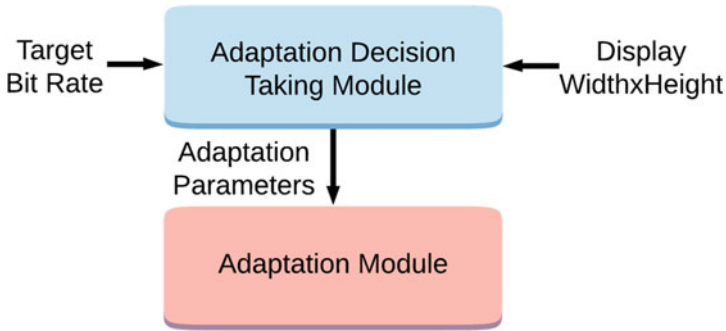


Fig. 3 The architecture of the proposed framework

$$UF_{CD} = UF_{VQ}UF_{DP} \tag{1}$$

where, UF_{CD} includes UF_{VQ} and UF_{DP} which are the 3D video quality and depth perception related UF functions, respectively. The development steps of these UF functions will be discussed in detail in the following sub-sections.

1) *Development of UF_{VQ}*

Envisaging the fact that the spatial resolution, motion, and structure of a color video are effective while forming the UF_{VQ} , they are used to develop UF_{VQ} . The effects of spatial resolution, motion, and structure of a color video on the 3D video QoE and how they are measured and/or integrated in UF_{VQ} will be discussed in the following sub-sections.

a) *Spatial Resolution of a Color Video*

In a color video plus depth map 3D video representation form, the spatial resolution of the color video part is a key factor for characterizing the HVS-related 3D QoE. In the proposed framework, this key factor should be associated with content related contexts effective on the HVS-related 3D QoE to be able to determine the most optimum adaptation parameters using UF_{VQ} . Therefore, the motion and structure of a color video are considered as content related contexts in UF_{VQ} .

b) *Motion of a Color Video*

The optical flow algorithm presented in the pyramidal Lucas&Kanade technique is used to calculate the motion intensities of the color videos. The iterative Lucas and Kanade optical flow algorithm is known as the pyramidal Lucas and Kanade method. This iterative version uses pyramidal representations, which reflect the various spatial resolutions of each video frame [11].

The optical flow computation in this approach begins with the very deep pyramidal resolution. This is because at this resolution, where the search range is less, extensive computations are not required. The optical flow computations are then transferred to the higher resolution computations. In the greater

resolution, this enables the use of a simpler motion estimating algorithm [12]. The optical flow measurements should rely on the prominent motion information points. Considering this, Shi and Tomasi method [13] is used to select prominent points in the optical flow estimations in this study.

Let k and l be two frames of a color video. If x and y are two pixel coordinates of the frames k and l , the motion intensity of a color video is computed as follows:

$$\Pi(i) = \sum_{d=0}^{NoP} \sqrt{(k_y^d - l_y^d)^2 + (k_x^d - l_x^d)^2} \quad (2)$$

In this equation, $\Pi(i)$ is the motion intensity of the i^{th} frame of a color video. NoP is the number of feature points. After the motion intensity measurement in each frame of a videos, these intensities are combined to determine the motion intensities for the original and distorted color videos, as follows:

$$M_{\alpha_C} = \sum_{i=1}^{NoF_C} \Pi_{\alpha_C}(i) \quad (3)$$

$$M_{\beta_C} = \sum_{i=1}^{NoF_C} \Pi_{\beta_C}(i) \quad (4)$$

where, M_{α_C} and M_{β_C} are original and distorted color videos, NoF_C is the number of frames in the videos. The comparison between the M s of the original and distorted color videos (i.e., $M(\alpha_C, \beta_C)$) is measured as follows:

$$M(\alpha_C, \beta_C) = \frac{M_{\alpha_D} - M_{\beta_D}}{NoF_C} \quad (5)$$

$M(\alpha_C, \beta_C)$ is normalized using the NoF_C to provide consistency for different color video pairs.

c) Structure of a Color Video

For the purpose of computing the structure of a color video in this work, contours in the frames are detected by the Canny edge detection technique, which is utilized to set edge pixels to 1. Every frame of the video is tallied for this purpose with the number of pixels set to 1 [14]. Then, the measured structure values are summed up to compute the structure measurements throughout the original or distorted color videos, as follows:

$$C_{\alpha_C} = \sum_{n=1}^{NoF_C} \delta_{\alpha_C}(n) \quad (6)$$

$$C_{\beta_C} = \sum_{n=1}^{NoF_C} \delta_{\beta_C}(n) \quad (7)$$

Here, $\delta(n)$ is the number of edge pixels in the n^{th} frame. The total of the values counted is then normalized considering the NoF to provide consistency across different videos, as follows:

$$C(\alpha_C, \beta_C) = \frac{C_{\alpha_D} - C_{\beta_D}}{NoF_C} \tag{8}$$

d) UF_{VQ}

Considering the integration of spatial resolution, motion, and structure of a color video, UF_{VQ} is formed as follows:

$$UF_{VQ} = S_C M(\alpha_C, \beta_C) C(\alpha_C, \beta_C) \tag{9}$$

where, S_C is the spatial resolution of a color video.

2) *Development of UF_{DP}*

The spatial resolution, relative depth, and aerial perspective of a depth map video which are envisaged as significant on the depth perception part of the 3D video QoE are exploited to form UF_{DP} . The effects of spatial resolution, relative depth, and aerial perspective of a depth map video on the 3D video QoE and how they are measured and/or integrated in UF_{DP} will be discussed in the following sub-sections.

a) *Spatial Resolution of a Depth Map Video*

Similar to the spatial resolution of a color video, the spatial resolution of the depth map video part of a color + depth map 3D video representation form is likewise a factor for describing the HVS-related 3D QoE. In order to find the most ideal adaption parameters utilizing UF_{DP} , this key factor in the suggested framework should be in line with content-related circumstances that have an impact on the HVS-related 3D QoE. Therefore, the relative depth and aerial perspective of a color video are considered as content related contexts in UF_{DP} .

b) *Relative Depth of a Depth Map Video*

The relative depth is connected with the depth level variations in depth map gray level values [15]. Standard deviation [16] is used to measure the relative depths of the original and distorted depth maps in this study as follows:

$$\sigma_{\alpha_D} = \sum_{y=1}^{NoF_D} \sqrt{\frac{\sum_{k=1}^{n_D} (x_{\alpha_D k}^y - \mu_{\alpha_D k}^y)^2}{n}} \tag{10}$$

$$\sigma_{\beta_D} = \sum_{y=1}^{NoF_D} \sqrt{\frac{\sum_{k=1}^{n_D} (x_{\beta_D k}^y - \mu_{\beta_D k}^y)^2}{n}} \tag{11}$$

where, σ_{α_D} and σ_{β_D} are the relative depths of the original and distorted depth maps, respectively. x_{α_D} and x_{β_D} represent each pixel depth value in the original and distorted frames, respectively. μ_{α_D} and μ_{β_D} are the mean of the pixel depth values in the original and distorted depth map frames, respectively. n_D is the number of pixels in the original or distorted depth map frame, which is equal to width \times height of the frame.

The individual relative depth calculations for each frame are integrated together and normalized by NoF_D , which represents the number of frames in the depth maps, to determine the depth maps of the original and distorted depth maps (i.e., $B(\alpha_D, \beta_D)$) as follows:

$$B(\alpha_D, \beta_D) = \frac{\sigma_{\alpha_D} - \sigma_{\beta_D}}{NoF_D} \quad (12)$$

c) *Aerial Perspective of a Depth Map Video*

When the objects and background provide the appearance of being at various depth levels, the brightness contrast difference that is linked with aerial perspective arises [17]. The Median Absolute Deviation (MAD) differences between the original depth map video and the warped depth map are used to calculate the overall luminance contrast difference between the two:

$$\theta_{\alpha_C} = \sum_{y=1}^{NoF_C} \sum_{k_C=1}^{n_C} \frac{(t_{\alpha_C k_C} - med(t_{\alpha_C}))}{n_C} \quad (13)$$

$$\theta_{\beta_C} = \sum_{y=1}^{NoF_C} \sum_{k_C=1}^{n_C} \frac{(t_{\beta_C k_C} - med(t_{\beta_C}))}{n_C} \quad (14)$$

where, θ_{β_D} and θ_{α_C} are the MADs of the distorted and original depth map videos, respectively. t_{α_C} and t_{β_C} represent each luminance value in the original and distorted frames, respectively. $med(t_{\alpha_C})$ and $med(t_{\beta_C})$ are the median of the luminance values in the original and distorted depth map video, respectively. n_C is the number of pixels in each original or distorted depth map video, which is equal to width \times height of the frame. The MADs are calculated for each frame individually to have uniform MADs in the frames. Then, the calculated MADs for the frames are integrated together to calculate the MADs across the depth map videos. NoF_C represents the number of frames in the depth map videos. Using the MAD measures in (13) and (14), the $A(\alpha_C, \beta_C)$ is measured as follows:

$$A(\alpha_C, \beta_C) = \frac{\theta_{\alpha_C} - \theta_{\beta_C}}{NoF_C} \quad (15)$$

d) UF_{DP}

Considering the integration of spatial resolution, relative depth, and aerial perspective of a depth map video, UF_{DP} is designed as follows:

$$UF_{DP} = S_D B(\alpha_C, \beta_C) A(\alpha_C, \beta_C) \quad (16)$$

3) Adaptation Decision Taking Process

The color + depth map videos can be encoded separately at various spatial resolutions as illustrated in Fig. 1 with the support of SVC.

Determining the optimal color + depth map video spatial resolution combinations require a trade-off between the 3D video QoE and rate-distortion performance. The objective of the adaptation decision taking process is to find the optimal color + depth map spatial resolution combinations that keep the 3D video QoE at a high level while satisfying a set of constraints (e.g., display size, network bandwidth, etc.). The optimum combinations are determined by solving the given optimization problem below:

$$\begin{aligned} &\text{Maximize : } \{UF_{CD}\} \Rightarrow \text{Optimization Constraint} \\ &\text{Subject to : } \{\text{bitrate} \leq \text{target bit rate; width} \leq \text{display width; height} \leq \text{display height}\} \\ &\Rightarrow \text{Limitation Constraint} \end{aligned}$$

Here, the optimization constraint relies on the maximization of the UF_{CD} . Bit rate is the bit rate of the 3D video, target bit rate represents the network bandwidth. Moreover, width and height present the horizontal and vertical dimensions of the encoded 3D videos, respectively. The width and height are the horizontal and vertical resolutions of the display, respectively.

A. Adaptation Module

The adaptation module presented in Fig. 3 performs the adaptation process considering the adaptation parameters determined by the adaptation decision taking module.

4 Results and Discussions

In order to derive results with the proposed framework, each of the color + depth map videos of the 3D videos including; Advertisement, Ballet, Breakdance, Chess, Football, Farm, and Windmill are encoded with four spatial layers (i.e., CIF, SD, HD, and Full HD) and two Medium Grain Scalability (MGS) layers [10] for each spatial layer with Quantization Parameters of 30, 35, and 40 using JSVM 9.13.1 [18]. Then, UF_{CD} s for all of these encoded 3D videos are computed. After that, the target network bandwidths = {700, 1000, 2000, and 2500} kb/s, display width = 1920 pixels, and display height = 1080 pixels are considered as constraints for enabling constraints on the adaptation decision taking parameters. Then, from the encoded 3D videos,

color + depth map videos with bit rates lower than the combined bandwidth of each target network are retrieved. Candidate adaptation parameters are the characteristics of these retrieved videos. The parameters that maximize UF_{CD} are then determined among these candidate parameters for each target bandwidth of network. These parameters refer to the adaptation decision taking results of the proposed framework. Using the adaptation decision taking results, adaptation experiments are carried out. Then, using Video Quality Metric (VQM), a widely accepted and defined video quality evaluation metric in the literature [18], the color + depth map movies of the converted 3D videos are evaluated. VQM values are between 1 and 5. When the VQM value is 1, the video quality is low; when it is high, the video quality is extremely high. The average VQM values are calculated utilizing these VQM values after the VQM values for the color + depth map videos are separately derived.

The advertisement, ballet, chess, and football 3D movies' outcomes employing the suggested framework are shown in Table 1. As can be observed from the table, all of the converted films' average VQM values for color video spatial resolution, depth map spatial resolution, and bit rate are rather encouraging given the network bandwidth restrictions. For instance, the adapted advertisement video's adaptation decision-taking outcomes when the network bandwidth is 1000 kbps are as follows: the color video spatial resolution is SD, the depth map spatial resolution is CIF, the adapted 3D video's bit rate is 973 kbps, and the resulting VQM value is 4.87.

Table 1 Results derived using the proposed framework

3D videos	Constraints and derived results				
	Target network bandwidth (kb/s)	Color video spatial resolution	Depth map spatial resolution	Bit rate (kb/s)	Average VQM
Advertisement	700	CIF	CIF	677	4.75
	1000	SD	CIF	973	4.87
	2000	HD	SD	1984	4.90
	2500	Full HD	HD	2332	4.94
Ballet	700	SD	CIF	694	4.76
	1000	SD	CIF	981	4.83
	2000	HD	SD	1979	4.92
	2500	Full HD	Full HD	2478	4.98
Chess	700	CIF	CIF	662	4.78
	1000	SD	SD	976	4.81
	2000	HD	SD	1894	4.86
	2500	Full HD	HD	2314	4.90
Football	700	SD	SD	681	4.71
	1000	HD	SD	984	4.75
	2000	HD	HD	1923	4.88
	2500	Full HD	HD	2322	4.93

5 Conclusions and Future Work

Given that the spatial resolution plays a significant role in the perception of 3D video quality on HVS, a framework for 3D video adaption has been created in this study. While developing this framework, content-related factors, such as the motion and structure of a color movie and the relative depth and aerial perspective of a depth map, were taken into consideration. The performance evaluation results show that, in the event that a particular set of requirements has been met and the 3D video QoE has been maintained at an ideal level, the proposed framework has been quite effective in determining the best spatial resolution combinations for the color + depth map 3D video representation. Future implementations of the proposed framework will also incorporate contextual elements to help Future Networks' smart service management even further.

References

1. Hewage C, Ekmekcioglu E (2020) Multimedia quality of experience (QoE): current status and future direction. *Future Internet* 12(7):121. <https://doi.org/10.3390/fi12070121>
2. Nur G, Kodikara Arachchi H, Dogan S, Kondoza AM (2012) Advanced adaptation techniques for improved video perception. *IEEE Trans Circ Syst Video Technol* 22:225–240
3. Ramakrishna M, Fernandes RC, Karunakar AK (2017) Estimation of adaptation parameters for scalable video streaming over software defined networks. *Procedia Comput Sci* 115:715–722
4. Ginimav I (2020) Live streaming architectures for video data—a review. *J IoT Soc Mob Anal Cloud* 2(4):207–215
5. Raj JS, Vijesh Joe C (2021) Wi-Fi network profiling and QoS assessment for real time video streaming. *IRO J Sustain Wirel Syst* 3(1):21–30
6. Mysirlidis C, Dagiuklas T, Politis I, Ekmekcioglu E, Dogan S, Kotsopoulos S (2014) Quality evaluation of 3D video using colour-plus-depth & MDC over IP networks. *IEEE 3DTV*
7. Lie W-N, Lu Y-H (2015) Fast encoding of 3D color-plus-depth video based on 3D-HEVC. In: *International conference on image processing*
8. Malekmohamadi H, Fernando A, Kondoza A (2014) A new reduced reference metric for color plus depth 3D video. *J Vis Commun Image Represent* 25(3):534–541
9. Peng WH, Zao JK, Huang HT, Wang TW, Huang LS (2008) A rate-distortion optimization model for SVC inter-layer encoding and Bitstream extraction. *J Visual Commun Image Represent* 19:543–557
10. Quality of service enhancement for multimedia applications using scalable video coding. In: *Second international conference on intelligent computing and control systems (ICICCS)*
11. Fleet DJ, Wiess Y (2006) Optical flow estimation in Paragios. In: *Handbook of math. models in comp vision*. Springer
12. Nur G, Dogan S, Kodikara Arachchi H, Kondoza AM (2011) Extended VQM model for predicting 3D video quality considering ambient illumination context. In: *IEEE 3DTV conference: the true vision - capture, transmission and display of 3D video*, Antalya, Turkey, 16–18 May 2011.
13. Shi J, Tomasi C (2004) Good features to track. In: *IEEE conference on computer vision and pattern recognition*, Seattle, pp 593–600
14. Grigorescu C, Petkov N, Westenberg MA (2004) Contour and boundary detection improved by surround suppression of texture edges. *Image Vis Comput* 22:609–622

15. Nur Yilmaz G, Battisti F (2018) Depth perception prediction of 3D video for ensuring advanced multimedia services. In: IEEE 3DTV conference: the true vision - capture, transmission and display of 3D video, Stockholm-Helsinki, Sweden-Finland, 3–5 June 2018
16. Hassani H, Howell G (2010) A note on standard deviation and standard error. *Teach Math Appl* 29(2):108–112
17. Nur Yilmaz G (2018) Depth perception prediction of 3D video QoE for future internet services. In: IEEE 32nd international conference on information networking, Chiang Mai, Thailand, 10–12 January 2018
18. JSVM (n.d.) 9.13.1 Software, downloaded from CVS Server, garcon.ient.rwth-aachen.de/cvs/jv

Improved Lightweight Cryptography Authentication Based Secure Data Transmission in IoT Networks



S. Hariprasad and T. Deepa

Abstract The internet of things (IoT) devices is becoming more common in our daily lives. At the same time, attacks on these devices lead people's lives in danger and cost much money. Man-in-the-Middle (MITM) attacks have been primarily examined, especially in IoT networks. In this paper, the framework is constructed using the smart environment monitoring system (SEMS), created with the help of ten nodes, and communicated through the message queuing telemetry transport (MQTT) protocol. An enhanced lightweight encryption algorithm (ELWEA) is proposed to mitigate the MITM attack on IoT devices. The ELWEA comprises (i) key scheduling and (ii) encryption and decryption. The performance metrics of the proposed method provide less time complexity of 10.5 ns compared to the literature. Substantial cryptography methods are applied to maintain the secrecy of sensor data. The proposed method protects against the MITM attack and works against a wide range of typical applications.

Keywords Internet of things (IoT) · Man-in-the-middle attack · Message Queuing Telemetry Transport (MQTT) · Lightweight encryption · Key scheduling

1 Introduction

The internet of things (IoT) has become an increasingly influential technology where several physical objects are connected to the internet to exchange data and gather information. Secure communication in IoT is necessary to secure infrastructure and services [17]. IoT components based on 6A's type are anything, any device, anybody, any service, any business, any path, any network, anywhere, anytime, or any connect. In 2025 more than 30.9 Billion IoT devices worldwide will be connected, according

S. Hariprasad (✉) · T. Deepa
SRM Institute of Science and Technology, Kattankulathur, Tamilnadu, India
e-mail: hs6512@srmist.edu.in

T. Deepa
e-mail: deepat@srmist.edu.in

to the analysis by Statista [21]. Connectivity across infrastructure and services is not possible now without the development of IoT. Cyber security looks at the overall safety of data in software and its systems. At the same time, IoT security has been left to its own devices.

The three levels of security aspects are (1) Design security: The security framework must be built from the basic level with security as an essential and integrated component. (2) Hardware security: The devices must be made tamper-resistant so that users can use them in challenging environments, (3) Data security: Cryptography is required to prevent and protect IoT data. It helps to safeguard IoT privacy and establish trust between end devices and users.

Data security has many challenges, such as handling enormous data and managing the many IoT devices. For more extensive data, the security algorithms such as advanced encryption standards (AES) [15] and data encryption standards (DES) [18] consume more time to secure the data in IoT networks.

Therefore, lightweight cryptography is needed, which can support a smaller number of 32 or 64 bits of data with more minor keys up to 32 or 64 bits. Lightweight cryptography is classified into block ciphers, hash functions, message authentication codes, and stream ciphers. The block cipher-based lightweight cryptographic primitives have performance advantages over other cryptographic standards based on the power, energy consumption, latency and throughput. Table 1 describes the various block cipher techniques with their block and key sizes.

Many existing works of literature used a heavy computation protocol such as advanced message queuing protocol (AMQP) to transfer the sensor data to the cloud. The IoT devices are resource-constrained to reduce the communication overhead. The lightweight protocol called message queuing telemetry transport (MQTT) is a protocol for publishing/subscribing messaging to client-server communication [7]. It is designed as light and easy to implement. These protocols can be used in restricted communications environments in the machine to machine (M2M), and less network bandwidth is needed.

The various security aspects, application-related and compared protocols with MQTT are discussed below. In [16], an IoT-based decision device for intelligent irrigation systems was developed using two protocols such as MQTT and HTTP. These two protocols communicate between the crop field sensor data and the cloud. Initially, the MQTT broker was created using a Mosquito light on a computer to

Table 1 Block cipher comparison

Block cipher	Type of construction	Block size	Key size
HIGHT [11]	Feistel network	64	128
PRESENT [3]	Substitution and permutation network	64	80/128
KTANTAN [5]	Stream/Block	32/48/64	80
LED [8]	Substitution and permutation network	80	64/128
AES-128 [15]	Feistel network	128	128
DES [18]	Balanced Feistel network	64	64

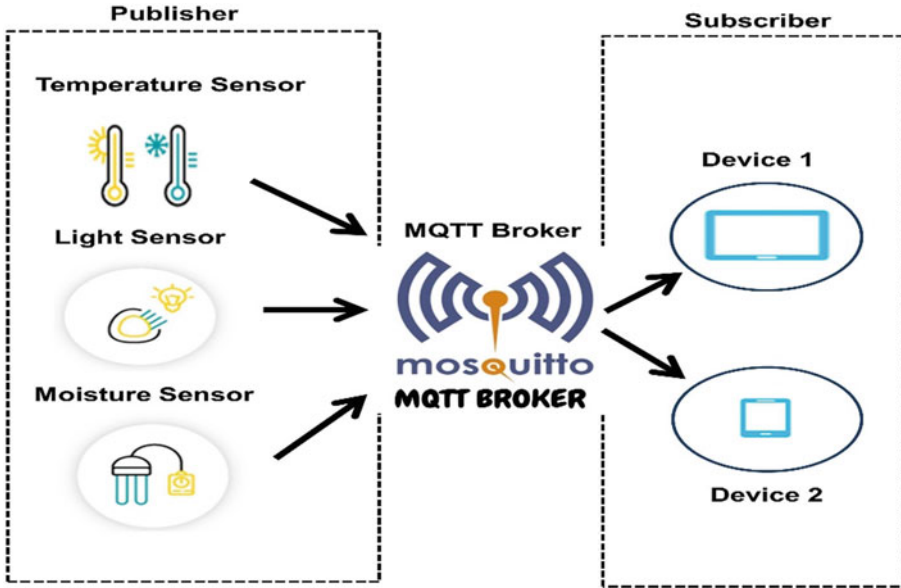


Fig. 1 General publish/subscribe architecture of the MQTT protocol

transfer the sensor data from crop yield. The same crop field sensor data is sent to the server using HTTP. MQTT is preferred over HTTP because it is lightweight and has a small message size, requiring less bandwidth. Figure 1 shows the general publish/subscribe architecture of the MQTT protocol.

The numerous benefits that have been found after the study of IoT with MQTT protocol are discussed in [14]. In addition, it has enabled components of smart grids to communicate and process data on a real-time basis. A multi-tier edge computing model was developed [22] using MQTT as a remote broker in a fog node, and the primary broker was placed in the cloud. As part of this [19], MQTT messages are exchanged between several nodes, including MQTT Control Packets. There are three principal parts in an MQTT control packet: the fixed header, the variable header, and the payload. MQTT ensures that messages are delivered when the networks are unreliable or unresponsive, and acknowledgement systems allow both parties to know if data has been received correctly.

2 Related Works

In recent years, edge computing has gained popularity among cloud service providers, such as Amazon, Google, and Microsoft [1], due to its low latency IoT connectivity and large data processing capacity. The control mechanisms, certificate service providers (CSP), enforce security policies to meet the demands of IoT applications.

The MQTT protocol is weak, making it vulnerable to attacks. According to previous literature, Attacks can exploit MQTT applications. The study [2] eventually led to a more robust protocol development for IoT-related applications. The usage of the MQTT protocol in IoT with encryption the data transfer-related applications are demonstrated in [9]. The fiestal network structure divides a 64-bit input block into two 32-bit output blocks. Recursive Pared Parity uses logical XOR to decode data. Transformation of Bits encrypts and decrypts data via bit swapping using recursive positional substitution on prime-nonprime of a cluster, bits are exchanged, and all bits in between constitute cypher text. Merging these two sub-blocks creates 64-bit ciphertext, as discussed in [6]. The insubstantial nature of the MQTT protocol makes the transfer of information at high speed. Hence, this protocol has been various applications related to IoT, Wireless sensor networks (WSN), and Machine to machine (M2M) communications. Data from many heterogeneous IoT devices with attack detection modelling named the SENMQTT-SET has been discussed [20]. The SENMQTT-SET features are evoked by an ensemble statistical multi-view cascade feature generation algorithm, detecting the attack in less time and improving accuracy.

A study of alternative media transport mechanisms in IoT networks is conducted in [10] using constrained application protocol (CoAP) and MQTT-SN for media propagation in low power lossy networks (LLNs). Computer vision neural network using lightweight convolution neural network provides the point estimation is the specific domain which estimates the uncertainty, improve prediction and forecasting has been discussed in [23]. Ordering and re-sending the lost message are essential for reliable message communication in IoT networks [12]. This system [13] monitors air quality and noise pollution in a specified geographic area and securely transmits data over the network to overcome the security issues in the IoT.

The application layer protocols are vulnerable to three major attacks: denial of service (DoS), distributed denial of service (DDoS), and man in the middle (MitM). In DoS, an attacker can make it unavailable to its intended users by disrupting the normal functioning of the IoT device. In DDoS, the intended effect is to overwhelm the infrastructure of the targeted IoT devices by causing abnormal traffic patterns. The MitM attack mitigates the user's ability to know that a third party is interfering with their communication.

2.1 Problem Statement

Nowadays, any data or information can be transmitted worldwide in milliseconds, but the data sent over the cloud can be hacked or spoofed by others in the interconnected devices. So, nodes must securely transfer the information to the user, and data must encrypt the information or message. The embedded devices need fewer hardware components, less area, cost, and low power consumption. Hence, in the software aspect, the secured data can be done with the help of lightweight cryptography.

The nodes in a monitoring system are resource-constrained devices on the heavy computation cryptography algorithm, leading to computational complexity and time decay. The proposed Enhanced Lightweight Encryption Algorithm (ELWEA) proves that eavesdropping has been a complicated task and cannot hijack the nodes. The main aim of the proposed method is to use lightweight cryptography because it requires a small footprint, less computational complexity and can be communicated with the secured data over the MQTT protocol.

2.2 Objectives

The main research contribution is described as follows.

- A novel smart environment monitoring system (SEMS) framework is created with ten resource-constrained nodes.
- A novel ELWEA is applied to protect the heterogeneous sensor data for secure transfer.
- Finally, an assessment is done to show the performance analysis of the proposed scheme, and it is compared with the existing system metrics such as computational time complexity.

3 Methodology

The complete system methodology comprises two (a) Framework of SEMS, (b) proposed ELWEA with key scheduling and encryption and decryption.

3.1 Framework of SEMS

The sensed data of all nodes is fetched and communicated to the fog server using the MQTT protocol. The collected sensor information is encrypted using the proposed ELWEA as illustrated in Fig. 2, and transferred through the MQTT protocol.

The secure sensor data acts as the publisher, and the actuators will serve as subscribers. The cloud works as a broker MQTT, which will check the publisher and subscriber's identity and topics to send and receive the data. There are more chances to prevail against more threats in the fog layer for the data. These threats could be controlled and precise from the forging of collected data. Sometimes these threats can destroy the devices by the following attacks: (a) node trap like MITM. In this work, the ELWEA system is used for more real-time security without modifying any existing service architecture. The received encrypted data will be decrypted at actuators using a micro python script with the help of the key of each node.

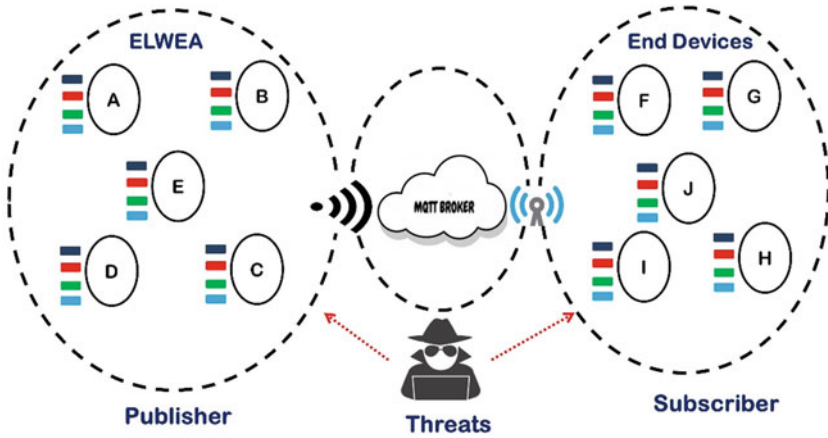


Fig. 2 Framework for smart environment monitoring system (SEMS)

3.2 Enhanced Lightweight Encryption Algorithm (ELWEA)

The proposed robust hybrid cipher is unbreakable and compatible with lightweight devices. The ELWEA is used to secure the sensor data of the nodes and transmit it to the receiver for actuating the end devices. The proposed ELWEA system consists of a lightweight cipher. It consists of (i) Key scheduling and (ii) Encryption/Decryption.

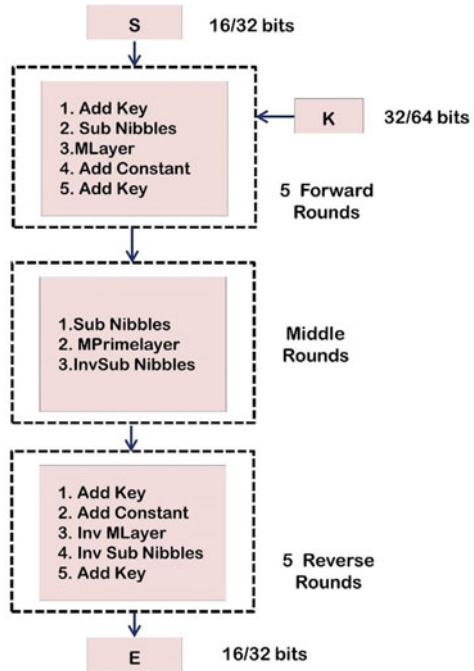
Key Scheduling

In key scheduling, the PRINCE algorithm [4] generates the key and is a lightweight cipher and a low latency block code. PRINCE Key has an alpha reflection property that allows the reuse of encoded information for decryption. This method uses resources of low computational cost, and the encryption time is less than one clock cycle. The block size is 64 bits, the key length is 128, and the number of iterations will be 11–13. PRINCE is best suited for the requirement that it uses a 32-bit key for securing 32/64-bit data to make it very secure. The key scheduling generates an individual private key for personal node data.

The length of data (N, t) will increase the size of bits for a secret stake in the encryption/decryption process. Here N is denoted as the number of nodes P_N , and ranges from 1, 2...10 and 't' is used for splitting the difference between plain text or key. If plain text is 'p', then the number of original bits will append 0 at the end until the data length is 16 bits. If a key is 'k', the number of actual bits will append 0 at the back until the data length is 32 bits.

For example: The secret share of length for node 1 of plain text $data(1, 'p')$ will be [1,0,0,0,0,0,0,0] and key [1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0].

Fig. 3 Proposed flow diagram for ELWEA



Encryption and Decryption

The encryption/decryption is based on a light encryption device (LED), and block cipher is a lightweight algorithm with an input block length of 32 bits and the key size is 64 bits which are in multiples of 4 [8]. The LED block cipher (P) uses a prince key $_K$ to encrypt all rounds, and a 32-bit plain text is encoded. The LED implementation provides compact hardware in gate count, and the power consumption is deficient. Figure 3 shows the proposed flow diagram ELWEA.

Table 2 shows the added constant value for each round. The plain text block of 32-bit size P is added with the PRINCE Key generation $_K$ with XOR operation, and results are stored in P. Then, the round constants are added to the consecutive 4×2 matrix cells containing eight nibbles of P. Then, subcells are added from the left to right of the cells and substituted with S-Box. This block is divided into half, the first half is fed from the left, the next is provided from the right, and the rows are shifted for encrypting. Here the linear feedback shift keying method is used to move the bits. Finally, the column is mixed once again for shuffling the bits to get a more robust encrypted code, and each constant value is combined using a linear feedback shift register and updated with the new P-value. The above operations will be repeated for eight rounds. Then the key ($_K$) undergoes XOR operation with the new P-value for 12 rounds to get the final encrypted E value.

Table 2 Add round constant value

RC0	0X00	0X00	0X00	0X00	0X00	0X00	0X00	0X00
RC1	0x31	0x91	0xa8	0xe2	0x30	0x07	0x37	0x44
RC2	0x4a	0x90	0x83	0x22	0x92	0xf9	0x13	0x0d
RC3	0x80	0xe2	0xaf	0x89	0xce	0xe4	0xc6	0x98
RC4	0x54	0x82	0x12	0x6e	0x83	0x0d	0x31	0x77
RC5	0xeb	0x45	0x66	0xfc	0x43	0x9e	0xc0	0xc6
RC6	0xe7	0x8f	0xf4	0x87	0xdf	0x59	0xc5	0x1b
RC7	0x58	0x48	0x80	0x15	0x1f	0xca	0x34	0xaa
RC8	0x8c	0x28	0x3d	0xf2	0x52	0x23	0xc3	0x45
RC9	0x46	0x5a	0x11	0x59	0x0e	0x3e	0x16	0xd0
RC10	0x3d	0x5b	0x3a	0x99	0xac	0xc0	0x32	0x99
RC11	0x0c	0xca	0x92	0x7b	0x9c	0xc7	0x05	0xdd

In the encryption process, initially, the node is connected to the MQTT protocol using the broker I.P. address and port number. Each node sensor data is encrypted using ELWEA and sent to the MQTT broker with specific topics. The end-user can receive the data using the particular case. In the decryption process, initially, the node is connected to the MQTT protocol using the broker I.P. address and port number. The encrypted data in the MQTT broker can be decrypted using the ELWEA decryption process, and the original data has been sent to the actuators for other operations.

4 Results and Discussion

The experimental setup consists of 10 nodes using \$ESP8266\$ and is programmed using a micropython script. The ESP8266 is a limited resource-constrained device with a processor L106 32-bit RISC and RAM of 32 K. The subscriber is the actuator who receives all the data on a specific topic. Out of ten devices, five are taken for a publisher who sent the sensor data (A, B, C, D, E), and five are taken for the actuators who receive the sensor data from the sender (F, G, H, I, J). Figure 4 shows the attack illustration on the IoT nodes.

Figures 5 and 6 shows the data transmitted and received between sender nodes to the actuator nodes without encryption and the proposed ELWEA algorithm.

The proposed encryption algorithm has been deployed in the nodes A, B, C, D, and E. Sensor scripts of all esp8266 node was developed using Micro Python in Thonny IDE Editor. Here ESP 8266 MQTT uses the three process actuator nodes: subscriber, sensor nodes, publisher, and a broker. Initially, the broker system runs on the data centre to receive the data from the subscriber system, and the publisher uses ESP8266.

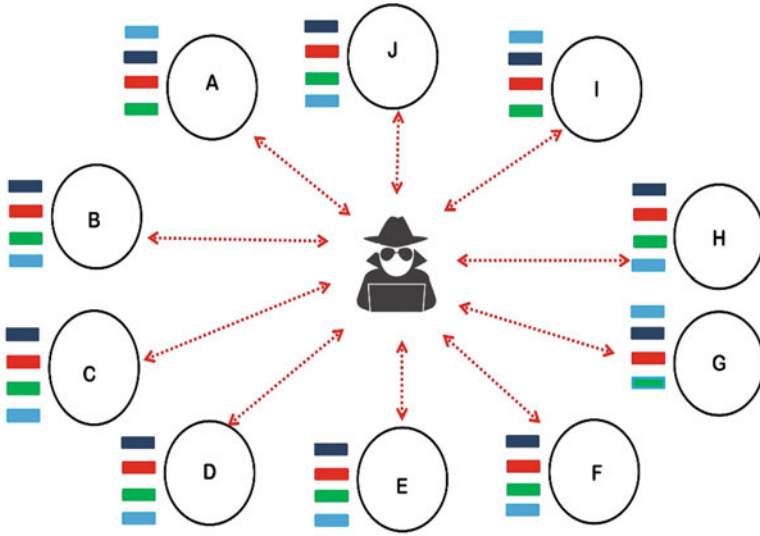


Fig. 4 Attack illustration on IoT nodes

```

Connected to MQTT Broker!
Received `[57, 54, 0, 0, 0, 0, 0, 0]` from `sens/room5/temp` topic
Received `[56, 57, 0, 0, 0, 0, 0, 0]` from `sens/room5/light` topic
Received `[57, 51, 0, 0, 0, 0, 0, 0]` from `sens/room5/motion` topic
Received `[57, 57, 0, 0, 0, 0, 0, 0]` from `sens/room4/temp` topic
Received `[56, 49, 0, 0, 0, 0, 0, 0]` from `sens/room4/light` topic
Received `[56, 51, 0, 0, 0, 0, 0, 0]` from `sens/room4/motion` topic
Received `[57, 49, 0, 0, 0, 0, 0, 0]` from `sens/room4/temp` topic
Received `[49, 49, 50, 0, 0, 0, 0, 0]` from `sens/room4/light` topic
Received `[56, 55, 0, 0, 0, 0, 0, 0]` from `sens/room4/motion` topic

```

Fig. 5 Message received to the actuators without encryption

```

Connected to MQTT Broker!
Received `[9, 245, 194, 6, 54, 217, 78, 105]` from `sens/room4/temp` topic
Received `[79, 85, 140, 248, 105, 93, 218, 84]` from `sens/room4/light` topic
Received `[115, 160, 146, 209, 135, 238, 218, 210]` from `sens/room4/motion` topic
Received `[8, 25, 229, 63, 136, 192, 123, 53]` from `sens/room5/temp` topic
Received `[172, 5, 118, 92, 106, 229, 210, 64]` from `sens/room5/light` topic
Received `[209, 155, 158, 214, 75, 25, 5, 51]` from `sens/room5/motion` topic
Received `[246, 82, 10, 23, 9, 132, 134, 231]` from `sens/room2/temp` topic
Received `[246, 82, 10, 23, 9, 132, 134, 231]` from `sens/room2/light` topic
Received `[5, 229, 63, 162, 69, 177, 203, 140]` from `sens/room2/motion` topic

```

Fig. 6 Message received to the actuators with proposed ELWEA

Table 3 Comparison with various algorithms

Method	Block (bits)	Key (bits)	Time
AES-128	128	128	1.988 s
DES	64	64	1.022 s
HIGHT	64	128	106 us
PRESENT-80	64	80	7.2 us
KATAN32	32	80	12.245 us
PROPOSED	32	32/64	10.499 us

Table 4 Proposed ELWEA performance metrics

Method	Encryption time (ns)	Decryption time (ns)	Encryption throughput (kbps)	Decryption throughput (kbps)
Proposed	10.5	17.02	20.12	18.36

Table 3 shows the various comparison like block size and key in terms of bits and time in seconds compared to the proposed method. Table 4 shows the Proposed ELWEA performance metrics such as encryption time (ns), decryption time (ns), encryption throughput (kbps), and decryption throughput (kbps). The table shows that the proposed method's performance metrics are better than other lightweight cryptography algorithms.

5 Conclusion

This paper gives an overview and explains the concept of a manually designed MITM attack on MQTT-based IoT devices. The framework of SEMS has been designed with ten nodes using the MQTT protocol. An ELWEA was proposed to protect the data from spoofing attacks. The five nodes are considered for both scenarios, such as without encryption and ELWEA. The performance metrics of encryption time, decryption time, encryption throughput and decryption throughput have been computed and tabulated. Compared to the other literature, the proposed method provides less time complexity of 10.5 ns. Future work could include exploring different attack mechanics for MITM attacks without using a WiFi Pineapple to show more attack possibilities.

References

1. Ahmad T et al (2021) Extending access control in AWS IoT through event-driven functions: an experimental evaluation using a smart lock system. *Int J Inf Secur.* <https://doi.org/10.1007/s10207-021-00558-3>
2. Akhtar S, Zahoor E (2021) Formal specification and verification of MQTT protocol in PlusCal-2. *Wirel Pers Commun* 119(2):1589–1606. <https://doi.org/10.1007/s11277-021-08296-4>
3. Bogdanov A et al (2007) PRESENT: an ultra-lightweight block cipher. In: *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Springer, Heidelberg, pp 450–466. https://doi.org/10.1007/978-3-540-74735-2_31
4. Borghoff J et al (2012) PRINCE - a low-latency block cipher for pervasive computing applications. In: *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*. LNCS, vol 7658, pp 208–225. https://doi.org/10.1007/978-3-642-34961-4_14
5. De Cannière C et al (2009) KATAN and KTANTAN - a family of small and efficient hardware-oriented block ciphers. In: *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*. LNCS, vol 5747, pp 272–288 (2009). https://doi.org/10.1007/978-3-642-04138-9_20
6. Chatterjee R et al (2019) Design of lightweight cryptographic model for end-to-end encryption in iot domain. *IRO J Sustain Wirel Syst* 1(04):215–224. <https://doi.org/10.36548/jsws.2019.4.002>
7. Edited by Andrew Banks and Rahul Gupta: MQTT Version 3.1.1. <http://docs.oasis-open.org/mqtt/mqtt/v3.1.1/os/mqtt-v3.1.1-os.html>
8. Guo J et al (2011) The LED block cipher. In: *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*. LNCS, vol 6917, pp 326–341. https://doi.org/10.1007/978-3-642-23951-9_22
9. Gupta V et al (2021) MQTT protocol employing IOT based home safety system with ABE encryption. *Multimed Tools Appl* 80(2):2931–2949. <https://doi.org/10.1007/s11042-020-09750-4>
10. Herrero R (2020) MQTT-SN, CoAP, and RTP in wireless IoT real-time communications. *Multimed Syst* 26(6):643–654. <https://doi.org/10.1007/s00530-020-00674-5>
11. Hong D et al (2006) HIGHT: a new block cipher suitable for low-resource device. In: *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*. LNCS, vol 4249, pp 46–59. https://doi.org/10.1007/11894063_4
12. Hwang HC et al (2016) Design and implementation of a reliable message transmission system based on MQTT protocol in IoT. *Wirel Pers Commun* 91(4):1765–1777. <https://doi.org/10.1007/s11277-016-3398-2>
13. Khrijji S et al (2021) Design and implementation of a cloud-based event-driven architecture for real-time data processing in wireless sensor networks. *J Supercomput* 0123456789. <https://doi.org/10.1007/s11227-021-03955-6>
14. Kondoro A et al (2021) Real time performance analysis of secure IoT protocols for microgrid communication. *Futur Gener Comput Syst* 116:1–12. <https://doi.org/10.1016/j.future.2020.09.031>
15. Moradi A et al (2011) Pushing the limits: a very compact and a threshold implementation of AES. In: *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*. LNCS, vol 6632, pp 69–88. https://doi.org/10.1007/978-3-642-20465-4_6
16. Nawandar NK, Satpute VR (2019) IoT based low cost and intelligent module for smart irrigation system. *Comput Electron Agric* 162:979–990. <https://doi.org/10.1016/j.compag.2019.05.027>
17. Ray PP (2018) A survey on Internet of Things architectures. *J King Saud Univ Comput Inf Sci* 30(3):291–319. <https://doi.org/10.1016/j.jksuci.2016.10.003>
18. Satoh A, Morioka S (2003) Hardware-focused performance comparison for the standard block ciphers AES, Camellia, and triple-DES. In: *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol 2851, pp 252–266. https://doi.org/10.1007/10958513_20

19. Seoane V et al (2021) Performance evaluation of CoAP and MQTT with security support for IoT environments. *Comput Netw* 197:108338. <https://doi.org/10.1016/j.comnet.2021.108338>
20. Siddharthan H et al (2022) SENMQTT-SET: an intelligent intrusion detection in IoT-MQTT networks using ensemble multi cascade features. *IEEE Access* 10:33095–33110. <https://doi.org/10.1109/ACCESS.2022.3161566>
21. Stastica: Stastica. www.stastica.com
22. Veeramanikandan M, Sankaranarayanan S (2019) Publish/subscribe based multi-tier edge computational model in Internet of Things for latency reduction. *J Parallel Distrib Comput* 127:18–27. <https://doi.org/10.1016/j.jpdc.2019.01.004>
23. Vivekanandam B (2021) Speedy image crowd counting by light weight convolutional neural network. *J Innov Image Process* 3(3):208–222. <https://doi.org/10.36548/jiip.2021.3.004>

Modeling and Control of Induction Machine and Drive in the Combined Domain with New Chaotic Gorilla Troop Optimizer



Rahul Chaudhary and Souvik Ganguli

Abstract In this work, a new chaotic version of gorilla troop optimizer is developed. The position equation of the algorithm is modified with the help of chaotic maps. Around ten widely cited chaos maps, one-dimensional in nature, are considered to develop new chaotic algorithms. Two unimodal and three multi-modal test functions are employed to validate the efficacy of the proposed technique. A 50 hp induction motor model is also reduced and further its controller is designed utilizing the benefit of delta operator with the help of this proposed algorithm. Finally, a practical test system of induction motor drive is taken up for modeling and control action as well. The controller realization in both cases is carried out using approximate model matching framework. The convergence speed and accuracy of the proposed techniques are better as compared to the standard and latest methods. Thus, the results in all experimentations performed show great promise.

Keywords Gorilla Troop Optimizer (GTO) · Chaotic Gorilla Troop Optimizer (CGTO) · Induction motor and drives · Delta operator · Reduced order modeling · Controller synthesis

1 Introduction

Induction motors and drives are in wide usage in manifold industries owing to the fact that they are rugged and robust. They also require very less maintenance. Further they are also cheaply available [1].

But however their mathematical models mostly result in higher-order systems which are difficult to control. The higher-order controllers thereby obtained for these

R. Chaudhary · S. Ganguli (✉)
Department of Electrical and Instrumentation Engineering, Thapar Institute of Engineering and Technology, Patiala 147004, Punjab, India
e-mail: souvik.ganguli@thapar.edu

R. Chaudhary
e-mail: rchaudhary_me20@thapar.edu

systems may have more hardware requirements and might not be realizable as well. Hence order reduction of the original system is necessary [2].

Modelling and control in the combined domain of delta operator instead of regular discrete-time systems has several advantages. Some of them worthy to this work are reported here. The discrete-time operators cannot handle high speed digital data and thus results in numerical ill-conditioning. The delta domain operator however takes care of high speed computation with increased numerical stability. Further, the discrete-delta systems at high sampling limit convergences to its continuous-time counterpart thus giving a unified modelling approach [3].

The use of metaheuristic techniques is quite common for modelling and control of different systems utilizing the delta operator framework [4, 5]. Of late chaotic metaheuristic techniques have become popular and applied to solve different engineering problems. There are several methods by virtue of which chaos can be embedded in metaheuristic techniques. Normally, the position update rule is varied chaotically to get improved performance. Moreover, some controlling parameters of any metaheuristic algorithm may be varied with chaos maps to bring about better performance. Even random numbers used in an algorithm may be replaced by chaos maps for better results [6].

Gorilla troop optimizer (GTO) is a recently developed metaheuristic method imitating the behaviour of gorillas [7]. In GTO, the position update equation can be modified using chaotic maps to better the performance of the parent technique. Some of the popular chaos features used in [8] are made use of in this work to develop new chaotic gorilla troop optimizer (CGTO). These algorithms are validated first of all with some higher-dimension test functions. Two types of test functions viz. the unimodal and the multi-modal are taken up for the study. Further experiments are carried out with the help of practical test systems. Mostly, literature records reports order reduction and control of permanent magnet synchronous motor drives in this combined domain of analysis employing some hybrid metaheuristic approaches [9, 10]. Here, the modelling and control of the induction motor and drive model is thus attempted in this paper with the help of these new CGTO methods.

The remaining paper is organized as per the details given. Section 2 deals with the procedure for order diminution and controller realization. Section 3 showcases the new chaotic gorilla troop optimizer obtained by position update rule. In Sect. 4, results of the test functions as well as the practical systems of 50 hp induction motor and drive models are presented. Section 5 gives the main inferences of the work conducted.

2 Problem Statement

Two steps are required to solve the problem presented in this paper. Model order reduction is the first step, while controller synthesis is the second. In the model reduction part, a second order model of fixed structure is assumed. There are four decision variables in this lower order model which are optimized. The second order

system with unknown parameters is compared with that of higher order model. A pseudo random binary sequence (PRBS) is considered as an input for both the above mentioned systems. Reduced system unknown parameters are determined by minimising the integral of the time-weighted absolute error (ITAE). The following are the restrictions that must be met in order to obtain this fixed-structured reduced system:

- Getting the dc gain to match
- Maintaining the minimal phase characteristics
- Guaranteeing stability.

The model reduction is carried out applying the delta operator framework to get the advantage of unifying discrete and continuous-time systems. The detailed mathematical expression of the order reduction process is available in [4]. A controller is proposed of the lower-order model. An approximate model matching (AMM) [11] controller synthesis technique is employed here. As exact model matching (EMM) [12] does not ensure the physical realization of the controller, AMM is preferred over EMM. The concept of reference model is utilized in the AMM to determine the unknown controller parameters. The closed loop response of the controlled plant is compared with the reference model having set design specifications. The square error is minimized to determine the controller parameters. The popularly used PID controller is used for the purpose. The step response of the controlled plant in the delta domain is finally matched with that of the reference test system. A novel chaos version of gorilla troop optimizer is proposed in Section 3, which uses the merit of AMM to tune controller parameters in the delta domain. A number of standard heuristic approaches, including the parent GTO method, are used for comparison. The step-by-step procedure for the controller design can be seen from [5].

3 Proposed Method

The artificial gorilla troop optimization (GTO) is a relatively new metaheuristic optimization technique. It is used to solve non-linear, non-convex, and non-smooth problems that involve multiple types of variables. The social intelligence of gorilla troops served as a source of motivation for the development of this algorithm. Operators based on gorilla behaviour are used in the GTO algorithm for simulated optimization operations (exploration and exploitation). In the exploration phase, three different operators were used: migration to an unknown location to increase GTO exploration. The second operator, a gorilla movement, improves the balance between exploration and exploitation. Migration towards a known location, the third operator in the exploration phase, significantly improves the GTO's ability to search for different optimization spaces. In the exploitations phase, on the other hand, two operators are used, which significantly improves the search performance. The detailed description of this algorithm along with its mathematical modelling can be seen from [7].

Table 1 Nomenclature of the different CGTOs

Nomenclature of algorithm	Variation in the algorithm	Chaotic map used
CGTO-01	Position	Chebyshev map
CGTO-02	Position	Circle map
CGTO-03	Position	Gauss map
CGTO-04	Position	Iterative map
CGTO-05	Position	Logistic map
CGTO-06	Position	Piece wise map
CGTO-07	Position	Sine map
CGTO-08	Position	Singer map
CGTO-09	Position	Sinusoidal map
CGTO-10	Position	Tent map

The position update in the exploitation part of the GTO method is chaotically varied using one-dimensional chaotic maps bringing about significant improvement in the outcome of the original GTO algorithm. The proposed nomenclatures for the algorithms are defined by chaotic gorilla troop optimizer (CGTO). The chaotic maps taken up for the study are taken from [8]. Accordingly, the proposed CGTO algorithms are numbered as per their chaotic map number, as discussed in Table 1.

Three basic tests are conducted to prove the efficacy of the suggested method. Two types of test functions (unimodal and multi-modal) are taken up to validate the performance of the proposed CGTOs. A higher order model of an induction motor is reduced to a second order model and then its PID controller is developed using the proposed method. Finally a practical induction motor drive model is reduced to synthesize an implementable PI controller applying the proposed approach. A unified domain approach has been carried out in experiments 2 and 3 respectively. The controllers in these experiments are realized using approximate model matching procedure.

4 Simulation Results

Experiment 1: To test the efficacy of the proposed methods, two types of test functions viz. unimodal and multi-modal are considered for experimentation. Two unimodal test functions while three multi-modal test benchmarks are taken for the study. The mathematical descriptions of these test functions are available in [13]. Their search domain and ideal optimum values are also standardized. F1 and F2 considered are unimodal functions while F3, F4 and F5 represent the multi-modal test functions. In each of these test functions, hundred decision variables are optimized. The population size and the maximum number of iterations are considered

as 30 and 500 for these test problems. This means that the number of function evaluations (NFE) is $30 \times 500 = 15,000$ which is quite competitive in terms of number of decision variables considered for the study.

The test functions are optimized using the proposed techniques. Further, the algorithms like Gorilla Troop Optimizer (GTO), Marine Predator Algorithm (MPA), Black Widow Optimization Algorithm (BWOA), Chimp Optimization Algorithm (ChOA), Chameleon Swarm Algorithm (ChSA), Dingo Optimization Algorithm (DOA), Slime Mould Algorithm (SMA), Wild Horse Optimizer (WHO), and War Strategy Optimizer (WSO) are applied for comparison. Sample convergence plots of the test functions F2 and F3 are displayed in Figs. 1 and 2 respectively corresponding to methods CGTO-05 and CGTO-01.

It is clearly indicative from Figs. 1 and 2, that the proposed CGTO approaches provide better convergence speed and accuracy than the parent GTO algorithm as well as a host of techniques used for comparison.

Experiment 2: A 50 HP induction motor [14] is also considered for the study. The machine model is 5th order as given below by its transfer function.

$$G_1(s) = \frac{2085s^3 + 511000s^2 + 3.081e07 s + 4.676e09}{s^5 + 397.9s^4 + 184800s^3 + 4.151e07s^2 + 3.408e09 s + 4.076e10} \quad (1)$$

The corresponding delta transformed model represented in γ -domain with a sampling time of 0.0025 s is denoted by

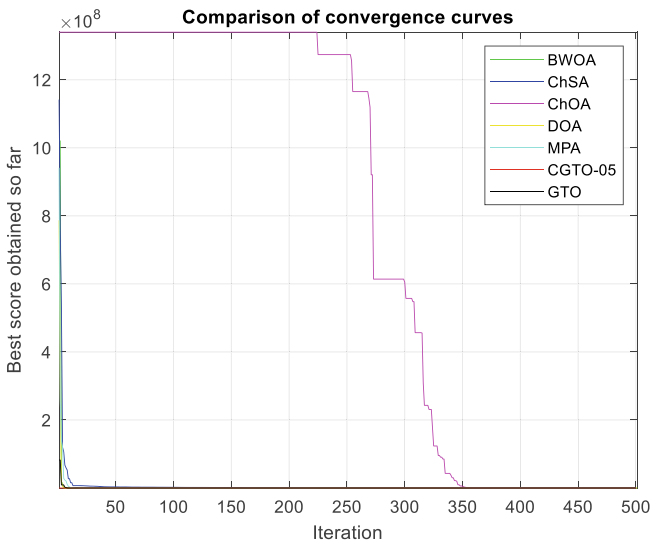


Fig. 1 Convergence curve of test function F2

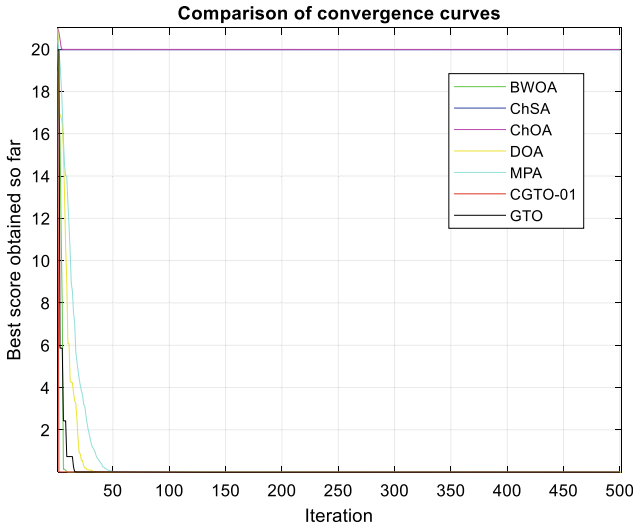


Fig. 2 Convergence characteristics of test function F3

$$G_1(\gamma) = \frac{2.131\gamma^4 + 2084.7\gamma^3 + 3.887e05\gamma^2 + 3.179e07\gamma + 2.71e09}{\gamma^5 + 609.98\gamma^4 + 2.17e05\gamma^3 + 3.454e07\gamma^2 + 2.125e9\gamma + 2.362e10} \tag{2}$$

It is quite difficult to develop an implementable controller corresponding to this higher order machine model. Thus, this model is reduced to a second order system by using the proposed CGTO methods. For this experiment however, 20 search agents and 100 iterations are selected as this involves finding only four decision variables. Several new algorithms like MPA, BWOA, ChOA, ChSA, DOA including GTO are utilized for comparison. The reduced models are provided in Table 2. Since only the heuristic technique is used to bring about the reduced model, hence average and standard deviation of the optimized error value viz. ITAE in this case is provided in this Table. Amongst the proposed techniques, only CGTO-06 is reported. The best error values are also bolded in this Table for the convenience of the readers.

It is found from Table 2 that the proposed CGTO-06 technique outperforms other methods in terms of average ITAE error optimized. The standard deviation of the ITAE error is however least in the MPA method indicating that the algorithm is more stable in terms of the other techniques used for comparison. Further, the convergence characteristics are drawn as given by Fig. 3.

Even an intelligent PID controller is also developed for this motor using the proposed technique. The sum of square error (SSE) is optimized to evaluate the controller parameters. The controller parameters are determined using approximate model matching method in the delta domain. A fixed reference model is taken up for the study adopted as in [5]. A handful of new algorithms are used for comparison. There are only three decision variables involved in the controller tuning process.

Table 2 Reduced models of the 50 hp induction motor in the unified delta domain, average and standard deviation of error function

Algorithms	Reduced systems in the delta domain	Avg. error	Std. error
CGTO-06	$\frac{1.284\gamma+20.37}{\gamma^2+23.83\gamma+161.62}$	0.0103297	1.8977e-09
GTO	$\frac{1.33\gamma+5.874}{\gamma^2+14.02\gamma+50}$	0.010337	8.87e-07
BWOA	$\frac{1.398\gamma+6.298}{\gamma^2+15.17\gamma+49.41}$	0.010341	1.29e-06
ChOA	$\frac{1.404\gamma+0.5254}{\gamma^2+11.26\gamma+4.287}$	0.010343	1.66e-06
ChSA	$\frac{1.329\gamma+5.744}{\gamma^2+13.93\gamma+49.04}$	0.010337	1.86e-06
DOA	$\frac{1.33\gamma+5.874}{\gamma^2+14.02\gamma+50}$	0.010337	1.24e-06
MPA	$\frac{1.33\gamma+5.874}{\gamma^2+14.02\gamma+50}$	0.010337	1.03e-10

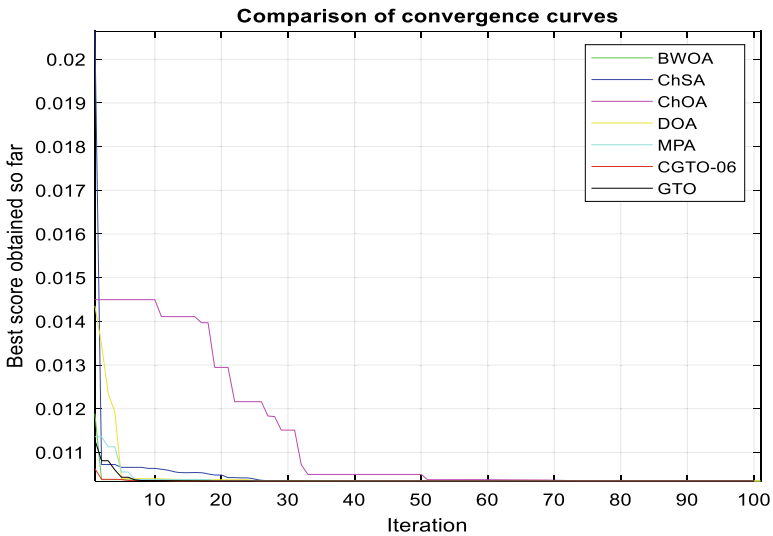


Fig. 3 Convergence plot of reduced 50 hp induction motor in the delta operator framework

Hence, the population size and the maximum number of iterations are taken up as 20 and 100 respectively for this optimization problem. The suggested CGTO-06 technique yields a controller in the delta domain given by

$$G_c(\gamma) = 42.58 + \frac{0.045}{\gamma} + 4.2\gamma \tag{3}$$

Moreover, the convergence plot of this controller tuning problem is shown in Fig. 4.

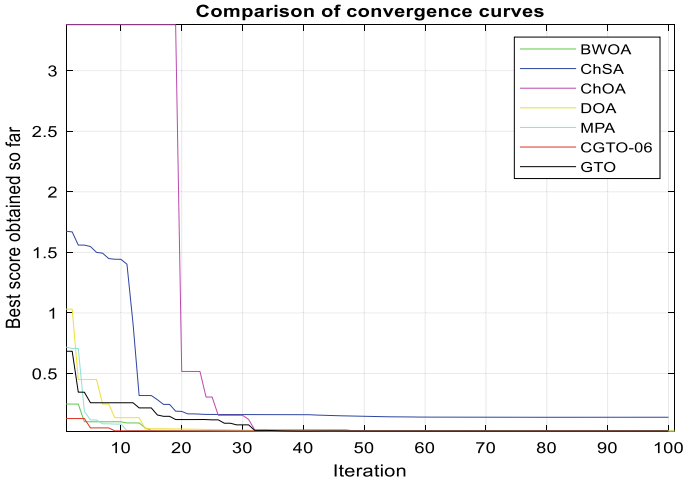


Fig. 4 Convergence plots of controller tuning problem for the 50 hp induction motor using the delta operator

From the convergence plot of Fig. 4, it is evident that the proposed CGTO-06 shows an appreciable good convergence when compared with a host of the latest techniques reported in the literature.

Experiment 3: The next test system chosen for the study is an induction motor drive model [15]. The transfer function of the drive model is denoted by

$$G_2(s) = \frac{13.381 s + 40.54}{8.58e-07s^3 + 0.003517s^2 + 4.802 s + 40.69} \tag{4}$$

The drive system is modelled in the delta domain as

$$G_2(\gamma) = \frac{1.128e03\gamma^2 + 4.643e05\gamma + 1.395e06}{\gamma^3 + 8.23e03\gamma^2 + 1.728e05\gamma + 1.402e06} \tag{5}$$

The model corresponds to a sampling period of 0.0025 s. Once again, the experimentation is performed to reduce this model to a lower order second order system choosing four unknown parameters. Hence the choice of 20 search agents and 100 iterations can be considered appropriate. The reduced model in the unified domain of the proposed method and their optimized fitness function values (mean and standard deviation) are depicted in Table 3. A handful number of methods are also applied for comparison purpose. The least error values are marked with the help of bold letters.

From Table 3, it is clear that the proposed method provides least error values in terms of both average as well as standard deviation. Only GTO method provides same average error. The standard deviation is significantly low proving that the proposed algorithm is quite stable. The convergence diagram is also plotted in Fig. 5 to validate the proposed technique.

Table 3 Reduced models of the induction motor drive in the unified delta domain, average and standard deviation of error function

Methods	Lower-order models in the delta domain	Avg. error	Std. error
CGTO-06	$\frac{1128.4616\gamma+3416.3041}{\gamma^2+414.73204\gamma+3428.5995}$	0.0000184	1.8273e-10
GTO	$\frac{2784.7\gamma+3325.1}{\gamma^2+1536.4\gamma+3031.7}$	0.0000184	0.000073
BWOA	$\frac{1998.54\gamma+1994.01}{\gamma^2+706.3\gamma+1994.386}$	0.017274	0.024938
ChOA	$\frac{13000\gamma+13000}{\gamma^2+4634.68\gamma+13000}$	0.035431	0.035431
ChSA	$\frac{4416.35\gamma+4222.24}{\gamma^2+1647.04\gamma+4430.1}$	0.015290	0.018626
DOA	$\frac{649.92\gamma+1266.36}{\gamma^2+240.2\gamma+1367.97}$	0.047036	0.108978
MPA	$\frac{4341.04\gamma+6758.52}{\gamma^2+1586.87\gamma+6971.5}$	0.000165	0.000057

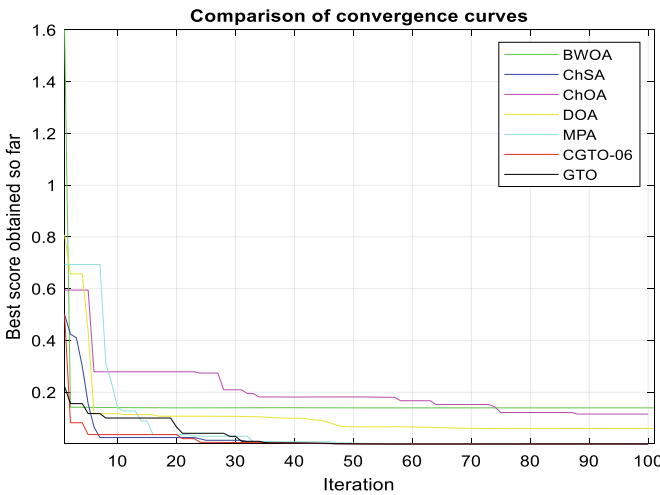


Fig. 5 Convergence plots of reduced induction motor drive in the combined domain

The proposed technique provides quite good convergence in terms of the algorithms being compared. Using the proposed technique, an intelligent PI controller is also constructed for this drive application. To evaluate the controller parameters, the sum of square errors (SSE) is optimised. The controller parameters are calculated in the delta domain using the approximate model matching method. The study employs a fixed reference model as discussed in [5]. For comparison, a number of latest algorithms are applied. The controller parameters optimized for the CGTO-06 approach whose transfer function is given by

$$G_{PI}(\gamma) = 0.21 + \frac{3.34}{\gamma} \tag{6}$$

The time and frequency domain responses of the controlled plant are compared with the reference model whose graphs are shown in Figs. 6 and 7 respectively.

The step response in Fig. 6 shows a close resemblance with the reference model compared. From the frequency domain plot in Fig. 7, it is evident that the controlled induction motor drive and reference model in the delta domain show close resemblance at lower frequency but there is some mismatch at relatively higher frequency

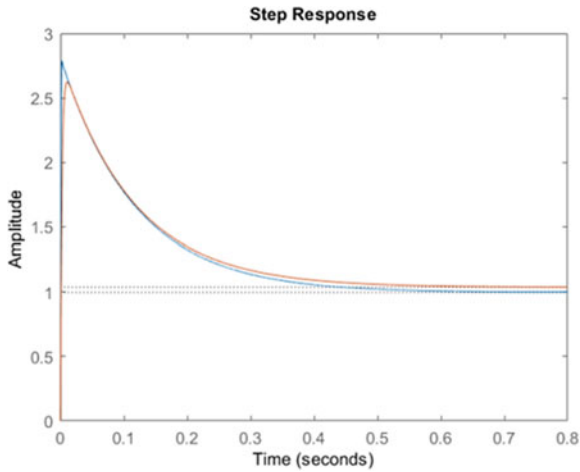


Fig. 6 Step responses of controlled induction motor drive and reference model

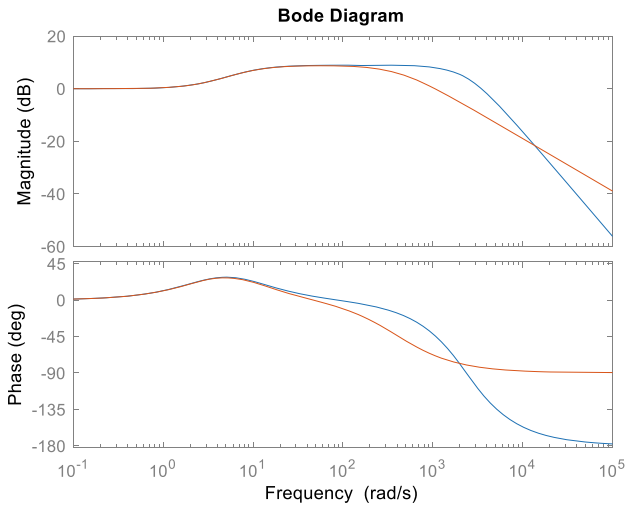


Fig. 7 Bode response of controlled induction motor drive and reference model

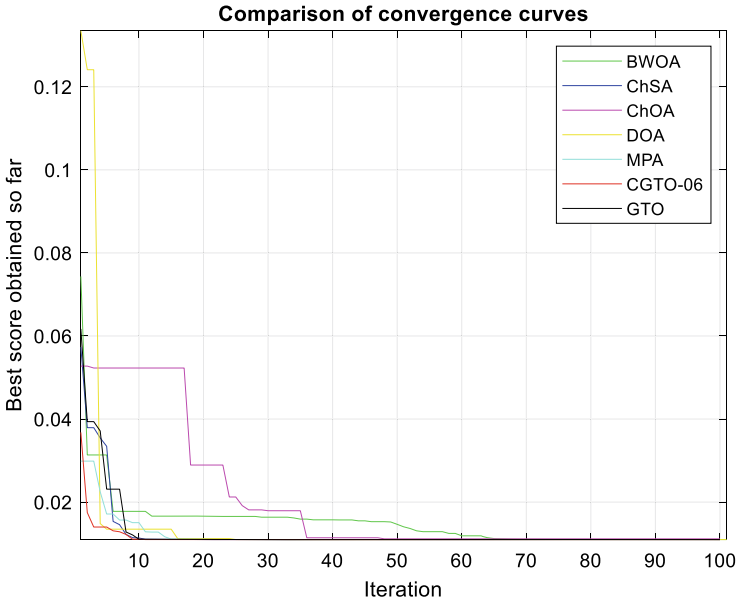


Fig. 8 Convergence diagram of controller tuning problem for the induction motor drive

levels. Further, the convergence curve is also plotted in Fig. 8 for the proposed CGTO-6 technique and is compared with some of the popular heuristic techniques available in the literature.

The Fig. 8 depicts good convergence behaviour of the proposed technique over the existing standard heuristic methods. Thus the proposed CGTO approaches prove successful for modelling and control in the combined domain of analysis. There are several ways by virtue of which new chaotic versions of GTO can be developed. There are three controlling parameters in the GTO algorithm which can also be chaotically varied. Further, there are some random parameters used in this GTO method which can be replaced by one-dimensional chaotic maps.

5 Conclusions

A new chaotic version of the gorilla troop optimizer is developed in this paper. The parent algorithm's position equation is updated using various one-dimensional chaotic maps. To develop new chaotic algorithms, ten popular and widely cited chaotic maps are being considered. To validate the efficacy of the proposed technique, two types of test functions are used: unimodal and multimodal. A 50 hp induction motor model is also reduced, and its controller is designed with the help of this proposed algorithm, utilising the benefit of the delta operator. Finally, an induction

motor drive is used for order reduction and controller synthesis. In both cases, the controller is realised using an approximate model matching framework. The proposed techniques outperform the standard and most recent methods in terms of convergence speed and accuracy. The statistical measures of the optimal values are also examined, taken into account the stability of the proposed algorithms. Selected results have been presented to the readers. Last but not the least, the results of all experiments performed is really promising. The proposed technique may be applied for other complicated engineering design problems.

References

1. Liang X, Ali MZ, Zhang H (2019) Induction motors fault diagnosis using finite element method: a review. *IEEE Trans Ind Appl* 56(2):1205–1217
2. Sarkar P, Pal J (2004) A unified approach for controller reduction in delta domain. *IETE J Res* 50(5):373–378
3. Ganguli S, Kaur G, Sarkar P (2022) Model order diminution of MIMO systems using the delta transform method with new firefly-based hybrid algorithms. *Soft Comput* 1–18. <https://doi.org/10.1007/s00500-021-06591-7>
4. Ganguli S, Kaur G, Sarkar P (2021) Global heuristic methods for reduced-order modelling of fractional-order systems in the delta domain: a unified approach. *Ricerche di Matematica* 1–29. <https://doi.org/10.1007/s11587-021-00644-7>
5. Ganguli S, Kaur G, Sarkar P (2021) An approximate model matching technique for controller design of linear time-invariant systems using hybrid firefly-based algorithms. *ISA Trans*. <https://doi.org/10.1016/j.isatra.2021.08.043>
6. Kaveh A (2017) Chaos embedded metaheuristic algorithms. In: *Advances in metaheuristic algorithms for optimal design of structures*, pp 375–398
7. Abdollahzadeh B, Soleimani Gharehchopogh F, Mirjalili S (2021) Artificial gorilla troops optimizer: a new nature-inspired metaheuristic algorithm for global optimization problems. *Int J Intell Syst* 36(10):5887–5958
8. Gupta J, Nijhawan P, Ganguli S (2021) Parameter estimation of different solar cells using a novel swarm intelligence technique. *Soft Comput* 1–31. <https://doi.org/10.1007/s00500-021-06571-x>
9. Ganguli S, Kumar A, Kaur G, Sarkar P, Rajest SS (2021) A global optimization technique for modeling and control of permanent magnet synchronous motor drive. *Innov Inf Commun Technol Ser* 074–081
10. Ganguli S, Srivastava T, Kaur G, Sarkar P (2022) Model reduction and controller scheme development of permanent magnet synchronous motor drives in the delta domain using a hybrid firefly technique. In: *Handbook of intelligent computing and optimization for sustainable development*, pp 537–547
11. Sarcar NC, Sarkar P, Bhuyan M (2010) Delta operator modelling and control by optimal frequency matching using GA. *Int J Model Simul* 30(4):434–444
12. Wolovich WA (1972) The use of state feedback for exact model matching. *SIAM J Control* 10(3):512–523
13. Jamil M, Yang XS (2013) A literature survey of benchmark functions for global optimisation problems. *Int J Math Model Numer Optim* 4(2):150–194
14. Waszynuk O, Diao YM, Krause PC (1985) Theory and comparison of reduced order models of induction machines. *IEEE Trans Power Appar Syst* 3:598–606
15. Krishnan R (2001) *Electric motor drives: modeling, analysis, and control*. Pearson

rSense: A Novel Gesture-Based Human Assistive Device



Vijay A. Kanade

Abstract The research paper discloses the implementation of rSense, a novel gesture-based human assistive device. The device comprises a ring and wireless earpiece, wherein the ring has Internet connectivity and is operated via gestures, while the earpiece is connected to the ring via Bluetooth. The proposed device utilizes sensory modalities of vision and hearing to perceive the environment and assist individuals by relaying relevant feeds over the earpiece in real time. It uses gesture-based controls and Internet-assisted intelligence to perform tasks such as reading, object recognition, speech recognition and others. The device is capable of providing high-level assistance to students, researchers, blind people, and people with reading disabilities.

Keywords rSense · Gesture-based controls · Assistive device · Sensory modality · Artificial Intelligence (AI) · Machine Learning (ML) algorithms

1 Introduction

Assistive technology has been around for a quite some time now, right from wheelchairs, hearing aids, memory aids, to spectacles, walkers, and prostheses. Most assistive devices have contributed significantly in improving the overall quality of human lives. According to a May 2022 report published by WHO and UNICEF, around 2.5 billion people across the globe need one or more assistive systems like wheelchairs, hearing aids, spectacles, etc. to lead a normal life [10].

The technology has revived the lives of all individuals, be it a healthy person or a person with some kind of a disability. It has ensured the well-being of everyone by allowing them to live a more productive, independent and easy life. Moreover, with the proliferation of IoT devices and advancement in technology, such assistive devices have become portable and are easily available in the market [9]. As a result, the demand for these devices has surged over the years.

V. A. Kanade (✉)
Pune, India
e-mail: kanade.science@gmail.com

The primary objective of any assistive device is to resolve the difficulties faced by people and make their lives easier. Each device addresses a specific problem or difficulty. For example, spectacles improve the vision of individuals having visual problems. However, the problem faced by avid readers including researchers, students or individuals who tend to consume a lot of content via books, research papers, or digital media in a limited time span has not been addressed in specific. Such problems are typically related to the difficulty in determining the meaning of the alien content and making sense of it while reading. Here, the alien content may refer to unknown words, mathematical formulas, equations, derivations, experimental data, technical terminology, scientific terms, chemical formulas, figures, words and phrases in foreign languages, etc.

Moreover, people with reading disabilities (i.e. dyslexia) and blind individuals seem to face a similar problem while reading. Although braille or books on tape are available for blind readers, it is highly unlikely that all books or scriptures may have such a functionality. Hence, there seems to be a long standing need to develop an assistive system that helps readers of all kinds to read and assimilate content.

2 Typical Challenges Faced by Readers

Reading comes easy to most individuals. And that's the reason we tend to overlook the complex process underlying it. Fundamentally, reading involves translation of visual cues into words and words to meanings. However simple reading may seem, readers from different spectrums and age groups including students, researchers, analysts or any layperson encounter problems while reading.

For example, sometimes a reader may not know the meaning of a word, formula, or scientific term. In such a case, the reader may have to look for the meaning of the unknown content elsewhere like in a dictionary, book, or an online source like Wikipedia and then come back to reading the text. This can be problematic at times as readers may lose concentration, reading rhythm and also face a problem while correlating the already consumed content with the meaning of the unknown matter. Such problems can double if the user is reading a foreign or native language where most of the content is unknown to him.

In a survey, the challenges faced by students while reading their class literature, research articles, and scientific papers were studied. According to the study, students found it difficult to follow up on experimental data, understand the scientific terms, and interpret unfamiliar techniques, methods and figures [1].

A similar problem is encountered by researchers as they find it hard to interpret the unfamiliar mathematical equations, derivations, or new scientific terminologies. Although the citations given in the papers suffice the cause, however, referring to the citations most likely derail the reading flow of the reader. Also, in some cases, content unrelated to the subject may surface through citations which can further act as an impediment as the reader is left with no option but to find a relevant source elsewhere that defines the unknown subject matter. The reader can then make sense of the entire content.

Moreover, reading is also a matter of concern for blind community and dyslexic society. The paper provides a solution to this non-mainstream problem of struggling readers.

3 Architecture of rSense

rSense is a human assistive device that comprises two components, namely, a ring and a wireless earpiece. The ring is to be worn on a finger. It is connected to the Internet and is gesture controlled, while the earpiece is in direct communication with the ring via Bluetooth and relays the audible information in the user's ear [8]. Gesture control implies the ring has motion sensors to detect the movement of the device at the minutest scale [2].

The device is designed to perceive its surroundings via sensory modalities of vision and hearing. The vision module is enabled via a miniature camera that captures the visual data, while the hearing module captures audible data through a speaker unit. The camera in combination with motion sensors help to capture finger movement and perform gesture recognition. Both the vision and hearing modules are triggered based on this very gesture input. The device also contains a wireless processing unit embedded in the ring component which processes data.

Moreover, the device is capable of performing recognition based functions such as word, character, object, speech, image and mathematical formula recognition using AI and machine learning algorithms [3–7]. It also aids in reciting the mathematical derivations of an unknown formula or breakdown chemical equations so that it assists the reader in knowing its relevance in the context of the text being read.

Let's consider a use case where the user is reading a book while wearing the device (ring and earpiece) and comes across a word or phrase which is unknown to him. In such a case the user can hover the camera of the ring over the unknown content and perform gestures while focusing on the content. Some examples of gesture controls include the following:

- Single tap—suggest synonyms for the word or phrase.
- Double tap—find the meaning of the tapped word or phrase.
- Fold finger—identify mathematical relevance of a formula or equation.
- Tap the adjacent finger—start recording the audio.
- Circular movement—explanation for figures.

After the hand gestures, the processing unit in the device (ring) determines the gesture type and performs the corresponding activity. In our case the activity corresponds to a search function as that of a search engine. The information for the unknown content is searched over the Internet and the most relevant information is broadcasted through the earpiece. The relevancy of the content is also determined by the processing unit prior to relaying it in the user's ear.

The working of rSense showcasing the visual (i.e. while reading a book) and hearing (i.e. while hearing sound) modalities is shown in the below figure (Fig. 1):

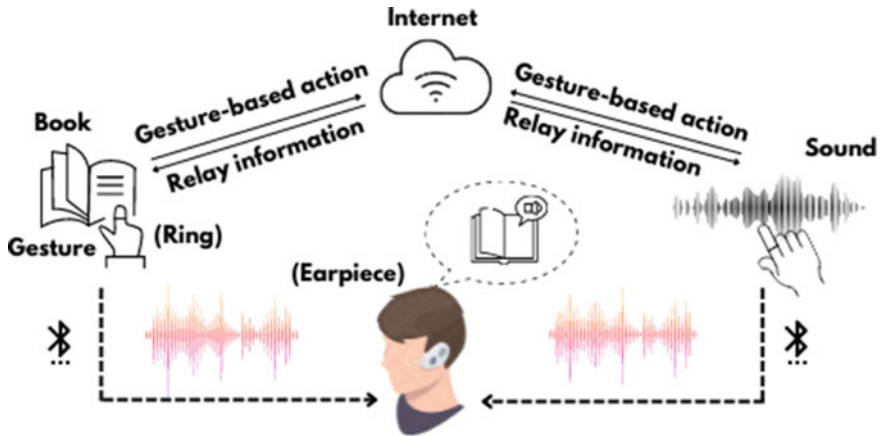


Fig. 1 Working of rSense device

3.1 Data Flow

The rSense device as in Fig. 2 follows a sequence of steps as shown below:

Step-I

In the first step, the rSense device is exposed to the visual or audible data. Visual content may refer to a book, image or any live visual scene. Similarly, audible data may refer to songs played at concerts, or audio of plays, music, etc.

Step-II

In the second step, the user performs a hand gesture that triggers an activity within the ring.

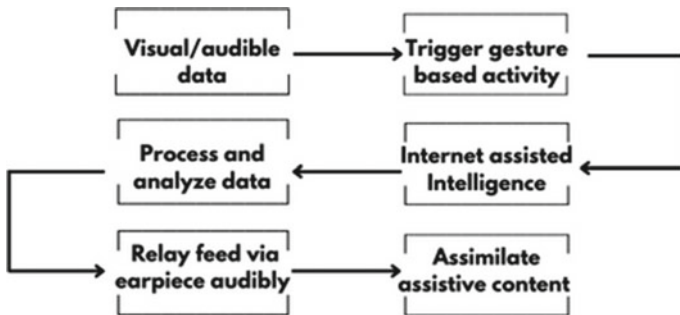


Fig. 2 Data flow diagram for rSense

Step-III

Next, the triggered activity (e.g. search) is transmitted to the server in the form of a request via the Internet and relevant information is thereby retrieved which is sent back to the rSense ring.

Step-IV

Next, the wireless processing unit within the ring processes and analyzes the data to determine the relevancy of the received content.

Step-V

Following step four, the most relevant content is relayed onto the wireless earpiece via Bluetooth mode of communication.

Step-VI

Lastly, the relayed content is ready for the user consumption. It may assist him in the ongoing task.

4 Preliminary Results

Our research presents preliminary findings for the visual module of rSense. We captured an image of the page being read by a reader. The book used by the reader is an old dilapidated one wherein some portion of text is unclear and unreadable. Older book was intentionally chosen to validate the concept of rSense.

Furthermore, we used an artificial intelligence tool that scans the captured image to retrieve the texts from it. The tool uses character and word recognition features to recognize and extract clear and unclear words observed in the image.

Following images show the implementation of this visual module:

In Fig. 3, red highlights reveal the unclear words that are difficult for the reader to make sense of. The AI tool comprehends the context of the entire text and fills the unclear words with words that seem most relevant in view of the entire text. This is represented in Fig. 4 below.

The AI tool replaces seven unclear words in Fig. 3 with seven appropriate words shown in Fig. 4. The replaced words closely match the context, thereby making the entire text logical.

5 Applications

rSense is typically useful to students and researchers who come across diverse literature from various technical streams. More so, the device assists blind individuals and people with reading disabilities. The researchers can also extend the application

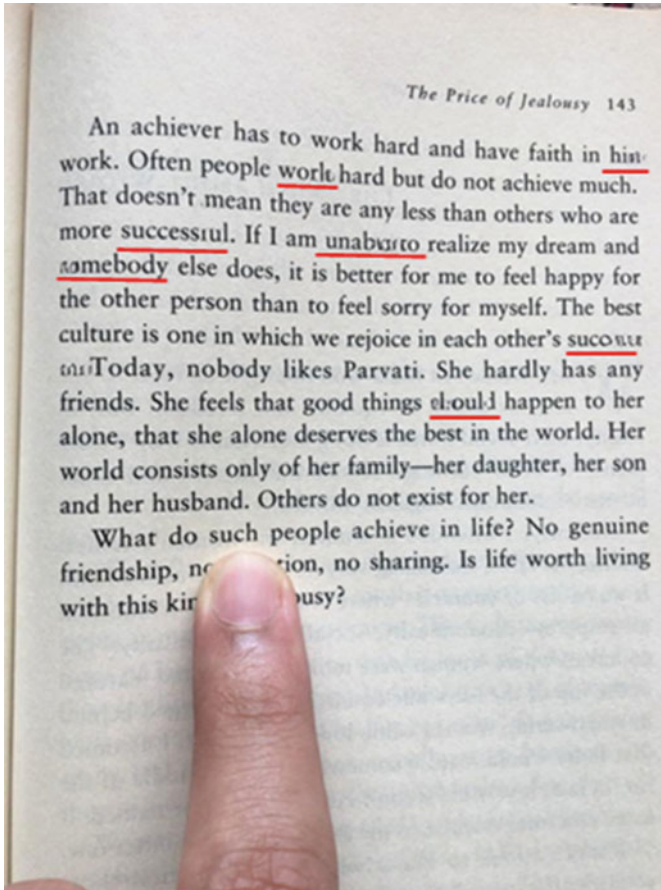


Fig. 3 Photo clicked using a camera (Color figure online)

of rSense by employing machine learning and artificial intelligence algorithms in the processing unit of the ring. Following are some key futuristic applications of rSense:

1. Medicine

Medical practitioners can use the speech and image recognition functionality to ascertain the symptoms in a patient [4, 5]. Through image recognition, rSense may scan the photos of a patient and suggest diagnosis or recommend treatments for it. Fundamentally, such treatments are a consequence of analysis done through ML or AL algorithms. The results of it are transmitted over the earpiece of rSense used by the medical professionals.

The process may help in identifying cancerous tissues, analyze bodily fluids or even correlate photos to rare genetic diseases by annotating them.

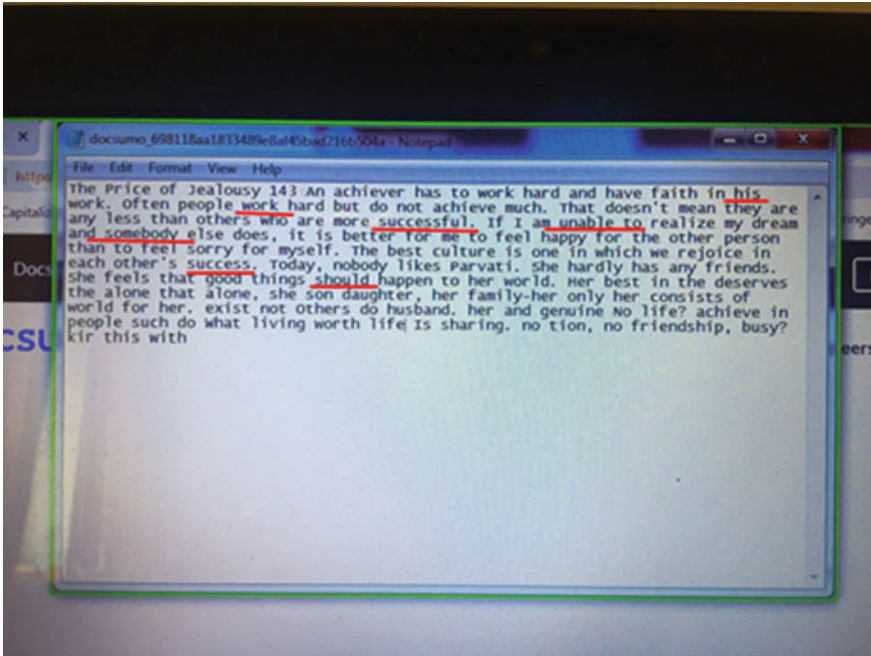


Fig. 4 Snapshot of text extracted from the captured image by the AI tool

2. Personalization

The technology can be used by professional scholars or academicians as they can decide upon certain paid databases that the rSense device may access to extract relevant data. For example, the user may manually configure the device during the setup process and give access rights to academic databases like Springer, IEEE Xplore, ACM, Elsevier, etc., which are not be freely available on the web. This may expand the scope of the rSense device.

3. Song Composition

The device can be used to generate variations of song notes that are fed into it. For example, suppose a user attends a live concert and likes the song's melody, rhythm or form. With the use of rSense, the user can apply and train ML, AI algorithms to create different versions of the song note, thereby aiding in song composition practices [11].

4. Connected Devices

rSense can be connected to various other smart devices such as smartphones, smart watches, smart home appliances, gadgets, etc. This allows the user to sync all the connected devices, user accounts and in turn permit crosstalk between them. For example, if the rSense ring is connected to the user's smartphone, he can access urgent emails on the go. The ML and AI algorithms analyze the email content and access

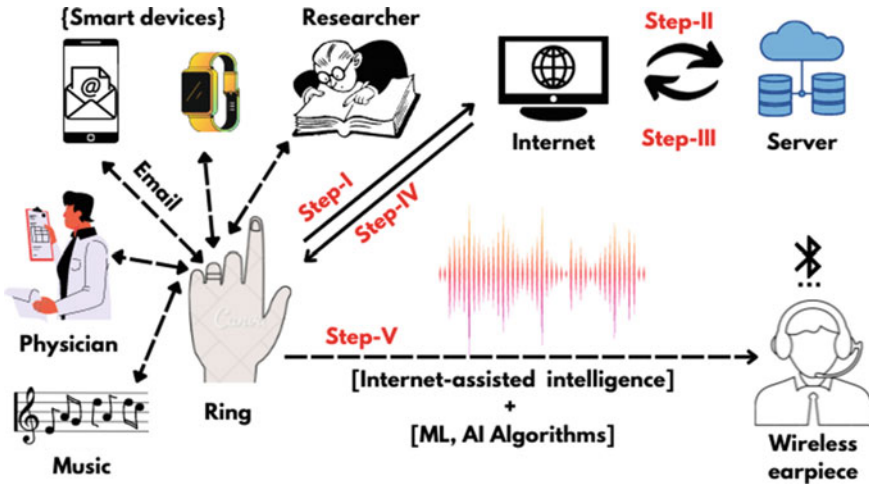


Fig. 5 rSense applications

the important ones requiring user attention. On determining the email’s relevance and importance, the earpiece reads, notifies and alerts the user about it.

The following Fig. 5 depicts the stepwise operation of rSense device along with various applications:

6 Conclusion

The research paper discloses the futuristic concept of rSense, a novel gesture-based assistive device. The device comprises a ring having Internet connectivity and a wireless earpiece connected to the ring via Bluetooth. rSense primarily addresses the problems of book readers, scholars and academicians as it searches for details of the unknown content over the web and transmits the relevant content via the earpiece in an audible format. Thus, the technology uses the Internet as an intelligence tool along with AI and ML algorithms to empower applications across the fields of medicine, academics, media and others.

7 Future Work

The experiments performed in the current research represent initial stages of rSense development. However, the camera component of the rSense ring is expected to perform the similar functionality as depicted in Figs. 3 and 4. The entire process

happens in real time with the involvement of the Internet. Although these are preliminary research findings, in future we intend to take this further and develop a full-fledged rSense device with the visual module (i.e. camera and AI/ML algorithms) embedded in the rSense ring along with a separate earpiece component. The AL and ML algorithms will be a part of the processing unit of the rSense ring that processes the visual image as seen in Fig. 4 and also relays the processed content in the audible form through the rSense earpiece.

Acknowledgements I would like to extend my sincere gratitude to Dr. A. S. Kanade for his relentless support during my research work.

Conflict of Interest The authors declare that they have no conflict of interest.

References

1. Lie R, Abdullah C, He W, Tour W (2016) Perceived challenges in primary literature in a master's class: effects of experience and instruction. *CBE Life Sci Educ* 15:1–12. ar77
2. Gesture Control Technology: An investigation on the potential use in Higher Education, University of Birmingham
3. Scheidl H (2018) Build a handwritten text recognition system using TensorFlow, towards data science, 15 June 2018
4. Alake R (2020) How does AI detect objects? (Technical), towards data science, 14 January 2020
5. Doshi K (2021) Audio deep learning made simple: automatic speech recognition (ASR), how it works, 26 March 2021
6. Javed M (2020) The best machine learning algorithm for handwritten digits recognition, towards data science, 22 November 2020
7. Bluche T (2010) Mathematical formula recognition using machine learning techniques, Worcester College, University of Oxford, 3 September 2010
8. How do true wireless earbuds work? RHA, 23 August 2019
9. Hasan M (2022) State of IoT 2022: Number of connected IoT devices growing 18% to 14.4 billion globally, IoT Analytics, Market Insights for IoT, 18 May 2022
10. WHO news (2022) Almost one billion children and adults with disabilities and older persons in need of assistive technology denied access, according to new report, 16 May 2022
11. Luck S (2021) Automatic music generation using AI, towards data science, 11 January 2021

Smart Boosted Model for Behavior-Based Malware Analysis and Detection



Saja Abu-Zaideh, Mohammad Abu Snober, and Qasem Abu Al-Haija

Abstract Malware analysis and detection are the most important activities to ensure system security. However, current attacks like polymorphic viruses and zero-day attacks that utilize signature-based methods complicate the detection process with accurate results. Therefore, this has—in turn—raised the need for more intelligent techniques to analyze the behavior of the malware rather than depending on the signature-based analyses. This paper proposes a machine learning-based model to analyze and detect the different types of malware. The system tries to determine the optimal feature representation and extraction and classification method that can lead to the best detection accuracy. Particularly, different machine learning algorithms were evaluated, including k-Nearest Neighbors (kNN), Multi-Layer Perceptron (MLP), Naive Bayes Classifier (NBC), Adaboost/XGBoost Decision Trees (ADT), and Support Vector Machines (SVM). The models were trained and tested using a new dataset that includes op-codes available in .asm format (generated using the IDA disassembler tool); it is a subset of data used in Kaggle for the Microsoft Classification challenges. Our empirical results revealed the superiority of the XGBoost-based model scoring an overall detection accuracy of 98.3%.

Keywords Malware · Machine learning · XG Boost · Malware analysis · Malware detection · Classification · Accuracy

1 Introduction

Obviously, it is justifiable that malware has become one of the major threats in the world. The prompt improvement of communication and the internet is the main reason. Malware or malicious software is an accumulated number of viruses developed to execute a malicious activity, information-stealing, reconnaissance, or others. Malware is mainly meant to produce massive destruction to the data and sources or

S. Abu-Zaideh · M. A. Snober · Q. A. Al-Haija (✉)

Department of Computer Science/Cybersecurity, Princess Sumaya University for Technology (PSUT), Amman, Jordan

e-mail: q.abualahaija@psut.edu.jo

to obtain unlawful access to the network. Also, it is defined as “a type of computer program designed to infect a legitimate user’s computer and inflict harm on it in multiple ways.” [1]

The main purpose of employing machine learning techniques instead of anti-virus scanners is that the assortment of malware types is expanding, resulting in millions of hosts being attacked. Therefore, anti-virus systems and other scanners cannot comply with the requirements of recognition and security. The latest research demonstrates that seven million various hosts were attacked, and in 2015, up to four million malware entities were discovered [2]. Nowadays, for the period of the COVID-19 epidemic, the number of malware attacks has grown; Malware expenditures are currently totaling over \$1 billion every year.

Additionally, the malware raises for the reason that various accessible tools are established that involve the lowest possible degree of proficiency. Recent research indicates that the majority of attackers nowadays are script-kiddies [3].

Consequently, malware safeguard is among the most important cybersecurity tasks to ensure user privacy and confidentiality; this is since, at the same time, an individual attack can lead to a sufficient loss in the organizational assets. Recurrent attacks impose the demand for sensible and precise recognition approaches. We realized that the existing static and dynamic approaches do not support effective detection dependent on modern attacks, such as zero-day attacks. For this reason, the application of machine learning-based methods was increased. Thus, in this paper, we will discuss the primary arguments and interests of machine learning-based malware detection. Also, we will cover the finest feature interpretation and classification procedures.

The primary objective is to determine the optimal feature selection process and how the features must be extracted. Such an accurate procedure can differentiate the malware types with the smallest possible error value. Hence, to achieve this goal, we will create the proof of concept for the machine learning-based malware classification, which will be utilized as an input to the machine learning systems. Consequently, we look at high-accuracy results to determine the best performant algorithm.

The leftover parts of this manuscript are structured as follows: Sect. 2 provides an essential theoretical background to fulfill the important knowledge required throughout this paper. Section 3 provides the details of the proposed model, including feature selection, machine learning modeling, the dataset, and cross-validation. Section 4 discusses the experimental results and Sect. 5 concludes the finding of this research paper.

2 Theoretical Background

This section delivers the basic knowledge which plays a significant role in understanding malware discovery and the necessity for machine learning techniques. First, we will describe the malware types relevant to the study, then the typical malware

detection techniques. Subsequently, we will discuss the necessity for employing machine learning based on the knowledge gained. Moreover, we will review some of the related work.

2.1 Malware Types

In this section, we will classify the malware to provide the best way to understand the methods and logic; we can divide Malware, depending on its intention, into several groups. As follows:

- **Worm:** This type can propagate across the network, similar to the virus. Also, it is able to reproduce in other machines.
- **Virus:** Viruses can declare as any part of the software which is inserted and released automatically; with user consent and permissions. This is the humblest structure of software. It can reproduce itself or infect (modify) other software [4].
- **Adware:** This type of malware can display advertisements on your computer. We can say that adware is a subclass of spyware.
- **Trojan:** This is a type of malware that intends to be seen as lawful software. Trojans may be engaged by cyber thieves and hackers attempting to retrieve the systems of the users. Social engineering normally deceived users into the insertion and execution of Trojans on their systems. When enabled, Trojans can activate cyber-criminals to snoop on your system, sneak your vulnerable data, and obtain backdoor entrance to your system [5].
- **Spyware:** Spyware is a type of malware that performs espionage and can be mentioned to as spyware. It can infect your PC or mobile. So, spyware can do several warm actions like gathering information about you, including tracking your search history, the websites you visited, the belongings you downloaded, your credentials (usernames and passwords), payment details, and the emails you send and receive, all that to send personalized announcements, or to sell them to the third parties subsequently [6].
- **Rootkit:** It's designed to enable access to your computer's data with more privilege than is allowed. So, it is used to give administrative access to an unauthorized user. Another important piece of information about Rootkits is that it is hard to detect or incredible to remove because they hide their existence.
- **Backdoor:** This type of malware indicates any procedure which enables authorized users and/or unauthorized users to gain high-level user access to the computer system. Provides an additional secret "entrance" to the system. It can manage to steal financial and private data or to hijack devices.
- **Ransomware:** Ransomware is malware that encrypts all the data in the victim's computer using an encrypted key. So the user cannot open any file in his machine until he gets the decrypted key, and he can get it by transferring money to the attacker.

- **Keylogger:** This type of malware stores all keys logged by the user, like usernames, passwords, and account numbers, so that the attacker can get this sensitive information [7].

2.2 *Detection Methods*

We can detect malware depending on the file's signature or by testing the behavior of executable files. But first, we must recognize between static and dynamic malware analysis. Static analysis is for non-executable files or without running the file. And dynamic analysis includes testing the executable file while running; this approach can be made using sandboxes.

The main job of static analysis is to infer the file's behavior properties by reading the malware's source code and infer the file's behavioral properties. Several techniques can Static analysis cover it, such as [8]:

1. **File Format Inspection:** file metadata can fork out utility information like Windows PE (portable executable) files.
2. **String Extraction:** Examining the software output (e.g., status or error messages).
3. **Fingerprinting:** which compromises cryptographic hash computation.
4. **AV scanning:** by comparing the inspected file and if it is a well-known malware, it can be detected by all anti-virus scanners.
5. **Disassembly:** trying to infer the software logic and intentions by reversing machine code to assembly language.

Static analysis is safer than dynamic analysis. Because the file will not be running while it is under testing, it cannot infect the system. Static analysis is a simple basic approach it gives us to predict all possible behavioral scenarios. But it is not usually used in the real world because it is more time-consuming [9].

On the other hand, the dynamic analysis. It is less safe because we execute the file in a virtual environment like sandboxes while we test it to monitor the behavior of the file. It also runs at high speed and takes less time than static analysis [10].

2.3 *The Need for Intelligent Models*

We can detect malware depending on signatures. But there are two types of signatures; the first one is a static signature, all malicious files with a static signature can be detected easily using anti-virus scanners. They compare file signatures with all signatures of malicious files they have. If the signature matches, then they decide this file is malicious.

There are also types of files that can change their signature continuously; we call that polymorphic signature. This type of file spreads at a high rate. This type of signature cannot detect using traditional scanners. From that, the need for an updated

tool arises, so in this paper, we will use machine learning by writing a python code that compares the feature of the file with the stored feature of the malicious file to try finding if this file is malicious or not with high efficiency and less time.

3 Proposed System Model

The data mining techniques are commonly rapidly developed, resulting in using machine learning in other fields like the security field. The way the computer program learns from its test is called machine learning. In 1959 Arthur Samuel declared machine learning a “field of study that gives computers the ability to learn without being explicitly programmed.” Machine learning depends on training the model to do its job by using some algorithm to do several functions: classification, clusterization, regression, etc. What machine learning does to train the dataset is take the dataset as input and build predictions by using certain models, then give us the output. The general workflow will give a good understanding of this process, as shown in Fig. 1 below.

The machine learning process contains five stages:

- Data intake. It is the process of loading the data from the file and storing it.
- Data transformation. In this step, we initialize the data by clearly and normalizing it to be appropriate for the algorithm. Also, feature extraction and selection are performed. In addition, the data is separated into sets—‘training set’ and ‘test set’—all done in this stage.

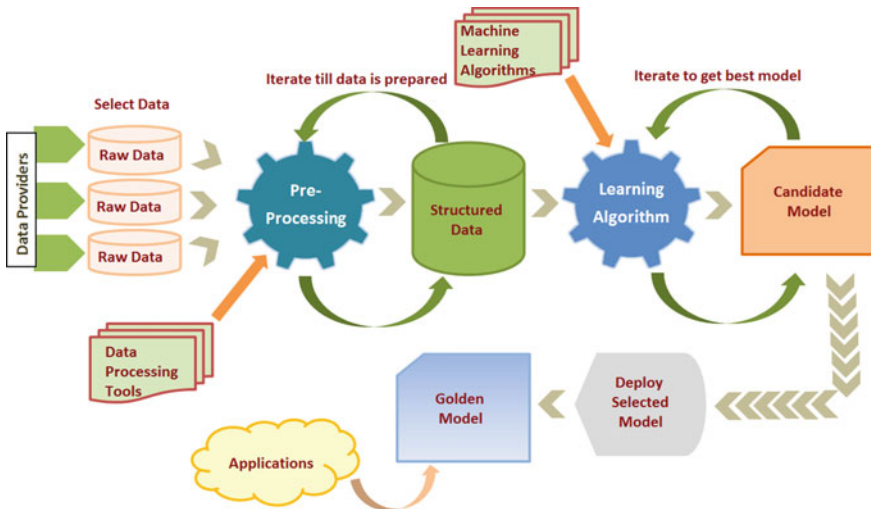


Fig. 1 The general process of machine learning

- **Model Training.** This step is responsible for selecting the appropriate algorithm and then building the model.
- **Model Testing.** The result produced from this step will be used for building new.
- **Model Deployment.** This stage is to select the final model.

3.1 Classification Methods

From the machine learning point of view, malware identification can be seen as a challenge of classification tasks where unknown malware categories must be clustered into various bunches according to specific attributes associated with the algorithm. In other contexts, having trained a model on the inclusive dataset of malicious and benign files, we can decrease this problem to classification. For common malware groups, this dilemma can be tightened down to classification only—having a restricted set of classes, to one of which malware instance certainly be in the right place, it is easier to recognize the appropriate class, and the result would be more accurate than with clusterization algorithms.

For instance, the random forest algorithm is used repeatedly in machine learning-based solutions. It is a simple algorithm that gives highly accurate results. As the name implies, it is called a forest because it is based on a large set of decision trees. It works over multiple decision trees. All these depend on an independent subset of datasets. It consists of n nodes. The main scheme of the algorithm is shown in Fig. 2 below. The advantages of this algorithm are that they are fit for classification and regression problems; they are easy to use and apply. They give a much better accurate result.

Also, one of the recent supervised machine learning algorithms is the XG Boost. XG Boost is considered a type of gradient boosted decision tree. With high speed and excellent accuracy. And in our experiment, we achieve high accuracy, which equals 98%, by using the XG Boost algorithm.

3.2 Data

In this research, we have used an open-source dataset from Microsoft; this dataset is roughly half a terabyte, supplied with a collection of common malware records indicating a combination of nine distinct groups. Every malware file has an identification number (ID). Every ID is composed of a 20-character hash value distinctively recognizing the file. The file also has a class label that is composed of integer values indicating one of nine malware types that the malware might fit in.

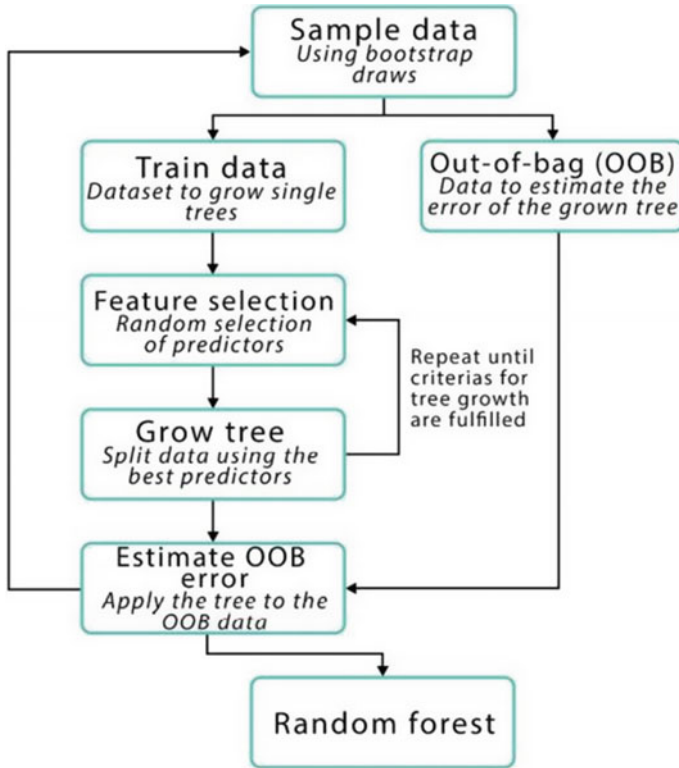


Fig. 2 Random Forest workflow

3.3 Cross-Validation

The cross-validation method is used to predict the way the model will perform on the new output data. The way it works is by splitting the dataset. Then the model takes the biggest part and will be trained on it. After that, the model will be evaluated using the small part.

1. Holdout method—it is a basic method, simply the origin dataset divided into two parts as shown in Fig. 3 below: the largest part, the training set, and the smallest part, the test set. And we trained the training date and evaluated it on test data. One of the advantages of this approach is that it is extremely fast.
2. The k-fold method is the improvement version of the Holdout approach. Here, the set is divided into numbers of subsets called k, and the holdout method is repeated k number of times. The disadvantage of this approach is that it is slow and more complex. We have used five-fold cross-validation in this research, as illustrated in Fig. 4 [11].

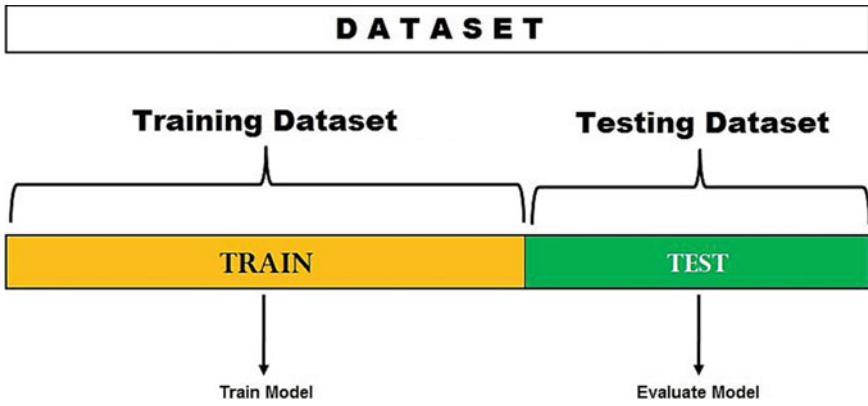


Fig. 3 The dataset in the holdout method

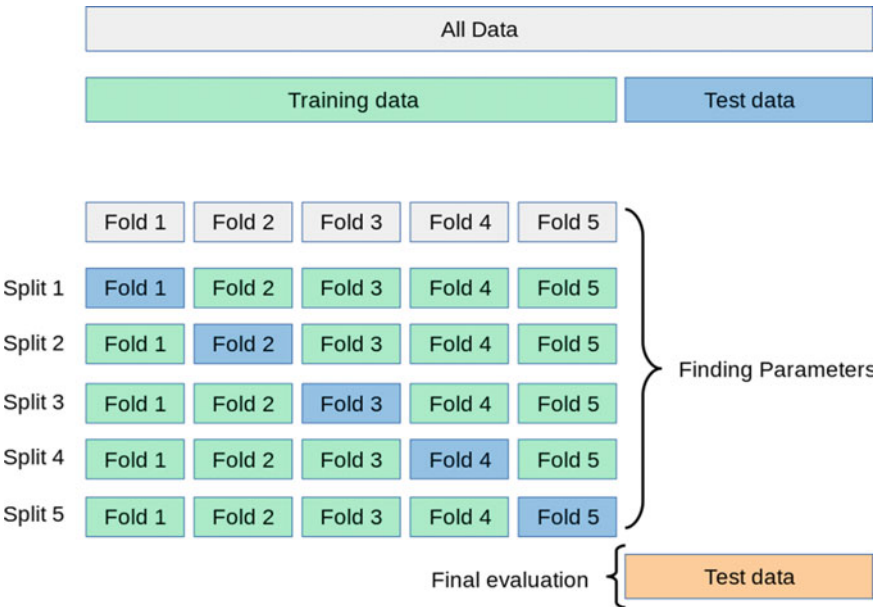


Fig. 4 Fivefold cross-validation method

4 Results and Discussion

Initially, feature selection is applied to eliminate redundant and inappropriate features. This in turn can enhance the predictive accuracy of the model. For our model, the feature set is tremendously big, and there is an indispensable necessity

for feature selection. Consequently, we applied a feature selection technique to end up with 56 features, as presented in Fig. 5.

After extracting the relevant features, and preprocessing the data records, then, we are capable of applying the machine learning approaches to the data we acquired. The machine learning techniques to be employed, as argued beforehand, include KNN, SVM, Naive Bayes, XG Boost, Random Forest, and Decision Tree. Figure 6 below shows data clustering and how they are divided into three groups. In the beginning, we applied an unsupervised machine learning, that is, the K-Means algorithm, to obtain the data classes where we obtained three classes 0 (Normal), 1 (Malware), or 2 (Unkown).

After that, ten different supervised machine learning techniques were applied to the labeled datasets: k-Nearest Neighbor (KNN), Linear SVM, RBF-SVM, Decision

```
Index(['Name', 'md5', 'Machine', 'SizeOfOptionalHeader', 'Characteristics',
      'MajorLinkerVersion', 'MinorLinkerVersion', 'SizeOfCode',
      'SizeOfInitializedData', 'SizeOfUninitializedData',
      'AddressOfEntryPoint', 'BaseOfCode', 'BaseOfData', 'ImageBase',
      'SectionAlignment', 'FileAlignment', 'MajorOperatingSystemVersion',
      'MinorOperatingSystemVersion', 'MajorImageVersion', 'MinorImageVersion',
      'MajorSubsystemVersion', 'MinorSubsystemVersion', 'SizeOfImage',
      'SizeOfHeaders', 'Checksum', 'Subsystem', 'DllCharacteristics',
      'SizeOfStackReserve', 'SizeOfStackCommit', 'SizeOfHeapReserve',
      'SizeOfHeapCommit', 'LoaderFlags', 'NumberOfRvaAndSizes', 'SectionsNb',
      'SectionsMeanEntropy', 'SectionsMinEntropy', 'SectionsMaxEntropy',
      'SectionsMeanRawsize', 'SectionsMinRawsize', 'SectionMaxRawsize',
      'SectionsMeanVirtualsize', 'SectionsMinVirtualsize',
      'SectionMaxVirtualsize', 'ImportsNbDLL', 'ImportsNb',
      'ImportsNbOrdinal', 'ExportNb', 'ResourcesNb', 'ResourcesMeanEntropy',
      'ResourcesMinEntropy', 'ResourcesMaxEntropy', 'ResourcesMeanSize',
      'ResourcesMinSize', 'ResourcesMaxSize', 'LoadConfigurationSize',
      'VersionInformationSize', 'legitimate'],
      dtype='object')
```

Fig. 5 The selected features

Fig. 6 Clustering and data labeling

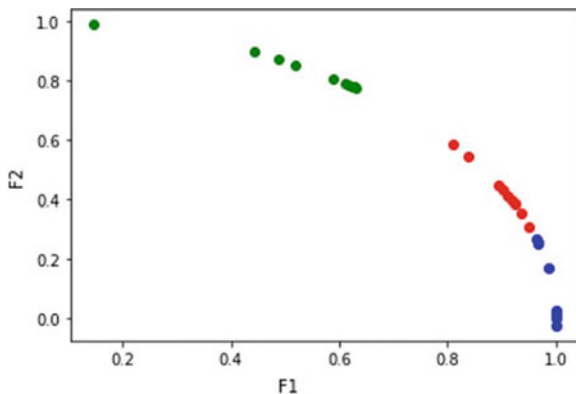


Table 1 Accuracy result

Accuracy	Name
0.943	KNN
0.773	Linear SVM
0.837	RBF SVM
0.937	Decision Tree
0.95	Random Forest
0.933	MLP Classifier
0.927	AdaBoost
0.9	Naive Bayes
0.787	QDA
0.98	XG Boost

Trees, Random forest, MLP Classifier, Adaboost, Naive Bayes, Quadratic Discriminant Analysis (QDA), and XG Boost. Table 1 shows the results of detection accuracy obtained from applying the stated models. From the results of different models, we can observe that the smallest accuracy rate was attained by Linear SVM (77.34%), followed closely by QDA and RBF SVM which obtained 78.7% and 83.7%, respectively. While the greatest accuracy rate was attained with the XG Boost, the XG Boost (Extreme Gradient Boosting) algorithm recorded the best detection accuracy result.

5 Conclusions and Future Work

In this paper, a new machine learning-based system for malware analysis and classification is proposed, implemented, and evaluated. The model uses ten machine learning techniques: k-Nearest Neighbor (KNN), Linear SVM, RBF-SVM, Decision Trees, Random forest, MLP Classifier, Adaboost, Naive Bayes, Quadratic Discriminant Analysis (QDA), and XG Boost. The experimental evaluation results for the detection accuracy showed that the model-based XG Boost is superior, with 98.3% accuracy. In the future, this experiment forms a good base and is applicable for larger datasets. Therefore, several future improvements related to the practical implementation of this project can be identified.

References

1. Kaspersky Labs (2017). What is malware, and how to defend against it? <http://usa.kaspersky.com/internet-securitycenter/internet-safety/what-is-malware-andhow-to-protect-againstit#.WJZS9xt942x>. Accessed 15 Feb 2017

2. Abu Al-Haija Q, Al-Dala'ien M (2022) ELBA-IoT: an ensemble learning model for botnet attack detection in IoT networks. *J Sens Actuator Netw* 11:18. <https://doi.org/10.3390/jsan11010018>
3. Aliyev V (2010) Using honeypots to study skill level of attackers based on the exploited vulnerabilities in the network. The Chalmers University of Technology
4. Horton J, Seberry J (1997) Computer viruses. An introduction. The University of Wollongong
5. Smith C, Matrawy A, Chow S, Abdelaziz B (2009) Computer worms: architectures, evasion strategies, and detection mechanisms. *J Inf Assur Secur*
6. Moffie M, Cheng W, Kaeli D, Zhao Q (2006) Hunting Trojan Horses. In: Proceedings of the 1st workshop on architectural and system support for improving software dependability
7. Chien E (2005) Techniques of adware and spyware. WWW document. <https://www.symantec.com/avcenter/reference/techniques.of.adware.and.spyware.pdf>
8. Lopez W, Guerra H, Pena E, Barrera E, Sayol J (2013) Keyloggers. Florida International University
9. Abu Al-Haija Q, Krichen M, Abu Elhaija W (2022) Machine-learning-based darknet traffic detection system for IoT applications. *Electronics* 11:556. <https://doi.org/10.3390/electronics11040556>
10. Prasad BJ, Annangi H, Pendyala KS (2016) Basic static malware analysis using open-source tools
11. Abu Al-Haija Q (2022) Top-down machine learning-based architecture for cyberattacks identification and classification in IoT communication networks. *Front Big Data* 4:782902

Quality Rating Application for Virtual Recipes Using Facial Analysis



C. Shyamala Kumari, Manoharan Pon Suresh, and K. Meena

Abstract In real world applications, people can find number of recipes online. Here using facial analysis method, food recipes rating application is developed. Some recipes might be authentic and few recipes might not be authentic. The recipes which you are choosing to order will never be similar as the visualized one. The quality and taste differs based on the preparation. So here a system is proposed, in order to get right recipes with healthy and good taste. The user has to select the variety and post the reviews of recipes online. All the recipes are reviewed, commented and rated by the visitors of the certain regional foods. So the customers may choose correct recipes online. These food items are being sorted according to the comments which are received from the users by applying facial analysis approach using machine learning named as Haar Cascade algorithm for mining the data. Here machine learning approach is used in which machine learns and analyses the facial expressions and emotions which reveals through the review. As the reviews are usually high in numbers, the visitors ends up with spending lots of time for searching non-casual food grounded on the peoples who participated by submitting reviews. This design will sort food items found on the bottom of facial analysis which represents the emoticon of the critics. The results will be shown with the help of a mobile application. The administrator will add watchwords and emoticon which are pertinent to the recipes in this application. In this methodology the watchwords which are matching up in the comments submitted by the visited customers and correspondingly the application will rate the recipes. On this application, based on the best-case emoticon reviews, the food will be listed on the top. This will make a more genuine

C. Shyamala Kumari (✉)

Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute Science and Technology, Chennai, India

e-mail: shyamalakumaric@veltech.edu.in

M. Pon Suresh

Department of Computer Science and Engineering, Kalasalingam Academy of Research and Education, Krishnankovil, Tamil Nadu, India

K. Meena

Department of Computer Science and Engineering, GITAM School of Technology, GITAM, Bengaluru, India

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023

815

P. P. Joby et al. (eds.), *IoT Based Control Networks and Intelligent Systems*,

Lecture Notes in Networks and Systems 528,

https://doi.org/10.1007/978-981-19-5845-8_59

review about the product. If this application has got implemented and installed in the device, further an update will be announced that users can review and view more than one particular product.

Keywords Recipes rating · Facial analysis · Haar Cascade algorithm · Reviewers

1 Introduction

The food recipes present on the paper will never come out exactly when it comes to the table. It never tastes as same as it looks. There are numerous channels which provide false reviews through their sites for the sake of money. To avoid such controversy the food must be rated by the end user. Even though there are many number of food recipes available online, as there are unauthenticated reviews, we are proposing a methodology, where the user can select grouping and post the food items. In order to find the quality recipe online, one must be visiting the review section which has been posted by visitors. Emotional Intelligence method and data abstraction approach is used in the process of mining the data. Emotional Intelligence is a data briefing tool in which the machine reads the reviews and analyzes the responses [1], emoticons [2] etc. about some data which is in text format like reviews on food varieties given by the customers who have tasted the food item in particular region. This application is based up on the food recipes rating using facial analysis [3, 4]. So the end user reaches out the correct food items through online. And we will sort according to the comments which are given by viewers by applying facial analysis [5] for mining data. Nowadays, Facial expression analysis is the latest approach which leads to the learning and analyzing of facial expressions [6, 7] and emoticons about the review given by the visitors. The existing system has some of the drawbacks like in the previous application user was not able to upload the recipes but they were able to comment and review the available food items which has already been uploaded by the admin.

2 Module Description

In order to get right recipes here is our proposed method, where user can select categories and post the recipes. Recipes are rated and commented by the visitors. So user may end up by finding correct recipe online. And it can be sort according to the comments which are given by viewers by applying facial analysis for data collection and mining process [8]. Facial analysis is a machine learning approach in which machine learns and analyze the facial expressions [9] and emotions [9] about the review.

2.1 Admin

Admin manages the whole system. Provides registration for the users to upload their own recipes. Planning and coordinating administrative procedures and systems. Ensures the smooth and adequate flow of information within the company to facilitate other business operations.

2.2 Registration

Users can register for this application by uploading two recipes. Once the registration is done then the users can upload any number of recipes.

2.3 Login

The users who already registered can login and upload their recipes.

2.4 List of Recipes

It contains all the recipes which are uploaded by the users and the admin. Every type of food recipe is available here in the list of recipes.

2.5 Search

It contains the search property where they can search the recipe they want to know.

2.6 Rating

It contains the rating for every recipe. Based on this rating we can come to know whether the recipe is good or bad.

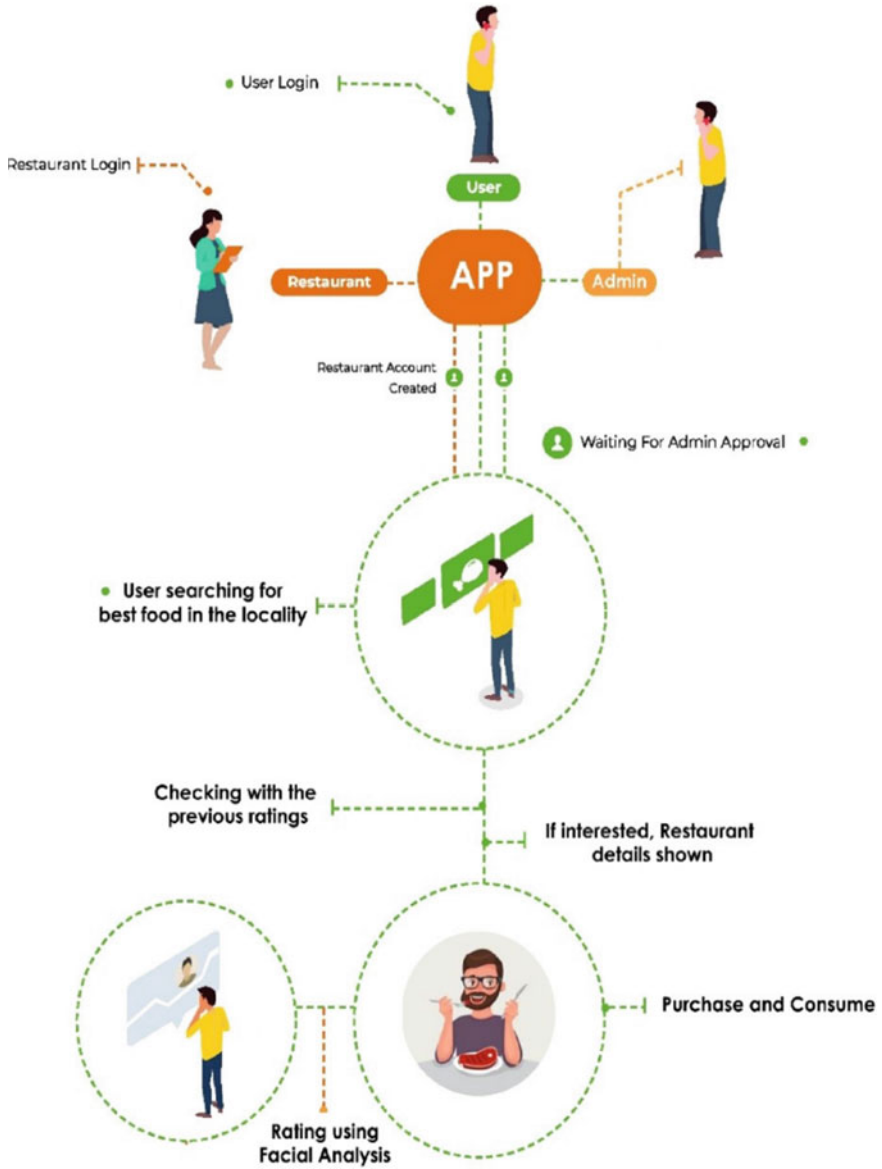


Fig. 1 Architecture diagram

3 Application Model

The three-dimensional design was there to improve code and content management and improve the performance of web-based applications. In the construction of the three dimensional blocks, there exists three stratum as shown in Fig. 1.

These are defined as follows...

- Demonstration.
- Trade Concept.
- Directory maintenance

3.1 Demonstration

The primary stratum demonstration carries particularly visible code, and this is exhibited to the user. This code might also comprise any generation that can be used on the purchaser side including HTML, JavaScript or VBScript and so on.

3.2 Trade Concept

The next stratum is trade concept which includes all of the server facet code. This stratum has a database touch code and query. This can manipulate, switch data to the user interface and manage properly any input from the UI.

3.3 Directory Maintenance

The third stratum data represents a records which are stored together with MSAccess, SQL Server, XML file, Excel document or textual content file containing statistics and other records information are inserted into the stratum. The directory with the above records has been maintained through the process of application.

4 Design Phase

Software design sits at the specialized kernel of the software engineering process and is applied no matter the event paradigm and area of play. Design is that the initiative within the development phase for any maneuvered product or system. The innovator's mark is to supply a model or representation of an integer which will thereafter be made. Incipency, once system necessary are specified and cut, system design is that the first of the three specialized exertion- design, law and test that's warranted to rear and validate software. The weightiness are often stated with one word "Quality". The skeleton was coming into thingness to refine superintendence law with contents and to refine the performance of the online rested uses. The login part stores the data of the user admin in an XML file. After logged into the app user can search for the best food in the locality and coding implementation for this part was done by Javascript. User can find the best food by it's rating and this sorting part was implemented by Quick sort algorithm. After tasting the food user can give his/her rating by showing his/her facial expression and this process is happening in the backend. Here 1265 dataset images of facial recognition is used to train the model in order to implementing the Local Binary Pattern Algorithm. Pattern is the most effective manner in which we are able to, as it should be translate and produced based on client perspectives into a completed product or software program [1] which serves as the idea for all destiny software program engineering steps as follows.

4.1 Process Diagram

The process diagram has three kinds of relationships as shown in Fig. 2. Association is the connection among the two classes. There may be a connection among both instructions if the instance of those elegance should recognize the other as a way to do its activity. In the above diagram, the organization is a link that connects the two sections. An association is an employer wherein one category belongs to a collection. The compound has a diamond give up that factors to the element that includes the whole thing. After that the hyperlink design that identifies one class is the bigger category of the alternative. A triangle pointing to a huge segment is done by generalization.



Fig. 2 Process diagram

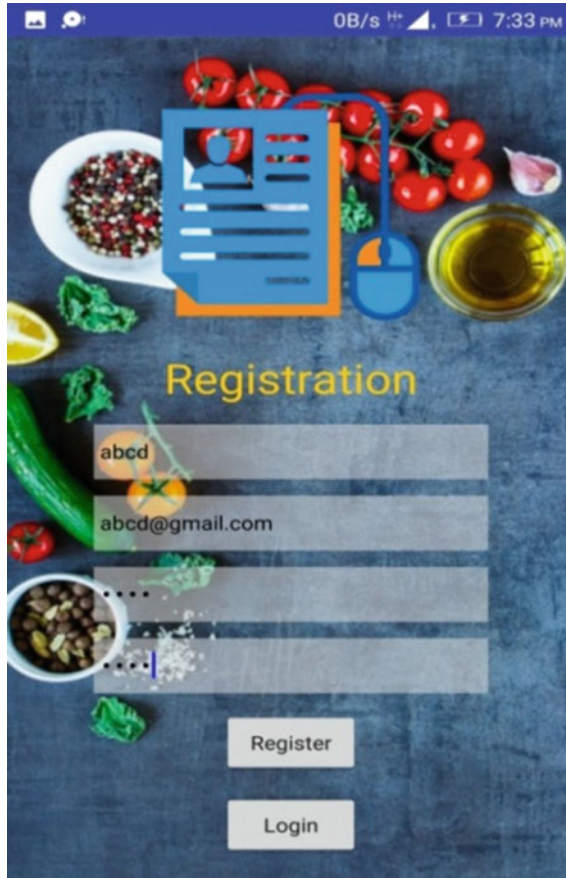


Fig. 3 Login page for recording rating and get results on rating

5 Implementation

5.1 Input and Output

The input is designed in the format of scanning the person face reaction with the help of camera access. Which is then post processed with help of LDA, HAAR algorithms and compares the reaction with pre-defined reactions [10]. HAAR algorithm is used because they're very fast at computing Haar-like features due to the use of integral images and is very efficient for feature selection. After the post process of the input, the output will be in the form of star rating [11], which is the rating of the particular food recipe. Accuracy level of the implementation is to be 85% of analysing the face and rating the food.

5.2 Login Page

The users can be the person who have tasted the food and the user who is registering to know the taste rate of food as shown in Fig. 3.

5.3 Food to Taste as Input of Application

After completing the login, food selection can be done to rate else food can be chosen for knowing the rate of it which is shown in below output screenshot as shown in Fig. 4.



Fig. 4 Input of application

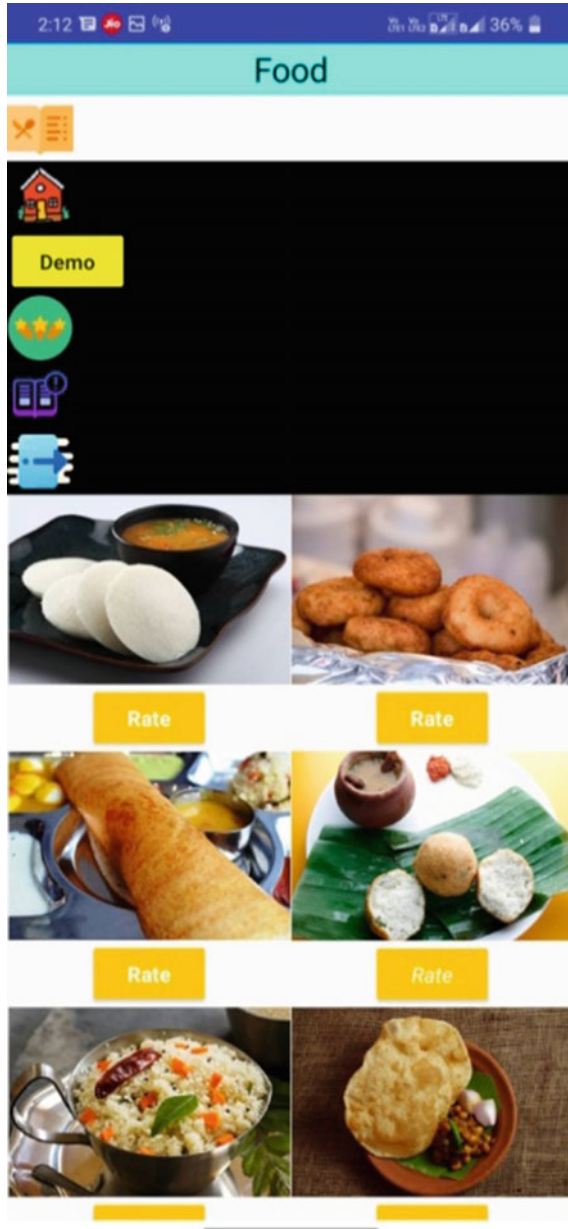


Fig. 5 Food varieties in region to rate

5.4 Food Varieties in Region to Rate for Taste

After the choice of food, the food rating can be viewed before getting into the restaurant or before having the food as shown in Fig. 5.

5.5 Face Recognition for Reaction Detection

The user of the application, after having the food in the particular region can give their rating based on their opinion through the camera as shown in Fig. 6 for facial expression recognition for the rate calculation. If the person is happy with the food, the smile expression can be taken for the rating as shown in Fig. 7, else it can be rated as lower range based on unhappy expression. Figure 8 shows the neutral rating when u switch on your camera or while you give permission for the app to use the camera [12] and if the neutral reaction continues while face reading is done the food taste is recorded as not so good and not so bad rating [13, 14]. Figure 9 shows the worst taste food in which the face reaction is read and rating is proved.



Fig. 6 Face recognition for reaction detection

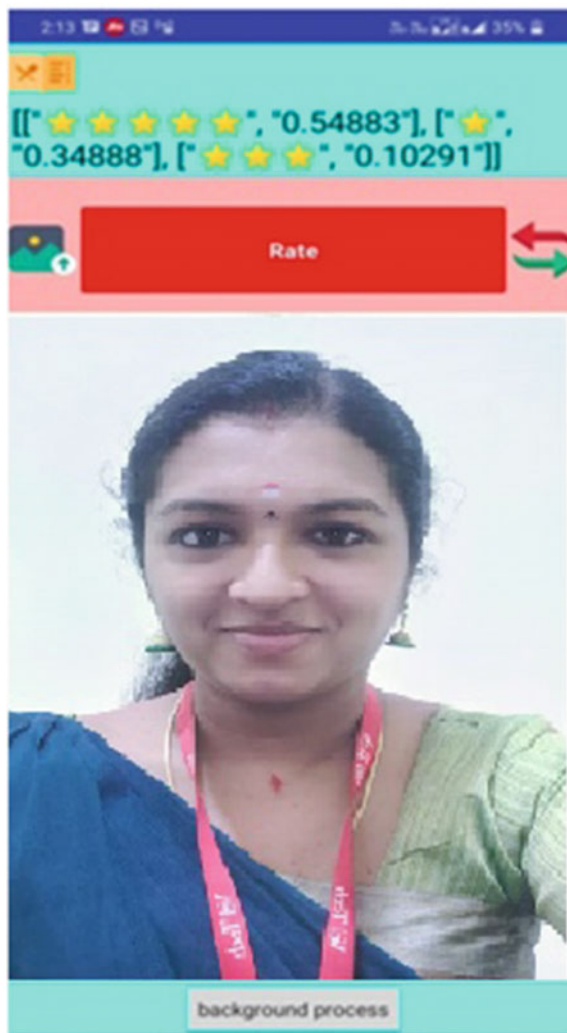


Fig. 7 Food rating based on the face reaction-good

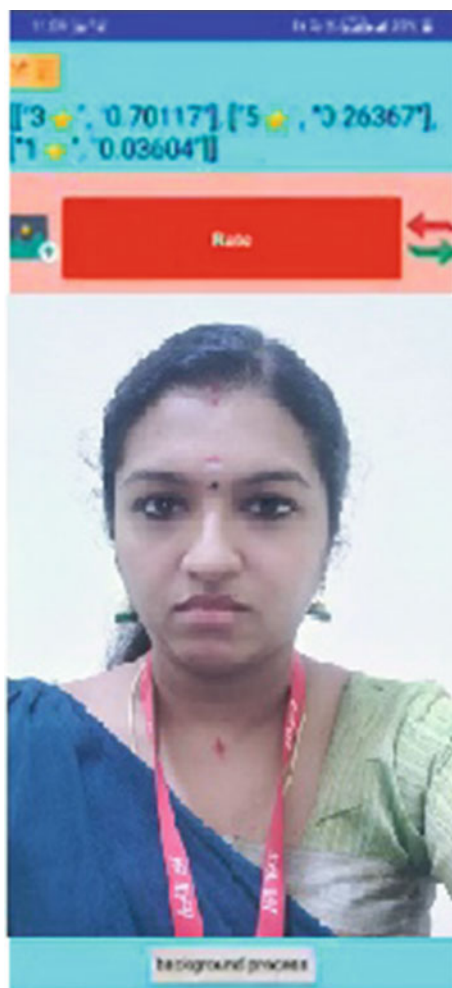


Fig. 8 Food rating based on the face reaction-neutral

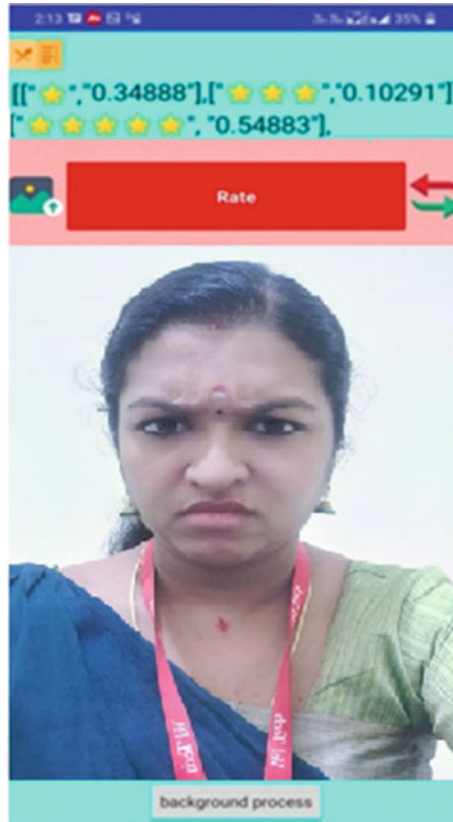


Fig. 9 Food rating based on the face reaction-worst

6 Conclusion

This application was mainly developed for the strangers of that particular location who is seeking for good healthy and tasty food and is very much user friendly and interesting. For that, they need to refer reviews and rating to know the best food. Here as we implemented an innovative and interesting method of facial analysis rating, people will come forward and give rating for the food based on their experience. So that they can find which food is famous over there. The proposed work contain the following add-on features which are detecting the eyes and lips along with cheeks movement for the analysis and based on the dimension added in facial expression, rating has been allotted. Additional feature is the rating range with star representation is improved and shading structure is included. So that they can find which food is famous over there. This helps the restaurants and hotels to find out which food is being liked by the customers and to it act accordingly. Also, kids might like it very much as it reflects the rating based on their expressions.

References

1. Raj JS, Vijesh Joe C (2021) Wi-Fi network profiling and QoS assessment for real time video streaming. *IRO J Sustain Wirel Syst* 3(1):21–30
2. Aruna S, Sasanka J, Vinay DA (2021) A brief study on analyzing student's emotions with the help of educational data mining. In: *Computer networks, big data and IoT*. Springer, Singapore, pp 785–796
3. Sharma R, Sungeetha A (2021) An efficient dimension reduction based fusion of CNN and SVM model for detection of abnormal incident in video surveillance. *J Soft Comput Paradigm (JSCP)* 3(02):55–69
4. Pandian AP (2021) Performance evaluation and comparison using deep learning techniques in sentiment analysis. *J Soft Comput Paradigm (JSCP)* 3(02):123–134
5. Kottursamy K (2021) A review on finding efficient approach to detect customer emotion analysis using deep learning analysis. *J Trends Comput Sci Smart Technol* 3(2):95–113
6. Patel K, Mehta D, Mistry C, Gupta R, Tanwar S, Kumar N, Alazab M (2020) Facial sentiment analysis using AI techniques: state-of-the-art, taxonomies, and challenges. *IEEE Vol 8:90495–90519*
7. Wang W, Xu K, Niu H, Miao X (2020) Emotion recognition of students based on facial expressions in online education based on the perspective of computer simulation. *Complexity* 2020. Article ID 4065207, 9 pages
8. Lu J, Plataniotis KN, Venetsanopoulos AN (2019) Face recognition using LDA based algorithms. *IEEE Neural Netw Trans*
9. Cruzal JEC, Shiguemorib EH, Guimar LNF (2019) A comparison of Haar-like, LBP and HOG approaches to concrete and asphalt runway detection in high resolution imagery. In: *Pan-American association of computational interdisciplinary sciences*
10. Shyamala Kumari C, Prema K, Florence S, Leema Priyadarshini L (2019) Enhancement of smart banking using biometrical security. *J Adv Res Dyn Control Syst* 11(1 Special Issue):609–611
11. Florence S, Kumari CS, Durai S (2018) Smart attendance marking system using face recognition. *J Comput Theor Nanosci* 15:2818–2821
12. Prema K, Leema Priyadarshini L, Shyamala Kumari C, Florence S (2018) Human intention detection with facial expressions using video analytics. *Int J Eng Technol (UAE)* 7(2):14–16
13. Prema K, Priyadarshini LL, Kumari CS (2018) Survey in detecting human beings behavior under video surveillances and its applications. *J Comput Theor Nanosci* 15(11–12):3550–3552
14. Torres AD, Yan H, Aboutaleb AH, Das A, Duan L, Rad P (2018) Patient facial emotion recognition and sentiment analysis using secure cloud with hardware acceleration, University of Texas at San Antonio, San Antonio, TX, USA
15. Boychuk V, Sukharev K, Voloshin D, Karbovskii V (2016) An exploratory sentiment and facial expressions analysis of data from photo-sharing on social media: the case of football violence. In: *ICCS-the international conference on computational science*, vol 80, pp 398–406
16. Rajakumari K, Nalini C (2015) Implementation of face recognition with LDA, PCA and Haar methods using 3d image based system. *Int J Innov Res Comput Commun Eng*
17. Michalski RS, Carbonell JG, Mitchell TM (2014) Machine learning – an artificial intelligence approach (Volume 1) in learning from observation, pp 345–360
18. Annuzzi Jr J, Darcey L, Conder S (2013) Introduction to android application development. In: *Android application design essentials*, pp 281–331
19. Viola P, Jones M (2001) Rapid object detection using a boosted cascade of simple features. In: *Accepted conference on computer vision and pattern recognition*

Survey of Text Document Summarization Based on Ensemble Topic Vector Clustering Model



G. Bharathi Mohan  and R. Prasanna Kumar 

Abstract Text document analysis has recently emerged as a promising strategy in content summarization domain. The analysis can be carried out in two stages: Text abstraction, the process of summarizing a document by including the most critical information from the original document; Text summarization, the process of generalizing redundant information in order to determine the significance of the issue. Automated text summarization is a technique used for extracting the most meaningful information from a document or group of related papers and assembling it into a concise version by retaining the overall meaning of the text document. The text abstraction model is mostly associated with the content extraction process, whereas to perform text summarization, the Natural Language Processing (NLP) technique is used to extract the necessary information from a lengthy text document. To perform efficient and automated text processing and summarization, this research study suggests a novel ensemble topic vector clustering technique, which utilizes Semantic Analysis (SA) to analyze the content. Further, the proposed study concentrates on the process of topic summarization to investigate various strategies and perform problem identification. Finally, the proposed study examines the significance of summarization implementation by comparing it with similar existing approaches.

Keywords Topic summarization · Abstraction · Extraction · Content analysis · Semantic analysis · Document modeling · Cluster models

G. Bharathi Mohan (✉) · R. Prasanna Kumar
Department of Computer Science Engineering, School of Computing, Amrita Vishwa Vidyapeetham, Chennai, India
e-mail: g_bharathimohan@ch.amrita.edu

R. Prasanna Kumar
e-mail: r_prasannakumar@ch.amrita.edu

1 Introduction

A summary is a descriptive text generated to deliver key information from the source material. The purpose of designing and developing an automated text summarization is to deliver the information present in the entire source material with a short descriptive text by including the semantics. The most significant advantage of developing a summary is that it limits the reading time. Text summarization techniques are generally classified into two types, they are extractive and abstractive. An extractive summary approach involves the extraction of key phrases and paragraphs from the source text document and concatenating them into a shorter version. Whereas, an abstractive summarization involves the understanding of the major concepts in a document and then convey those concepts in simple natural language.

1.1 Types of Summarization

This study focuses on the abstraction and extraction-based text summary to create a redundant text summary from an unstructured document. The procedure includes text data processing, training, and verification. The data mining techniques are used to generate the feature-based text summary. This type of text document analysis minimizes the dimensionality of forums, the attention mechanism for big data, and data learning. The fundamental examination of the sentence case clustered syntactic comparison of different text summary models offer the future direction for text summarization.

There are three main steps to carry out text summarization, they are identification, interpretation, and summary generation.

- **Topic Identification:** The most prominent information present in the document text will be identified. Different topic identification techniques are available, they are position, cue phrases, and word frequency. Methods that are proposed based on the position of phrases are the most useful methods for performing topic identification.
- **Interpretation:** All the developed abstract summaries should be processed through the interpretation step. In this step, different subjects will get featured to form a general content.
- **Summary Generation:** In this step, the system uses the text generation method to generate the final summary.

A) **Abstractive Summarization**

The abstractive summarization is the process of obtaining the crucial information from multiple documents and generating an accurate overview of the lengthy text document. This particular method has gained popularity for developing a new sentence to convey critical information from a text document. This process helps to generate a structured and easily understandable summarized text. Quality of

readability and language are the major advantages of utilizing the abstractive summarization technique.

B) Extractive Summarization

Extractive text summarization involves the selection of a subset of the sentence present in the source text for generating a summary. This type of summarization includes the most important sentences/phrases present in the source text document. This method can process single or multiple documents by understanding its underlying theme and create a better summarized descriptive text. A summarizer performs three relatively independent tasks, they are as follows:

- 1) Build an intermediate representation of input text.
- 2) Score will be provided based on the sentence mentioned in the demonstration.
- 3) Select the summary and include many sentences.

2 Automated Text Summarization

Automated Text Summarization (ATS) system plays a crucial role in the Natural Language Processing (NLP) domain. A small source text generation identifies users and resources that have information connected to the original text. Text compression stage is classified based on different characteristics. The main issue with these high-dimensionality scales is the residual number of document inputs. Figure 1 depicts the segmentation of text and editing the classification tasks into single and multiple document editing. When there is a need to process a collection of papers, the summarization model concatenates several document summaries into a single document summary.

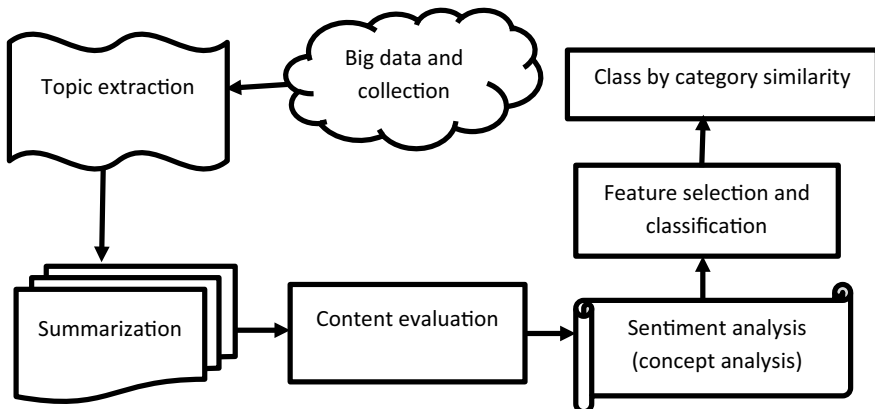


Fig. 1 Process of text classification

Figure 1 shows a brief overview of various recent technologies, as well as a more extensive description of the concept-based approach. When it comes to grouping high-dimensional data, the dimensionality challenges should be considered.

2.1 Important Steps for Text Summarization

To carry out the process of text summarization, a document will be processed in four main steps, they are topic identification, interpretation, summary generation, and extractive text summarization.

Topic Identification: This step identifies the essential information present in the text. Based on the position of the terms, this step identifies the topic.

Interpretation: Abstract summary is compiled through as many steps as required. This step combines different themes to form a standard content.

Summary Generation: In this step, the system creates a descriptive text by compiling the important sentences and phrases.

Extractive text summarization: This process is classified into two steps: The pre-processing and processing step. The text representation based on the pre-trained structure is usually included in the recognition of the sentence boundary. Removal of stop words is a common process involved in it. The processing step's objective is to concentrate on its significance and obtain stem cells or cardinality for each letter.

2.2 Observations of Text Summarization Methods

The compiled literature presents the following key observations.

- Before extracting the summary, the main task is to find the critical information present in the lengthy text document.
- The lengthy extracted statement sometimes contains non-essential information from the original document.
- It is not always feasible to capture all of the information included in the document in a summary of the data extraction.
- Summarization method might recommend unnecessary information.
- Extraction-based text summaries are undesirable because the content analysis is too sophisticated to predict equality, e.g., LSTM minimizes the corresponding format that is not close to the local rules of clustering.
- A summary of the extracted content from different text parts suddenly changes the document's subject, leading to a lack of workflow.
- In some cases, it is impossible to obtain semantic relationship between essential words present in a document via abstract generalization method.

- It is necessary to generate a summary that generalizes the NLP rule.
- Abstract semantic understanding requires a text. Depending on the abstract, it summarizes the quality of the profound language skills.

2.3 Features for Extractive Text Summarization

The extraction method is used by the majority of existing automated text summarization systems to generate a summary. Extraction summaries are frequently produced by using sentence extraction techniques. One way for obtaining appropriate sentences is to provide a numerical measure of a sentence for the summary, this process is known as sentence scoring, which then proceeds by selecting the best sentences to construct a document summary based on the compression rate. The compression rate is a crucial aspect in the extraction process since it defines the ratio between the length of the summary and source text. As the compression rate increases, the summary will become larger, and more insignificant content will be included. While the compression rate decreases, the summary becomes very short and as a result, more amount of information will be lost. When the compression rate is in the range of 5–30%, the summary quality will become acceptable.

3 Summarization Reviews and Their Implementations

Automatic text summarization intends to reduce the document text size by building a brief and high volume summary comprising the most important ideas present in that document. Through the years, many approaches have been proposed to improve the automatic text summarization results from which the graph-based method for sentence ranking is considered as one of the most important approaches in this field. The representation of text summarization tends to create a text document for retaining the general meaning of the significant information present in the input text. Text summarization approaches are classified into two types: abstraction and extraction. Abstract-based summarization creates a general concept by creating short sentences. The extractive-based summarization selects the informative phrases based on a specific evaluation carried out on the source document [1]. The main challenge is to determine the essential words, which could be included in the summary.

A clustering-based summary method reduces the text size and preserves its informative content [2]. However, it did not provide accurate results for text summarization obtained from the source text, and it faces an information overload challenge. Further, it does not relate to the homogeneous semantic similarity of the source text.

Automatic text summarization was used to extract and identify useful information from the source text to perform sentence ranking based on their weights. Automatic text summarization falls into two categories: extraction-based text summarization and abstraction-based text summarization [3]. The extractive-based text summarization

extracts the source text's important sentence or informative phrases. Abstraction-based text summarization extracts the useful information from the input document or source text. The extracted sentences will then be ranked based on their weights. However, the automatic text summarization will not provide a proper extraction result for ranking a sentence.

The Hidden Markov Model (HMM) was used for performing two important tasks: information sorting and extraction [4]. The first goal was to summarize the information, select the sequence to present the set of selected elements before constructing the conceptual text, summarize the multiple papers, and an important step in integrating the text into other issues. The second task was to extractively summarize the document by selecting the order of the sentences. This approach produces inaccurate input or source text results. In [5], authors are focused on developing an automatic text summarizer to extract the primary key of the input text document, and this extraction sentence uses the thesaurus to reduce the text size. The authors have mainly focused on reducing the text size to provide a satisfactory result.

Restricted Boltzmann Machine and Fuzzy Logic (RBMFL) algorithm selects an introductory sentence from the source text to extract the meaningful information. The first step of a pre-processing model was to remove stop words, i.e., tokenization. After pre-processing, the text feature extraction step is used to find the sentence score, where the RBMFL algorithm selects the highest from the source text. The increase of the text data on Internet allows the user to summarize the large amount of information, which will take more time to analyze and does not provide accurate results [6].

In [7], authors have introduced a cluster-based score calculation in the document based on latent characteristics. Then, the sentence-based scores trained into a regression model is used for improving the rouge scores based on the lexical terms, which has formulated a new summary. This process consumes more time to analyze the multi-document summarization models. In [8], authors have focused on performing a text summarization on online product reviews to detect the best products by using the definition representation technique. The paper is primarily focused on shortening the long sentences present in the online product reviews by utilizing the basic to advanced techniques i.e., Sequence2Sequence (S2S) with Long Short-Term Memory (LSTM). This method did not provide accurate results for summarizing the online product reviews.

The automatic text summarization reduces the text size and generates a summary from the original text as suggested by [9]. Different challenges are encountered during the text-reduction process. There are different automatic text summarization methods, they are: (i) Term Frequency (TF)-based method, (ii) Graph-based method, (iii) Time-based approach, (iv) Clustering-based approach, and (v) Topic-based approach. TF based method intends to discover the frequency of input to calculate the score. The graph-based method extracts the relevant information from the source text by using word graphs and characters, and it is also called as the text rank method. The time-based method is a concentrated form of the series document, which is used for extracting the summary object. The text summarizes the basics of access properties that depend on the length of the document. The time-based

method depicts how the strategy changes from time to time by using a sequence of different documents. The clustering-based method extracts the information from multiple document sources to create a single-document summarization. This method provides the main content of the summary from the source text. The topic-based method measures the topic themes for gathering the related information from source text.

Researchers [10] have introduced a two-level transformer for performing abbreviated text abstraction by using the Bidirectional Encoder Representations from Transformers (BERT) LSTM algorithm. In first stage, the input source text gets divided into segments by using the LSTM method, and then in the second stage, text segmentation is performed by extracting the most important sentence present in each section by using BERT. This method provides low accuracy and it does not offer the required performance.

The [11] automatic text summarization techniques are used to solve the overload problem by utilizing traditional S2S neural stretches and headline-aware decoder. An embedded encoder encodes the input text sentence structure and syntactic word information, and the decoder creates the quality of the headlines attention summary.

Table 1 compares the identification of rouge scores using the CNN/Daily Mail dataset to existing approaches. According to [7], different process and rouge values are used to assess components change at different levels. Text processing is the study of large amounts of text data that require an efficient application. This issue can be solved by using an automated abstraction system. As a result, the existing automatic summarization system need innovative solutions to satisfy the expanding user demand for data. A neural network-based abstract text summary technique is presented to collect data on creating text summaries and reducing the text size of the input text [12].

In [13], fuzzy logic, multi-feature, and Genetic Algorithm (GA) are used for performing an automatic news text summarization. The first process is to extract the key features of the word, such as news-related place, characters, and time, and scores given for the removed words by performing multi-feature analysis. Then, each feature extracted sentence calculates the weights by using GA. Finally, fuzzy logic was used to calculate the concluding scores of the automatic news text summarization. However, this process consumes more time and it will not provide accurate results for the news text.

Table 1 Identification of rouge scores using CNN/Daily mail dataset

Method	Rouge-1 (%)	Rouge-2 (%)	Rouge-n (%)
LSTM	28.1	20.4	27.4
TF	30.5	21.3	26.3
Time-based	34.6	18.6	35.7
Clustering-based	38.5	16.3	35.3
S2S based encoder and decoder	39.15	19.34	36.21

The author of [14] claimed that automated summarizing decreases the original content while improving student lecture material knowledge. The crucial text and essential pages were retrieved from course materials by using automatic text summarization. However, with the constrained student's attention, this is not necessarily a desirable outcome.

Based on the Structure Cosine Similarity (SCS) Document compression method, clustering generates a single document text summary to improve readability and consistency as explained in [15]. ExDoS is the first approach used to combine both supervised and unsupervised algorithms in a single framework and an interpretable approach is used for document summarization purpose.

Most of the researchers have explored the extractive text summarizer for performing text summarization by using Latent Semantic Analysis (LSA). LSA reduces long text to short text; it also extracts relevant content of the topic, and BERT uses sentences to retrieve the information [16]. Table 2 describes the reviews and limitations of the text summarization by using different techniques and methods.

The evaluation process initiated by confusion matrix [17] is where the LSA delivers best results and varies from other approaches. The topic model was utilized

Table 2 Reviews and limitations

Author name	Year	Title	Techniques used	Limitations
Y. Du et al. [13]	2020	News Text Summarization based on Multi-Feature and Fuzzy Logic	Fuzzy logic and Multi-feature	This method took more time and did not provide accurate results of news text
M.-H. Su et al. [10]	2020	A Two-Stage Transformer-based Approach for Variable-Length Abstractive Summarization	BERT and LSTM	This method provided low accuracy and did not offer the appropriate performance
R. Boorugu et al. [8]	2020	A Survey on NLP based Text Summarization for Summarizing Product Reviews	S2S and LSTM	It did not provide accurate results while processing the online product reviews
N. S. Shirwandkar et al. [6]	2018	Extractive Text Summarization using Deep Learning	RBMFL	It consumes more time to analyze and did not provide accurate results
H. Gupta et al. [16]	2021	Method of Text Summarization using LSA and Sentence based Topic Modelling With BERT	Latent Semantic Analysis (LSA) and BERT	It did not provide accurate results

for performing topic recognition and tracking. The main purpose of this model is to perform detection and monitoring. However, without clustering, the extended LM does not face the same difficulty.

In [18], authors have introduced the optimization by using text summary-based fusion and the concept of fine-tuning BERT topic information. Initially, a significant amount of consideration is given to the topic information for generating the abstract, wherein the keyword's topic was extracted as a part of the input and then merged with the source content. Secondly, the calculated semantic similarity between the summary and the original text was similar to the source material, and the quality of abstraction was also increased. However, the focus of the text summary research was mostly dedicated to abstract generalization..

The CCTSenEmb model described in [19] has identified the Gaussian topic to know more about the information concealed in the embedding space. Integrating both sides of the case seamlessly, the statement gets embedded in a new framework to abstract the system. To facilitate sentence embedding, based on the semantically coherent task frame and predictive sentence, CCTSenEmb has considered the link between adjacent sentences. These are the tasks where complex words and the typical representative did not meet the overall requirements of the application.

Random Cluster and L-length random walk are the optimized random walk methods shown in [20]. The strategy involved in simulating the influence of the topic social network on a large number of users and identifying a limited set of key points on behalf of the user. A user representative was selected, and the influence of the further distributed social overview via a social network, such as subjects, was often accepted. The social connections between users influence and user inquiries, include the effects of social concerns on the social network that incorporates both user and query.

In [21], a new summary task to preserve the sequence of displaying information concurrently was offered. However, in many circumstances, some users, particularly new users, found it difficult to comprehend the most recent issue in the face of a problem that was not difficult, overpowering, or structured.

The topic anatomical model TSCAN is used to derive the main topic from the eigenvectors of the association matrix of the time block proposed in [22]. Then, the summary was subjected to noteworthy events and was extracted by checking the configuration of a feature vector. Finally, to form a map of the evolution theme, the removed event was related to similar proximity and time background. However, the major problem was that, the information in the block was usually not sufficient to determine the block interrelationship.

An unsupervised probability generation model, in short called query topic sum will be generated to characterize the quotation process by using the LDA style model [23]. Additionally to identify the Cited Text Span (CTS) from the reference, it must use the quotation. This method has provided an insight for highlighting the importance of the annotated bibliography by academia. The generation of automobile-related work will generate a scientific overview of the tasks related to scientific papers given in a multi-document.

The work in [24] has been explained before implementation, where the evaluation method of graph-based, and the chart enhancement of the current document constraints are combined to determine the meaning of the sentence to declare. It controlled the most prominent statement in the present document, and made sure that it updates the previous record. Then the problem was the attention of the text in the current document that becomes possible to establish and interfere with the document earlier.

The LIM Topic has been incorporated into the modeling of topics based on the importance of the link [25]. In the framework, for example, Rank Topic and HITSTopic are included into the PageRank and local HITS of the local The ranking approach was specifically utilized to evaluate the geographical relevance of the first document. However, some topic models identify the relevance of another subject file.

The emerging framework explained in [26] has improved the ranking result of the text by using the clustering result of the sentence. Under this framework, it proposes to create a new method to rank the directly integrated cluster. However, the cluster-based way of the summary led to incomplete or sometimes biased result, and can be further applied to clustering and rank separation of the resulting analysis.

In [27], researchers provide a brief introduction and a standard definition on fuzzy ontology and its uses. The image of ontology extension is always unclear. To overcome the problem of uncertainty inference, information from other domains might be used to describe a domain ontology. An abstract, which is often closely connected to the content, may not be included.

The fuzzy logic method was used to extract the sentence with a critical function based on fuzzy logic [28]. The pre-processing step prepares each document by segmenting each phrase, removing stop words, and stemming words. However, determining the difference between insignificant and significant functions is a difficult, unreliable, and uncertain task.

In [29], authors have described new general techniques for improving the effectiveness of the search engine. A partially hierarchical output summary structure includes a specific file. Both configuration information and content are displayed in two summaries different from previous methods, which have been selected in a query bias mode. However, the existing search engines such as Google are composed of query terms and the surrounding text. They display short text fragments and two rows of the search results.

The abstract-based application is one of the most advanced compression systems that articulate item-package-based abstraction (items), which can be used in E-learning environments, was explained in [30]. For most of the part, the results and student expectations show that these are reflected in the automatically generated abstract. Therefore, it can support the learning activities in computer science courses at the university level. However, without comment, most of the existing abbreviations generally consider the document's rationale and lack user capabilities.

In [31], the Embra system was used when represented by the term co-occurrence matrix of the singular value decomposition. This improved reliability performance

indicates the present invention method. However, to extract a query-oriented multi-document summarization system, it is essential to calculate the relevance and redundancy to shape the meaning of a text.

The soft computing technique [32] is then defined by receiving the triangular coefficient of local accuracy mentioned in global precision and recall measurement. The value is then called to create a set of keys, which can generate a summary of methods used as input. It refers to merging or fusing the data and combining the common data problem.

Text classification was mainly concerned with a neural network, which applies to the general problem of supervised inductive learning in text summarization [34]. A set of training documents, classified into one or more predefined categories learn to automatically organize new documents. Clustering constitutes a significant class of data mining algorithms. The algorithm attempts to automatically partition the data space into a set of regions or clusters [35]. The examples in the table are allocated deterministically or probabilistically—similar to the entropy weighting approach for high-dimensional sparse data subspace clustering. Various ontologies have recently been developed manually, semi-automatically, or automatically. Some are constantly enhanced.

In the case explained in [36], the ontology development process has involved too many resources and consume more time and by including enormous scale ontology, the quality of the ontology may get affected. Furthermore, the misplaced elements must be discovered to make the factorization of the ontology possible. The review study [36] has proposed a method for spotting the most susceptible misplaced features by using a natural language technique for finding text similarity. The selection of sentence features has improved the performance of extractive summarizers [37]. Another graph-based text summarizer has used a trigraph to extract the important trigrams. Trigrams and words are linked and mapped to generate a summary [38]. Gramer rules were employed in functionality-based machine learning algorithms to anticipate Tamil tweets appropriate for generating an abstractive summary [39]. In text summarization, Latent Dirichlet Allocation (LDA) was used to uncover hidden topics in documents [40]. Topic modelling assists in the selection of key sentences based on the extracted topics.

4 Problem Identification Factors

Accurately measuring semantic similarity between words is a major challenge for web mining, information retrieval, and natural language processing models. Web mining, network extraction, and applications such as identifying the ambiguity entities must quantify the semantic similarity between specific ideas and entities. One of the most difficult tasks in information retrieval is providing a group of documents to the user that are semantically related. A number of NLP activities, such as automated summarization rely on accurate assessment of semantic similarity between words,

Word-Sense Disambiguation (WSD), and text meaning. Based on the abstract derivation about the user with the approach of selecting the most relevant high data source keyword search, the graph that represents the keyword relationship in the underlying database is summarized.

5 Single and Multi-Document Text Summarization

Single Document Text Summarization (SDTS): This type of summarization reduces the usage of long sentences and extracts the important content from the original document. It has been used at the beginning of a text summary.

Multi-Document Text Summarization (MDTS): This is a process where multiple documents are provided as input for aggregation techniques since the input length for a particular topic is too long. When compared to the SDTS, it is aggregated. In many cases, it is not easy to merge multiple documents into a single file. The difficulty here is that there may be diverse themes in different files. Excellent summary of technology tends to condense the subject to maintain readability and integrity and consider the essential sentences.

5.1 Term Frequency-Inverse Document Frequency (TF-IDF) Method

The number of statistics reflect the importance of a word in a given document. TF-IDF value will proportionally increase with the number of words present in the document. This method is mainly applied to the weighted term frequency and inverse sentence frequency paradigm, where the sentence represent the frequency of a number of documents, where the project's sentence include. These sentence carriers are acquired through a similar sentence with the highest score as a part of the summary.

5.2 Cluster-Based Approach

This approach represents three semantic characteristics of a captured document (subject, verb, and object). The cluster can use the same information to express the detected three semantic characteristics in natural language. The semantic characteristics are considered to be the basic unit of summary process. More similar to these three characteristics, the information becomes uselessly repeated; thus, it can be constructed by using the sequence summary sentences associated with the calculated clusters.

5.3 Automatic Text Summarization Based on Fuzzy Logic

The fuzzy logic method considers the length of the input sentence present in fuzzy systems like a slight similarity, similar terms, and each text feature. Next, the information will enter all the rules necessary for the aggregation of knowledge-based system. Subsequently, a method for obtaining a value from zero to one in each sentence is based on the sentence output characteristics and available knowledge base of rules. The degree of importance in determining the value of the statement's output is obtained in the final summary.

6 Result Analysis and Discussion

Results are analyzed from other recent research works obtained from various author-handled text summarization techniques. The dataset is collected from various document forums like Doc-CCN/Daily (1), Gigaword corpus (2), and Reuters corpus (3).

Table 3 describes the highest data score that has been selected for each sub-topic [15]. To define the categories of the dataset and topic summarizations determined by different algorithms, redundancy is removed from both the cases to generate each sub summary. For each data selected as a sub-summary, the dataset is compared before sub summaries.

Table 4 describes the sequential summary. It compares the dataset with readability, sequence, and novelty along with the evaluation, baseline, and generated values. This represents the sequential summary obtained from the baseline observation of readability, series, and novelty.

Table 5 defines the analysis of the text summarization results, which are compared with other algorithms. The table describes the LSTM, Fuzzy logic, RBMFL, and LSA analysis of text summarization results, where, P denotes Precision, R denotes Recall, and F1 represents F1-measure.

Table 3 Categories of the dataset and topic summarizations

Categories	No. of topics	Example of topics
Science	6	CHINA
News	5	Libya release
Technology	2	Apple iPhones
Lifestyle	8	Goose Island
Entertainment	3	Club
Sports	4	Bbcfl
Total	28	-----

Table 4 Evaluation of sequential summary (%)

Sequential summary	Baseline	Evaluation	Generated
Readability	2.10	3.15	5.02
Sequence	1.15	2.12	5.10
Novelty	3.10	3.10	6.12

Table 5 Comparative analysis of text summarization results (%)

Doc	LSTM (%)			Fuzzy logic-based clustering (%)			RBMFL (%)			LSA (%)		
	P	R	F1	P	R	F1	P	R	F1	P	R1	F1
1	59	60	61	62	63	64	65	66	67	68	69	70
2	60	61	62	63	64	65	66	67	68	69	70	72
3	65	68	70	71	74	75	76	77	79	80	83	83

Table 6 Evaluation of the difference between the changed topic and the original topic

Divergence	AVG	Successive rate	Sum	Max
LIMTopic	1.12	13.33	16.10	20.12
Clustering	0.8	12	10.02	12.02
Parameter learning	0.9	11.15	0.9	10.05
HITS	0.2	10.90	0.05	0.9

Table 6 describes how conversion topic distribution is calculated from the topic model while the original one is calculated from the link structure. The difference between the changed topic and the actual topic dependencies is based on text relevance, average value, successive rate, sum, and mean max rate. The contents of the document text will not reflect various aspects of the document grouping on related texts.

7 Conclusion

The automated text summarization technique can be implemented in the emerging fields of biomedical engineering, product evaluations, academics, emails, and blogs. A huge source of information is available in these areas, particularly on the World Wide Web. From the study, it is concluded that Natural Language Processing (NLP) models have the ability to generate a summary of one or more texts automatically. Text summarization using a Neural Network, a graph-based technique, fuzzy logic, and a cluster-based approach to some range resulted in a meaningful summary of the original text. Both extractive and abstractive methods have been investigated

in this research study. The extractive method underpins the majority of summary techniques. The abstractive process is equivalent to how humans reach conclusions. Abstract generalisation currently necessitates extensive machine learning language generation and is difficult to replicate in a given subject area.

References

1. Andhale N, Bewoor LA (2016) An overview of text summarization techniques. In: 2016 international conference on computing communication control and automation (ICCUBEA), pp 1–7. <https://doi.org/10.1109/ICCUBEA.2016.7860024>
2. Zhang P, Li C (2009) Automatic text summarization based on sentence clustering and extraction. In: 2009 2nd IEEE international conference on computer science and information technology, pp 167–170. <https://doi.org/10.1109/ICCSIT.2009.5234971>
3. Madhuri JN, Ganesh Kumar R (2019) Extractive text summarization using sentence ranking. In: 2019 international conference on data science and communication (IconDSC), pp 1–3. <https://doi.org/10.1109/IconDSC.2019.8817040>
4. Barzilay R, Lee L (2004) Catching the drift: probabilistic content models, with applications to generation and summarization. In: Human language technology conference of the North American chapter of the association for computational linguistics, pp 113–120
5. Patil P, Dalmia S, Abu Ayub Ansari S, Aul T, Bhatnagar V (2014) Automatic text summarizer. In: 2014 international conference on advances in computing, communications and informatics (ICACCI), pp 1530–1534. <https://doi.org/10.1109/ICACCI.2014.6968629>
6. Shirwandkar NS, Kulkarni S (2018) Extractive text summarization using deep learning. In: 2018 fourth international conference on computing communication control and automation (ICCUBEA), pp 1–5. <https://doi.org/10.1109/ICCUBEA.2018.8697465>
7. Celikyilmaz A, Hakkani-Tur D (2010) A hybrid hierarchical model for multi-document summarization. In: Proceedings of the 48th annual meeting of the association for computational linguistics, pp 815–824
8. Boorugu R, Ramesh G (2020) A survey on NLP based text summarization for summarizing product reviews. In: 2020 second international conference on inventive research in computing applications (CIRCA), pp 352–356. <https://doi.org/10.1109/ICIRCA48905.2020.9183355>
9. Bhatia N, Jaiswal A (2016) Automatic text summarization and its methods - a review. In: 2016 6th international conference - cloud system and big data engineering (Confluence), pp 65–72. <https://doi.org/10.1109/CONFLUENCE.2016.7508049>
10. Su M-H, Wu C-H, Cheng H-T (2020) A two-stage transformer-based approach for variable-length abstractive summarization. *IEEE/ACM Trans Audio Speech Lang Process* 28:2061–2072. <https://doi.org/10.1109/TASLP.2020.3006731>
11. Cheng J, Zhang F, Guo X (2020) A syntax-augmented and headline-aware neural text summarization method. *IEEE Access* 8:218360–218371. <https://doi.org/10.1109/ACCESS.2020.3042886>
12. Gaol SFL, Matsuo T (2021) A survey of the state-of-the-art models in neural abstractive text summarization. *IEEE Access* 9:13248–13265. <https://doi.org/10.1109/ACCESS.2021.3052783>
13. Du Y, Huo H (2020) News text summarization based on multi-feature and fuzzy logic. *IEEE Access* 8:140261–140272. <https://doi.org/10.1109/ACCESS.2020.3007763>
14. Okubo SF, Yin C, Ogata H (2018) Automatic summarization of lecture slides for enhanced student preview technical report and user study. *IEEE Trans Learn Technol* 11(2):165–178. <https://doi.org/10.1109/TLT.2017.2682086>
15. Ghodrathnama S, Beheshti A, Zakershaharak M, Sobhanmanesh F (2020) Extractive document summarization based on dynamic feature space mapping. *IEEE Access* 8:139084–139095. <https://doi.org/10.1109/ACCESS.2020.3012539>

16. Gupta H, Patel M (2021) Method of text summarization using LSA and sentence based topic modelling with Bert. In: 2021 international conference on artificial intelligence and smart systems (ICAIS), pp 511–517. <https://doi.org/10.1109/ICAIS50930.2021.9395976>
17. Yang P, Li W, Zhao G (2019) Language model-driven topic clustering and summarization for news articles. *IEEE Access* 7:185506–185519. <https://doi.org/10.1109/ACCESS.2019.2960538>
18. You F, Zhao S, Chen J (2020) A topic information fusion and semantic relevance for text summarization. *IEEE Access* 8:178946–178953. <https://doi.org/10.1109/ACCESS.2020.2999665>
19. Gao Y, Xu Y, Huang H, Liu Q, Wei L, Liu L (2020) Jointly learning topics in sentence embedding for document summarization. *IEEE Trans Knowl Data Eng* 32(4):688–699. <https://doi.org/10.1109/TKDE.2019.2892430>
20. Li J, Liu C, Yu JX, Chen Y, Sellis T, Culpepper JS (2016) Personalized influential topic search via social network summarization. *IEEE Trans Knowl Data Eng* 28(7):1820–1834. <https://doi.org/10.1109/TKDE.2016.2542804>
21. Gao D, Li W, Cai X, Zhang R, Ouyang Y (2014) Sequential summarization: a full view of twitter trending topics. *IEEE/ACM Trans Audio Speech Lang Process* 22(2):293–302. <https://doi.org/10.1109/TASL.2013.2282191>
22. Chen CC, Chen MC (2012) TSCAN: a content anatomy approach to temporal topic summarization. *IEEE Trans Knowl Data Eng* 24(1):170–183. <https://doi.org/10.1109/TKDE.2010.228>
23. Wang P, Li S, Zhou H, Tang J, Wang T (2020) ToC-RWG: explore the combination of topic model and citation information for automatic related work generation. *IEEE Access* 8:13043–13055. <https://doi.org/10.1109/ACCESS.2019.2959056>
24. Li X, Du L, Shen Y (2013) Update summarization via graph-based sentence ranking. *IEEE Trans Knowl Data Eng* 25(5):1162–1174. <https://doi.org/10.1109/TKDE.2012.42>
25. Duan D, Li Y, Li R, Zhang R, Gu X, Wen K (2014) LIMTopic: a framework of incorporating link based importance into topic modeling. *IEEE Trans Knowl Data Eng* 26(10):2493–2506. <https://doi.org/10.1109/TKDE.2013.2297912>
26. Cai X, Li W (2013) Ranking through clustering: an integrated approach to multi-document summarization. *IEEE Trans Audio Speech Lang Process* 21(7):1424–1433. <https://doi.org/10.1109/TASL.2013.2253098>
27. Lee C-S, Jian Z-W, Huang L-K (2005) A fuzzy ontology and its application to news summarization. *IEEE Trans Syst Man Cybernet Part B (Cybernet)* 35(5):859–880. <https://doi.org/10.1109/TSMCB.2005.845032>
28. Suanmali L, Salem M, Binwahlan, Salim N (2009) Sentence features fusion for text summarization using fuzzy logic. *IEEE*, pp 142–145
29. Canan Pembe F, Güngör T (2007) Automated query-biased and structure-preserving text summarization on web documents. In: *Proceedings of the international symposium on innovations in intelligent systems and applications, Istanbul*
30. Baralis E, Cagliero L (2016) Learning from summaries: supporting e-Learning activities by means of document summarization. *IEEE Trans Emerg Top Comput* 4(3):416–428. <https://doi.org/10.1109/TETC.2015.2493338>
31. Hachey B, Murray G, Reitter D (2006) Dimensionality reduction aids term co-occurrence-based multi-document summarization. In: *SumQA 2006: proceedings of the workshop on task-focused summarization and question answering*, pp 1–7
32. Van Britsom D, Bronselaer A, De Tré G (2015) Using data merging techniques for generating multidocument summarizations. *IEEE Trans Fuzzy Syst* 23(3):576–592. <https://doi.org/10.1109/TFUZZ.2014.2317516>
33. Mathew R, Gupta H, Patel M (2021) Method of text summarization using LSA and sentence based topic modelling with Bert. In: 2021 international conference on artificial intelligence and smart systems (ICAIS), pp 511–517. <https://doi.org/10.1109/ICAIS50930.2021.9395976>
34. Manoharan JS (2021) Capsule network algorithm for performance optimization of text classification. *J Soft Comput Paradigm (JSCP)* 3(01):1–9

35. Haoxiang W, Say S (2021) Big data analysis and perturbation using data mining algorithm. *J Soft Comput Paradigm (JSCP)* 3(01):19–28
36. Says S, Wang H (2021) Naïve Bayes and entropy-based analysis and classification of humans and ChatBots. *J ISMAC* 3(01):40–49 (2021)
37. Chiney RP, Prasanna Kumar R (2020) Extractive summarization approach for news articles based on selective features. *Int J Adv Sci Technol* 29:8215–8224
38. Raj D, Geetha M (2018) A trigraph based centrality approach towards text summarization. In: 2018 international conference on communication and signal processing (ICCSP), Chennai, India
39. Prasanna Kumar R (2021) Grammar rule-based sentiment analysis techniques for Tamil tweets classification using machine learning. *CNC Comput Mater Continua*
40. Bharathi Mohan G, Prasanna Kumar R (2021) A comprehensive survey on topic modeling in text summarization. In: 5th international conference on micro-electronics and telecommunication engineering, Springer book series on “Lecture Notes in Networks and Systems”

A Hybrid Approach on Conditional GAN for Portfolio Analysis



Jun Lu and Danny Ding

Abstract Over the decades, the Markowitz framework has been used extensively in portfolio analysis though it puts too much emphasis on the analysis of the market uncertainty rather than on the trend prediction. While generative adversarial network (GAN), conditional GAN (CGAN), and autoencoding CGAN (ACGAN) have been explored to generate financial time series and extract features that can help portfolio analysis. The limitation of the CGAN or ACGAN framework stands in putting too much emphasis on generating series and finding the internal trends of the series rather than predicting the future trends. In this paper, we introduce a hybrid approach on conditional GAN based on deep generative models that learns the internal trend of historical data while modeling market uncertainty and future trends. We evaluate the model on several real-world datasets from both the US and Europe markets, and show that the proposed HybridCGAN and HybridACGAN models lead to better portfolio allocation compared to the existing Markowitz, CGAN, and ACGAN approaches.

Keywords Synthetic series · Hybrid CGAN · Autoencoding conditional GAN (ACGAN) · Conditional GAN · Portfolio analysis and allocation · Sharpe ratio · Markowitz framework

1 Introduction

Financial portfolio management is largely based on linear models and the Markowitz framework [20, 21] though the underlying data and information in today's market has increased countless times over that of many years ago. The framework, often known as the modern portfolio theory (MPT), has become one of the cornerstones of quantitative finance. The fundamental idea behind the MPT is to create portfolio diversification while reducing specific risks and assessing the risk-return trade-offs

J. Lu (✉)
Trexquant, Stamford, USA
e-mail: jun.lu.locky@gmail.com

D. Ding
JJ Capital Fund, Beijing, China

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
P. P. Joby et al. (eds.), *IoT Based Control Networks and Intelligent Systems*,
Lecture Notes in Networks and Systems 528,
https://doi.org/10.1007/978-981-19-5845-8_61

849

for each asset. The MPT, on the other hand, has been criticized for making ideal assumptions about the financial system and data: the expected mean returns, and the covariance matrix of the return series are estimated from historical observations and assumed constant in the future. This is such a strong assumption, though, that it will be impossible for the market to actually achieve this demand.

As the evaluation outcomes of cross-section risk, the traditional portfolio assessment approach creates portfolio risk indicators based on asset price series over the previous period, such as variance, value at risk, and expected loss. The conventional method where the classical mean-variance optimization approach employed in MPT, however, has two clear shortcomings. First, historical data typically cannot be simply utilized to indicate the future due to the capital market's quick-change since financial returns are notoriously stochastic with an extremely low signal-to-noise ratio; When a reliable long-term prognosis is made available in a highly efficient market, traders immediately act on this forecast, which directly affects the price at hand; while future price variations are unpredictable again [1, 7, 10, 11, 24]. Second, the linear components in the historical series are typically all that are included in the risk measuring indicators evaluated using conventional methods, leaving out the nonlinear information. This causes a discrepancy between the evaluation results and the actual situation [25].

The financial sector, on the other hand, has been significantly impacted by advances in AI. Machine learning has been used in a variety of applications, including forecasting, series generation, risk management, customer service, and portfolio management [9, 12, 23]. Specifically, generative adversarial networks (GANs) are a sort of neural network architectures that have shown promise in image generation and are now being used to produce time series and other financial data [3, 5, 6]. Models of the ARCH and GARCH families, which use classical statistics to explain the change in variance over time in a time series by describing the variance of the current error component as a function of prior errors [2, 4, 18]. GANs are being used to address the problem of paucity of real data, as well as to optimize portfolios and trading methods which achieve better results [19, 23]. However, due to its highly stochastic, noisy, and chaotic nature, market price forecasting is still a major topic in the time series literature. While previous work have tried to generate financial data based on historical trend, the generation still lacks guidance on the potential trend of the future series [17, 19].

In this light, we focus on GANs for better portfolio allocation that can both capture historical trends and generate series based on past data. We present a novel framework about portfolio analysis based on conditional GAN (CGAN) that incorporates a proposer providing a potential mean value of future series for data normalization to achieve stable strategies, hence the name *HybridCGAN*. Similar to the CGAN and ACGAN models for portfolio analysis [17, 19], HybridCGAN can also directly model the market uncertainty via its complex multidimensional form, which is the primary driver of future price trends, such that the nonlinear interactions between different portfolios can be embedded effectively. We evaluate the proposed HybridCGAN method on two separate portfolios representing different markets (the US and the European markets) and industrial segments (e.g., Technology, Healthcare,

Basic materials, and Industrials sectors). The empirical results show that the proposed approach is capable of realizing the risk-return trade-off and outperforms the classic MPT, CGAN- and ACGAN-based methodologies considerably.

2 Related Work

As aforementioned, there are several methods delving with portfolio allocation, including the Markowitz framework, the CGAN and ACGAN methodologies [17, 19, 20]. The Markowitz framework relies on the assumption that the past trend can be applied in the future. While the CGAN and ACGAN methodologies partly solve the drawback in the Markowitz framework by simulating future data based on historical trends, it still lacks full ability to capture the information and features behind the past data. The proposed HybridCGAN (and HybridACGAN) model introduces an extra *proposer* that can help the constructed networks to capture historical features and propose future trends.

2.1 Markowitz Framework

Portfolio allocation is a kind of investment portfolio where the market portfolio has the highest Sharpe ratio (SR) given the composition of assets [20]. For simplicity, we here only consider long-only portfolio. Denote \mathbf{r} as the return on assets vector, Σ as the asset covariance matrix, \mathbf{w} as the weight vector of each asset, and r_f as the risk-free interest rate. If we measure portfolio risk by variance (or standard deviation), then the overall return and risk of the portfolio are:

$$r = \mathbf{w}^\top \mathbf{r}; \quad \sigma^2 = \mathbf{w}^\top \Sigma \mathbf{w}. \quad (1)$$

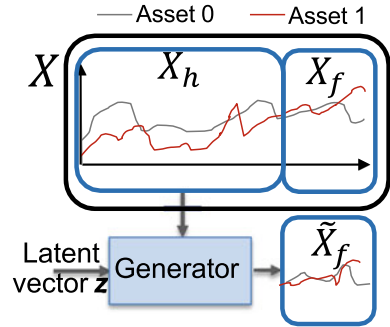
And the Sharpe ratio [22] can be obtained by

$$\text{SR} = \frac{r - r_f}{\sigma} = \frac{\mathbf{w}^\top \mathbf{r} - r_f}{\sqrt{\mathbf{w}^\top \Sigma \mathbf{w}}}. \quad (2)$$

According to the definition of portfolio allocation, the weight of each asset in the market portfolio is the solution to the following optimization problem:

$$\begin{aligned} & \arg \max_{\mathbf{w}} \frac{\mathbf{w}^\top \mathbf{r} - r_f}{\sqrt{\mathbf{w}^\top \Sigma \mathbf{w}}}; \\ \text{s.t. } & \sum_{i=1}^N w_i = 1; \quad 0 \leq w_i \leq 1, \forall i \in \{1, 2, \dots, N\}, \end{aligned} \quad (3)$$

Fig. 1 A conceptual overview of the CGAN, ACGAN, and the proposed HybridCGAN and HybridACGAN generators' inputs and outputs



where N is the number of assets, and w_i is the i -th element of the weight vector w .

2.2 Portfolio Analysis with GAN

We consider the matrix X to span the whole analysis length: $w = h + f$. The matrix X contains two components, the known historical series X_h of length h , and the unknown future X_f of length f . Given the number of assets N , the matrix X is of size $N \times w$; X_h has shape $N \times h$; and X_f is of shape $N \times f$.

Considering the (known) historical series $X_h \in \mathbb{R}^{N \times h}$ and a prior distribution of a random latent vector $z \in \mathbb{R}^m$, we use a generative deep-neural network G to learn the probability distribution of future price trends X_f within the target future horizon f . Figure 1 provides a conceptual representation of the matrix X , and the inputs and outputs of the generator G . Formally the generative model simulates a fake future matrix \tilde{X}_f by

$$\tilde{X}_f = G(z, X_h), \tag{4}$$

where $z \in \mathbb{R}^m$ is the latent vector sampled from a prior distribution (e.g., from a normal distribution). In practice, The unanticipated future events and phenomena that will affect the market are represented by the latent vector z . Based on the most recent market conditions, the known historical series X_h is used to extract features and condition the probability distribution of the future X_f . Given the historical observation X_h and following the Wasserstein GAN-GP (WGAN-GP) by Gulrajani et al. [8], the generative G is trained in adversarial mode against a discriminator network D with the goal of minimizing the Wasserstein distance between the real future series X_f and the fake series \tilde{X}_f . Formally, the procedure is described by the following optimization problem:

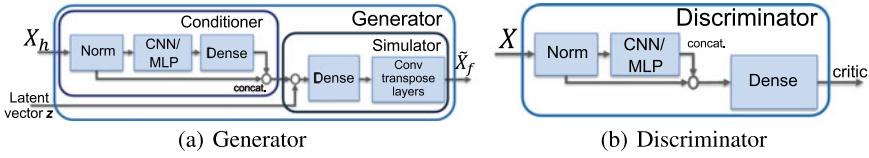
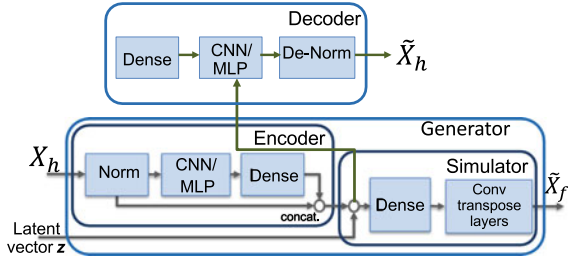


Fig. 2 Architectures of the CGAN generative and discriminative models for portfolio analysis

Fig. 3 Architecture of the ACGAN generative model for portfolio analysis



$$\begin{aligned} \max_D \quad & \mathbb{E}_{\mathbf{x} \sim p(\text{data})} \left\{ D(\mathbf{x}) - \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} [D(G(\mathbf{z}, \mathbf{x}_h))] \right\} \\ & - \lambda_1 \cdot \mathbb{E}_{\tilde{\mathbf{x}} \sim p(\epsilon \text{ data} + (1-\epsilon)G(\mathbf{z}))} [\|\nabla_{\tilde{\mathbf{x}}} D(\tilde{\mathbf{x}})\|_2 - 1]^2; \quad (5) \\ \max_G \quad & \mathbb{E}_{\mathbf{x} \sim p(\text{data})} \left\{ \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} [D(G(\mathbf{z}, \mathbf{x}_h))] \right\}, \end{aligned}$$

where \mathbf{x}_h contains the historical parts of the data \mathbf{x} ($\mathbf{x}_h \in \mathbf{x}$), $G(\mathbf{z}, \mathbf{x}_h)$ indicates that the generator depends on the (historical) data \mathbf{x}_h , and λ_1 controls the gradient penalty. Theoretically, the optimization process finds the surrogate posterior probability distribution $p(\tilde{\mathbf{X}}_f | \mathbf{X}_h)$ that approximates the real posterior probability distribution $p(\mathbf{X}_f | \mathbf{X}_h)$.

Discriminator. The discriminator shown in Fig. 2(b) (for both CGAN and ACGAN) takes as input either the real data matrix $\mathbf{X} = [\mathbf{X}_h, \mathbf{X}_f] \in \mathbb{R}^{N \times w}$ or the synthetic data matrix $\tilde{\mathbf{X}} = [\mathbf{X}_h, \tilde{\mathbf{X}}_f] \in \mathbb{R}^{N \times w}$.

The main drawback of the CGAN methodology is in that it puts too much emphasis on the *conditioner* to extract features that can “deceive” the discriminator (Fig. 2(a)). When the discriminator is perfectly trained, this issue is not a big problem. However, in most cases, especially due to the scarcity of financial data, the discriminator works imperfectly such that the conditioner may lose important information of the historical data.

2.3 Autoencoding CGAN

The Autoencoding CGAN (ACGAN) model partly solves the problem in the CGAN methodology, in which case find a balance between the information extraction and generation for cheating the discriminator via an embedded autoencoder providing the capability of keeping the intrinsic information of historical data [17]. The ACGAN model has the same discriminator structure as the CGAN. However, it contains an extra *decoder* in the generator as shown in Fig. 3. And therefore we call the conditioner an *encoder* in the ACGAN context.

We use an *encoding* deep-neural network E to learn the features that can help the generator cheat the discriminator and can find the internal information itself; and a *decoding* deep-neural network F to reconstruct the historical series so as to force the encoder to do so. Formally the encoding and decoding models reconstruct the historical matrix by

$$y = E(X_h), \quad \tilde{X}_h = F(y).$$

This process is known as *autoencoding*, hence the name autoencoding conditional GAN (ACGAN). In a non-GAN context, the autoencoder is typically done by non-negative matrix factorization or general matrix decomposition via alternative least squares or Bayesian inference [13–16]. Since we need to use the encoding part of the autoencoder to help trick the discriminator as well, the autoencoder is then constructed by deep-neural networks instead. Formally the process is described by the following optimization problem:

$$\begin{aligned} \max_D \quad & \mathbb{E}_{\mathbf{x} \sim p(\text{data})} \left\{ D(\mathbf{x}) - \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} [D(G(\mathbf{z}, \mathbf{x}_h))] \right\} \\ & - \lambda_1 \cdot \mathbb{E}_{\tilde{\mathbf{x}} \sim p(\epsilon \text{ data} + (1-\epsilon)G(\mathbf{z}))} [\|\nabla_{\tilde{\mathbf{x}}} D(\tilde{\mathbf{x}})\|_2 - 1]^2; \\ \max_{G, E, F} \quad & \mathbb{E}_{\mathbf{x} \sim p(\text{data})} \left\{ \mathbb{E}_{\mathbf{z} \sim p(\mathbf{z})} [D(G(\mathbf{z}, \mathbf{x}_h))] \right. \\ & \left. - \lambda_2 \cdot g\left(\underbrace{F(E(\mathbf{x}_h))}_{\tilde{\mathbf{x}}_h}, \mathbf{x}_h\right) \right\}, \end{aligned} \tag{6}$$

where $g(\cdot)$ denotes the loss function; the authors apply the mean squared error as the loss function in Lu and Yi [17]. The parameter λ_2 controls how large the penalization by the autoencoder, and the term is thus known as the *autoencoding penalty (AP)*. In the original CGAN methodology, the conditioner is used to extract features that can help the generator to cheat the discriminator; however, it may lose some important information that captures the internal features of the market trend. The ACGAN then finds a balance between cheating the discriminator and keeping its market information.

2.4 Data Normalization

Following Mariani et al. [19]; Lu and Yi [17], given the frame window of $w = h + f$ days¹, we consider the *adjusted closing price* $\mathbf{p} \in \mathbb{R}^w$ series for each asset. Then we unit-normalize the price series \mathbf{p} for each asset to fill in the range $[-1, 1]$ for the initial h days. This normalization procedure can help us to expose the values of neural networks limited within a suitable range that removes price-variability over multiple assets within the specified window. In practice, the unit-normalization can be done by 3-sigma normalization: given the mean μ and standard deviation σ of $\mathbf{p}_{1:h} \in \mathbb{R}^h$, the normalization is done by

$$\tilde{\mathbf{p}} = (\mathbf{p} - \mu)/(3\sigma). \tag{7}$$

After generating the surrogate future series $\tilde{\mathbf{p}}_{h+1:w}$, we apply again a de-normalization procedure:

$$\hat{\mathbf{p}}_{h+1:w} = \tilde{\mathbf{p}}_{h+1:w} \times 3\sigma + \mu. \tag{8}$$

3 Hybrid Methods

3.1 CGAN with Eavesdropping

The proposed hybrid approaches highly rely on the data normalization procedure. Suppose in Eq. (7), we obtain the mean m of the whole asset $\mathbf{p} \in \mathbb{R}^w$ in the window w rather than the historical one $\mathbf{p}_{1:h} \in \mathbb{R}^h$, and we apply the data normalization by this value:

$$\tilde{\mathbf{p}} = (\mathbf{p} - m)/(3\sigma). \tag{9}$$

Since the $\mathbf{p}_{h+1:w}$ in \mathbf{p} cannot be obtained in practice and we thus call this method *eavesdropping*. A simple experiment on this approach, comparing the CGAN and CGAN with Eavesdropping, the return-SR (Sharpe ratio) plots in Fig. 4 show that this eavesdropping procedure can increase the performance to a large extent, enforcing the mean Sharpe ratio from about 1.0 to 2.5. And the random draws are more clustered with small deviations such that the end strategy is more stable. This is reasonable since the GAN finds the future means of the assets from this normalization and generates the series whose means are closer to these values.

¹ h for the historical length, f for the future length. The historical series is denoted by $\mathbf{p}_{1:h} \in \mathbb{R}^h$, and the real future series can be obtained by $\mathbf{p}_{h+1:w} \in \mathbb{R}^f$.

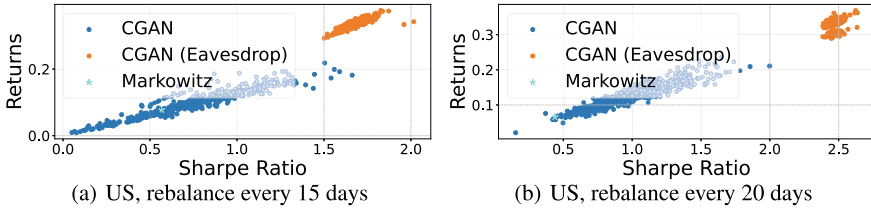


Fig. 4 Eavesdropping: (Annual) return-SR measured on the test period in US region by randomly sampling 1000 series for CGAN and **CGAN with Eavesdropping**. Similar results can be observed for ACGAN and ACGAN with Eavesdropping models

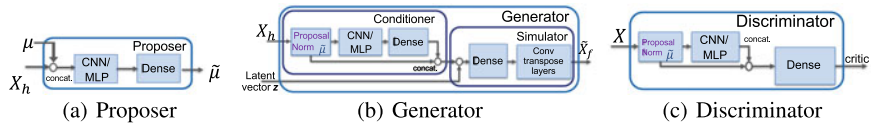


Fig. 5 Architectures of the HybridCGAN generative and discriminative models

3.2 Proposed Methodology

Though the eavesdropping methods are not practical since we use future data, the idea can be applied to the proposed hybrid methods. We incorporate an extra proposal network P to find the surrogate of the mean values for each asset in corresponding windows as shown in Fig. 5. The proposer P is trained by minimizing the mean squared error between the real mean value m of the whole period and the prediction $\tilde{\mu}$. Note that we incorporate the historical mean into the input of the proposal network to make the optimization easier:

$$\tilde{\mu} = P(X_h, \mu), \tag{10}$$

i.e., we concatenate X_h and μ to predict $\tilde{\mu}$. After finding the surrogate mean value $\tilde{\mu}$, the normalization follows:

$$\tilde{p} = (p - \tilde{\mu}) / (3\sigma). \tag{11}$$

And the de-normalization:

$$\hat{p}_{h+1:w} = \tilde{p}_{h+1:w} \times 3\sigma + \tilde{\mu}. \tag{12}$$

Since this approach combines the CGAN and deep neural network regression, we call it the *HybridCGAN* model. This hybrid approach can be easily extended into the ACGAN methodology in a similar way, termed the *HybridACGAN* model.

Table 1 Summary of the underlying portfolios in the US and EU markets, 10 assets for each market respectively. In each region, we include assets from various sectors to favor a somehow sector-neutral strategy

	Ticker	Type	Sector	Company	Curr.
US region	MSFT	Share	IT	Microsoft	USD
	GOOG	Share	IT	Alphabet	USD
	XOM	Share	Energy	Exxon Mobil	USD
	HES	Share	Energy	Hess	USD
	PFE	Share	Healthcare	Pfizer	USD
	WBA	Share	Consumer staples	Walgreens Alliance	USD
	KR	Share	Consumer staples	The Kroger	USD
	IYR	ETF	Real estate	iShares US Real Estate	USD
	IYY	ETF	Dow Jones	iShares Dow Jones	USD
	SHY	ETF	US treasury bond	iShares Treasury Bond	USD
EU region	VOW3.DE	Share	Automotive	Volkswagen	EUR
	BMW.DE	Share	Automotive	BMW	EUR
	VK.PA	Share	Industrials	Vallourec S.A.	EUR
	SOL.PA	Share	Industrials	Soitec S.A.	EUR
	DTE.DE	Share	Technology	Deutsche Telekom AG	EUR
	SAP.DE	Share	Technology	SAP SE	EUR
	BAS.DE	Share	Basic materials	BASF SE	EUR
	BAYN.DE	Share	Healthcare	Bayer AG	EUR
	^FCHI	Index	French market	CAC 40	EUR
	^GDAXI	Index	German market	DAX	EUR

4 Experiments

To evaluate the strategy and demonstrate the main advantages of the proposed Hybrid-CGAN and HybridACGAN methods, we conduct experiments with different analysis tasks; datasets from different geopolitical markets including the US and the European (EU) markets, and various industrial segments including Healthcare, Automotive, Energy and so on. We obtain publicly available data from Yahoo Finance². For the US market, we obtain data for a 17-year period, i.e., from 2005-05-24 to 2022-05-27, where the data between 2005-05-24 and 2019-03-28 is considered training data;

² <https://finance.yahoo.com/>.

while data between 2019-03-28 and 2022-05-27 is taken as the test set (800 trading days). For the EU market, we obtain data for a 16-year period, i.e., from 2006-07-18 to 2022-06-07, where the data between 2006-07-18 and 2019-04-09 is considered training data; while data between 2019-04-10 and 2022-06-07 is taken as the test set (800 trading days). The underlying portfolios are summarized in Table 1:

Algorithm 1 Training and testing process for the HybridCGAN, HybridACGAN, CGAN, and ACGAN models.

1: **General Input:** Choose parameters $w = h + f$; number of assets N ; number of epoches T ; latent dimension m ;

2: **Training Input:** Training data matrix $M \in \mathbb{R}^{N \times D}$;

3: Decide index set $S_1 = \{1, 2, \dots, D - w + 1\}$ and **draw without replacement**;

4: **for** $t = 1$ to T **do**

5: **for** $i \in \text{random}(S_1)$ **do**

6: $X = [X_h, X_f] = M[:, i : i + w - 1] \in \mathbb{R}^{N \times w}$;

7: Randomly sample latent vector $z \in \mathbb{R}^m$;

8: Backpropatation for generator in Eq. (6) or (5);

9: Generate surrogate $\hat{X}_f = G(z, X_h) \in \mathbb{R}^{N \times f}$;

10: Backpropatation for discriminator in Eq. (6) or (5);

11: **end for**

12: **end for**

13: **Inference Input:** Testing data matrix $A \in \mathbb{R}^{N \times K}$;

14: **Inference Output:** Testing data matrix $B \in \mathbb{R}^{N \times K}$;

15: Decide index set $S_2 = \{h + 1, h + f + 1, \dots\}$;

16: Copy the first h days data $B[:, 1 : h] = A[:, 1 : h]$;

17: **for** $i \in \text{ordered}(S_2)$ **do**

18: $X = [X_h, X_f] = A[:, i : i + w - 1] \in \mathbb{R}^{N \times w}$;

19: Randomly sample latent vector $z \in \mathbb{R}^m$;

20: Generate $B[:, i : i + f - 1] = G(z, X_h) \in \mathbb{R}^{N \times f}$ with de-normalization in Eq. (8);

21: **end for**

22: Output the synthetic series B ;

- *US market*: 10 assets of US companies from different industrial segments, i.e., GOOG and MSFT (from IT sector), PFE (from Healthcare sector), XOM and HES (from Energy sector), WBA and KR (from Consumer staples sector), and three ETFs (IYY, IYR, SHY).
- *EU market*: 10 portfolios of EU companies from different industrial segments, i.e., VOW3.DE and BMW.DE (from Automotive sector), VK.PA and SOL.PA (from Industrials sector), DTE.DE and SAP.DE (from Technology sector), BAS.DE (from Basic materials sector), BAYN.DE (from Healthcare sector), and two indices, ^FCHI and ^GDAXI, that track the German and French stock markets respectively.

The specific time periods and assets are chosen by following the four criteria. 1). *Data diversity*: in each market, we include companies from different sectors so that the final strategies are somewhat sector-neutral with fewer risks; 2). *Data availability*: we cover as a longer period as possible to make a decent prediction; the periods

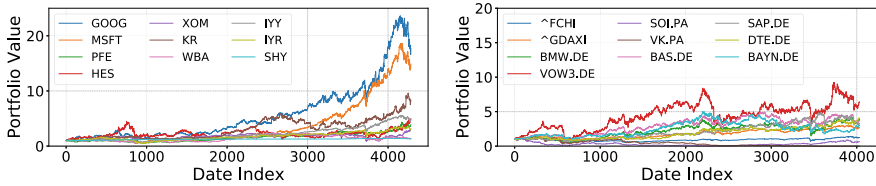


Fig. 6 Different portfolios for the US (left) and EU (right) markets with a unit initial value

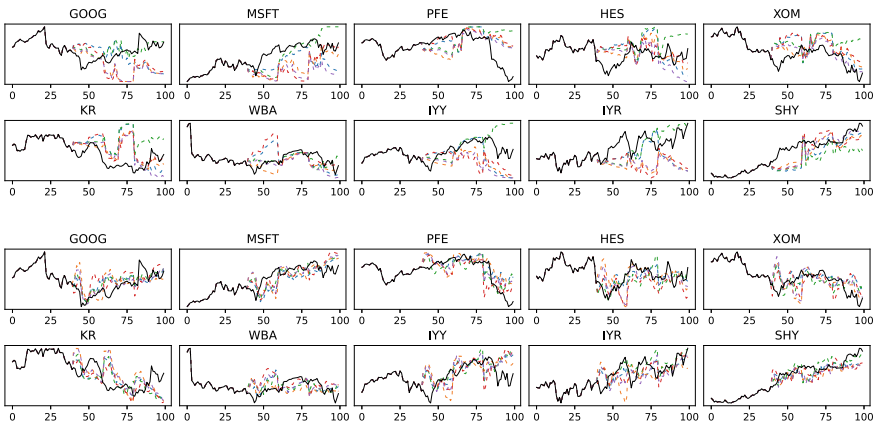


Fig. 7 Actual price trend (black solid line) of the US assets, five representative simulations (colored dashed lines) generated by HybridCGAN (upper two rows), and five representative simulations generated by CGAN (lower two rows) for the first 100 trading days in the test set

are selected to make all the assets have same frame length; 3). *Data correctness*: given the Yahoo Finance data source, we only include the data that do not have NaN values; 4). *Currency homogeneity*: in each region, the traded currencies are the same. Figure 6 shows the series of different assets where we initialize each portfolio with a unitary value for clarity.

In all experiments, the same parameter initialization is adopted when conducting different tasks. We compare the results in terms of performance of portfolio allocation and diversification of the assets. In a wide range of scenarios across various tasks, HybridCGAN and HybridACGAN improve portfolio evaluations, and lead to return-risks performances that are as good or better than the existing Markowitz framework, CGAN, and ACGAN methodologies.

Network structures for the conditioner (in HybridCGAN and CGAN), encoder, decoder (in HybridACGAN and ACGAN), generator, discriminator (in HybridCGAN, HybridACGAN, CGAN, and ACGAN), and proposer (in HybridCGAN and HybridACGAN) are provided in Appendix A. In all experiments, we train the network with 1,000 epochs. For simplicity, we set the risk-free interest $r_f = 0$ to assess the Sharpe ratio evaluations.

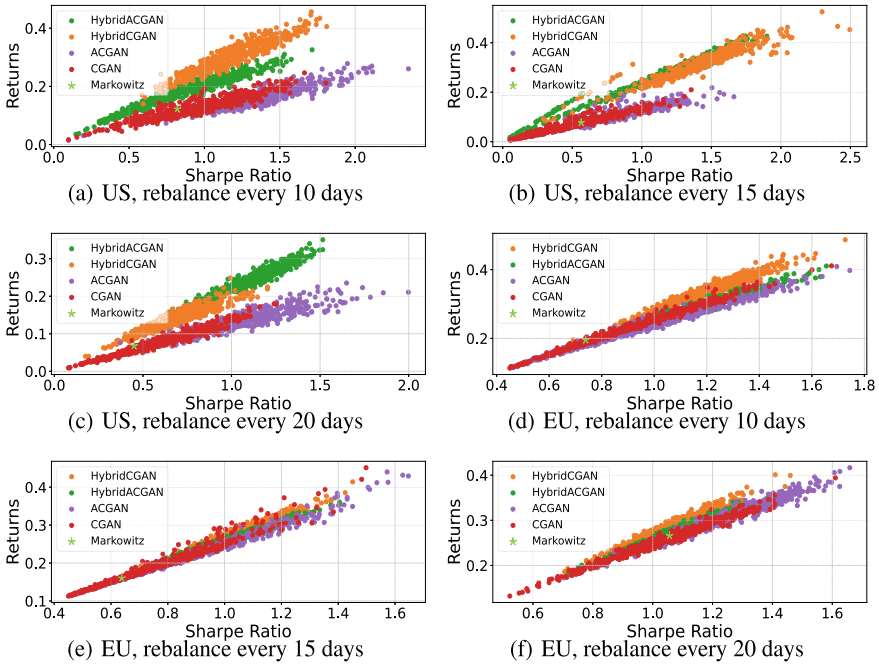


Fig. 8 (Annual) return-SR measured on the test period by randomly sampling 1000 series

4.1 Generating Analysis

We follow the training and testing procedures in Algorithm 1. Given the training matrix M of size $N \times D$ (where N is the number of assets and D is the number of days in the daily analysis context) and the window size w ($w = h + f$ where h is the length of the historical window and f is the length of the future window), we define the index set $S_1 = \{1, 2, \dots, D - w + 1\}$ so that $D - w$ samples can be extracted for each training epoch. While at the testing stage, given the testing matrix $A \in \mathbb{R}^{N \times K}$, the index set is obtained by $S_2 = \{h + 1, h + f + 1, \dots\}$ so that $(K - h)/f$ samples can be obtained (supposed here $(K - h)$ can be divided by f). The output $B \in \mathbb{R}^{N \times K}$ of Algorithm 1 is the financial market simulation of the N assets in K days (here $N = 10$ and $K = 800$ in our datasets for both US and EU regions). To be more concrete, the first h days of B are just copies of A , while the next f days are the synthetic series based on the data of the first h days; the next f days are the synthetic series based on the data between the f -th and $(f + h)$ -th days; and so on.

We set window size $h = 40$, $f = 20$ and $w = 60$ in all experiments. Figure 7 shows the actual price trend (black solid line) of the US assets for the first 100 trading days in the test set, and five representative simulations generated by HybridCGAN and CGAN models (colored dashed lines). The proposed hybrid methods are not

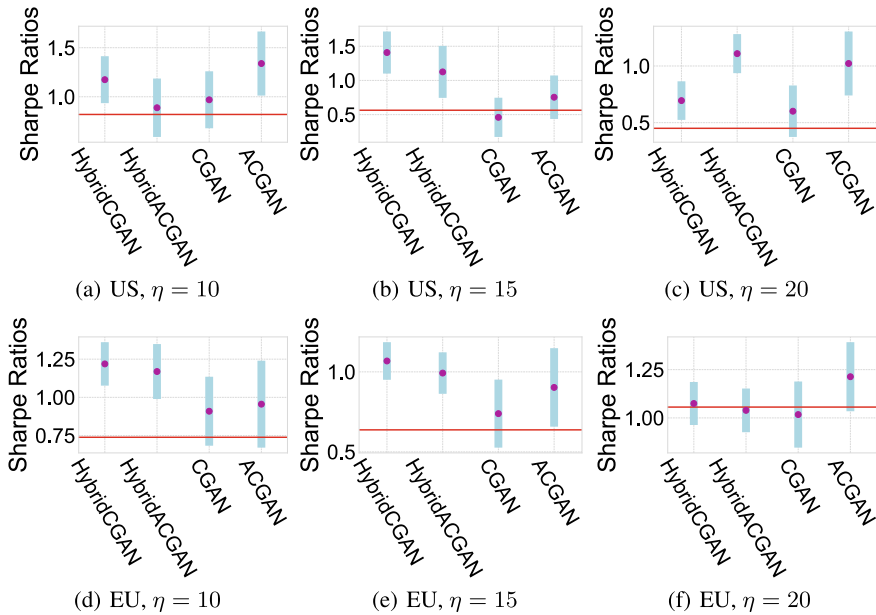


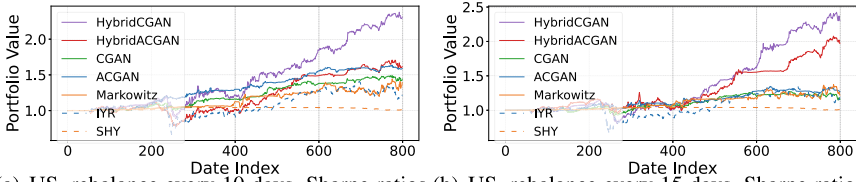
Fig. 9 The hybrid approaches surpass the non-hybrid alternatives reference approach on financial performance where the shaded bars are the standard deviation over means. Red horizontal lines are the relative Markowitz results

seeking simulations that are closer to the real series, but find the typical trends of the series, e.g., there is a big drawdown for GOOG, MSFT, PFE, HES, XOM, WBA, and IYY around 80-th day; and an increase for SHY around 80-th day.

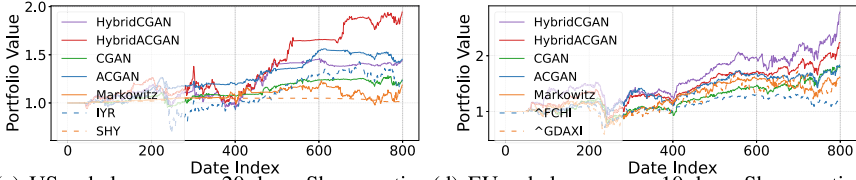
4.2 Portfolio Analysis

After generating the synthetic series for each asset, we optimize over the fake series to generate minimal Sharpe ratio weight allocations (for HybridCGAN, HybridACGAN, CGAN, and ACGAN). For Markowitz framework, the optimization is done over the past data (here we use h days). We consider three rebalance settings: a *defensive setting* with rebalancing every $\eta = 10$ days; a *balanced setting* with $\eta = 15$; and an *aggressive setting* with $\eta = 20$. Figure 8 presents the distribution of return-SR (Sharpe ratio) scatters with 1,000 draws from HybridCGAN, HybridACGAN, CGAN, and ACGAN models, and the one from Markowitz framework. The points in the upper-right corner are the better ones.

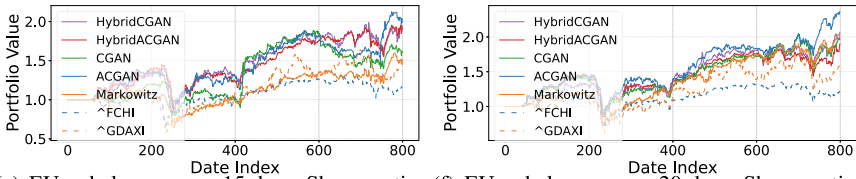
Figure 9 shows the distribution of Sharpe ratios over 1,000 draws where the shaded bars are the standard deviation over means. We observe that all the HybridCGAN models surpass the alternative CGAN approaches; and in most cases, the HybridACGAN models perform better than the ACGAN model except when $\eta = 10$ in the US



(a) US, rebalance every 10 days, Sharpe ratios of HybridCGAN, HybridACGAN, CGAN, ACGAN, and Markowitz are 1.32, 0.95, 1.22, **1.72**, and 0.82 respectively. (b) US, rebalance every 15 days. Sharpe ratios of HybridCGAN, HybridACGAN, CGAN, ACGAN, and Markowitz are **1.47**, 1.18, 0.49, 0.82, and 0.56 respectively.



(c) US, rebalance every 20 days. Sharpe ratios of HybridCGAN, HybridACGAN, CGAN, ACGAN, and Markowitz are 0.74, **1.20**, 0.71, 1.17, and 0.45 respectively. (d) EU, rebalance every 10 days, Sharpe ratios of HybridCGAN, HybridACGAN, CGAN, ACGAN, and Markowitz are **1.33**, 1.22, 0.95, 1.02, and 0.74 respectively.



(e) EU, rebalance every 15 days. Sharpe ratios of HybridCGAN, HybridACGAN, CGAN, ACGAN, and Markowitz are 0.96, **1.01**, 0.76, 0.96, and 0.64 respectively. (f) EU, rebalance every 20 days. Sharpe ratios of HybridCGAN, HybridACGAN, CGAN, ACGAN, and Markowitz are 1.09, 1.06, 1.05, **1.27**, and 1.06 respectively.

Fig. 10 Portfolio values for different diversification risk settings. Reference benchmarks are shown with dashed lines (Index or ETF assets). HybridCGA, HybridACGAN, CGAN, ACGAN, and Markowitz with solid lines

region and $\eta = 20$ in the EU region. The hybrid approach has a smaller variance such that the final strategies are more stable.

The end strategies from HybridCGAN, HybridACGAN, CGAN, and ACGAN are the ones by taking average weight from these 1,000 draws on each rebalancing date (we call it *mean strategy*). Figure 10 shows the portfolio value series of mean strategies for HybridCGAN, HybridACGAN, CGAN, ACGAN, and the one from Markowitz framework along the test period where we initialize each portfolio with a unitary value. The hybrid approaches dominate the other approaches in terms of the final portfolio values and Sharpe ratios.

When we apply the mean strategy in the US region with $\eta = 15, 20$, the hybrid versions of CGAN and ACGAN achieve both better return and Sharpe ratio evaluations compared to the mean strategies of non-hybrid approaches. Similar results are

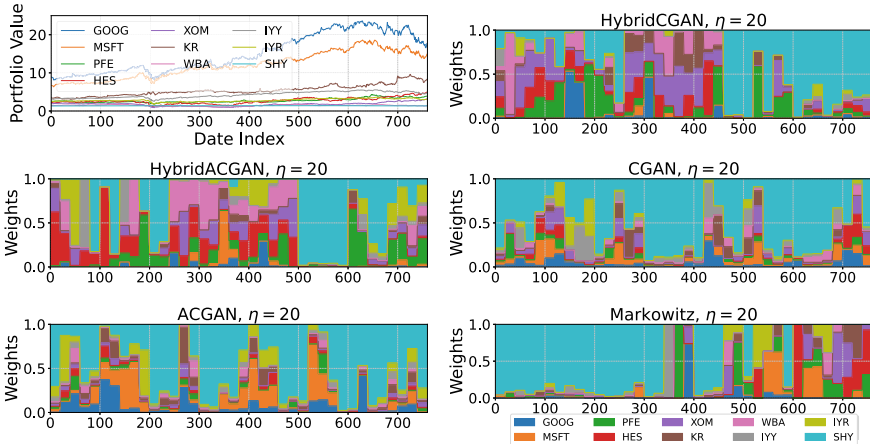


Fig. 11 Portfolio values of 10 assets, and weights distribution over time on the test period for HybridCGAN, HybridACGAN, CGAN, ACGAN, and Markowitz models for the US region with rebalancing every 20 days. Portfolios from the HybridCGAN and HybridACGAN are more diversified than those from CGAN, ACGAN, and Markowitz

observed in the EU region with $\eta = 10, 15$. The Sharpe ratio of ACGAN model for US region with $\eta = 10$ obtains best performance among other results (SR = 1.72); however, its final portfolio value is not the best where the HybridCGAN obtains the best final portfolio value with a decent Sharpe ratio of 1.32.

4.3 Weight Distribution

In Fig. 11, we present the distributions of weights over time on the test period for the US region with rebalancing every 20 days and the corresponding portfolio values of the 10 assets. Since we use the first 40 trading days as the historical series, weights of only 760 days are shown in the figure. We observe that the Markowitz model puts a large weight on the SHY asset for the first 300 trading days; while in this period, companies in IT sector (GOOG and MSFT) receive large positive returns making the Markowitz result less competitive.

Moreover, we find that there is a big drawdown for the IT companies since the 600-th day and 700-th day. The HybridCGAN and HybridACGAN perform well compared to their non-hybrid versions in that they put less weight on these companies during the drawdown periods. Further, one can easily observe that the portfolios from the HybridCGAN and HybridACGAN are more diversified than those from CGAN, ACGAN, and Markowitz. Hybrid methods are able to systematically improve the returns achievable.

For other rebalancing days and weights distribution for the EU region, results are provided in Figs. 12, 13, 14, 15, and 16; and we shall not repeat the details.

5 Conclusion

The paper aims to solve the issue of poor prediction ability in the CGAN and ACGAN methodology for portfolio analysis. We propose a simple and computationally efficient algorithm that incorporates a deep neural regression model and requires little extra computation. A potential future work on the HybridCGAN and HybridACGAN models is to further reduce the variance of different draws such that the end strategy will be more stable and consistent.

A Network Structures

We provide detailed structures for the neural network architectures we used in our experiments in this section. Given the number of assets N , historical length h , future length f ($w = h + f$), and latent dimension m for the prior distribution vector \mathbf{z} , we consider multi-layer perceptron (MLP) structures, the detailed architecture for each fully connected layer is described by $F(\langle \text{num inputs} \rangle : \langle \text{num outputs} \rangle : \langle \text{activation function} \rangle)$; for an activation function of LeakyRelu with parameter p is described by $LR(\langle p \rangle)$; and for a dropout layer is described by $DP(\langle \text{rate} \rangle)$. The *conditioner* in CGAN shares the same structure as the *encoder* in the ACGAN model (see Figs. 2 and 3). Then the network structures we use can be described as follows:

$$\begin{aligned} \mathbf{Conditioner} = \mathbf{Encoder} &= F(N \cdot h : 512 : LR(0.2)) \cdot F(512 : 512 : LR(0.2)) \cdot \\ &DP(0.4) \cdot F(512 : 16) \end{aligned} \quad (13)$$

$$\mathbf{Decoder} = F(16 : 512 : LR(0.2)) \cdot F(512 : 512 : LR(0.2)) \cdot DP(0.4) \cdot F(512 : N \cdot h) \quad (14)$$

$$\begin{aligned} \mathbf{Simulator (in CGAN or ACGAN)} &= F(m + 16 : 128 : LR(0.2)) \cdot F(128 : 256 : LR(0.2)) \cdot \\ &F(256 : 512 : LR(0.2)) \cdot F(512 : 1024 : LR(0.2)) \cdot F(1024 : N \cdot f : \text{TanH}) \end{aligned} \quad (15)$$

$$\begin{aligned} \mathbf{Simulator (in HybridCGAN or HybridACGAN)} &= F(m + 16 : 128 : LR(0.2)) \cdot \\ &F(128 : 256 : LR(0.2)) \cdot F(256 : 512 : LR(0.2)) \cdot F(512 : 1024 : LR(0.2)) \cdot \\ &F(1024 : N \cdot f : \text{TanH}) \cdot 100 \end{aligned} \quad (16)$$

$$\begin{aligned} \mathbf{Discriminator} &= F(N \cdot (h + f) : 512 : LR(0.2)) \cdot F(512 : 512 : LR(0.2)) \cdot \\ &DP(0.4) \cdot F(256 : 512 : LR(0.2)) \cdot F(512 : 1) \end{aligned} \quad (17)$$

$$\begin{aligned} \mathbf{Proposer} &= F(N \cdot (h + 1) : 512 : LR(0.2)) \cdot F(512 : 512 : LR(0.2)) \cdot \\ &DP(0.4) \cdot F(512 : 16), \end{aligned} \quad (18)$$

where the highlighted 1 in proposer network is the input of the historical mean values (Eq. (10)). Due to effect of the proposal network, the outputs of the generator (or simulator) are not constrained into the range of $[-1, 1]$, we also multiply the result by 100 in the simulator of HybridCGAN or HybridACGAN. We trained networks using Adam's optimizer with learning rate 2×10^{-5} , $\beta_1 = 0.5$, and $\beta_2 = 0.999$. We set the penalization parameters $\lambda_1 = 10$ and $\lambda_2 = 3$. The latent dimension is $m = 100$. And we trained models for 1,000 epochs.

B Weight Distributions Under Different Settings

As shown in the main paper, Fig. 11 presents the weight distribution over time on the test period for the US region with $\eta = 20$. Figures 12 and 13 show the weight distributions for the US region with $\eta = 10, 15$ respectively. Again, we observe that the HybridCGAN and HybridACGAN still have a more diverse portfolio allocation. Figures 14, 15, and 16 then present the weight distributions for the EU region with $\eta = 10, 15, 20$ respectively.

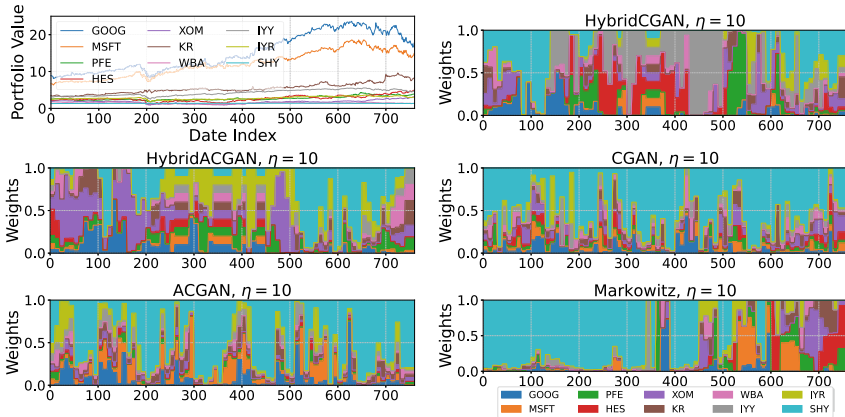


Fig. 12 Weights distribution over time on the test period for HybridCGAN, HybridACGAN, CGAN, ACGAN, and Markowitz models for the US region with rebalancing every 10 days

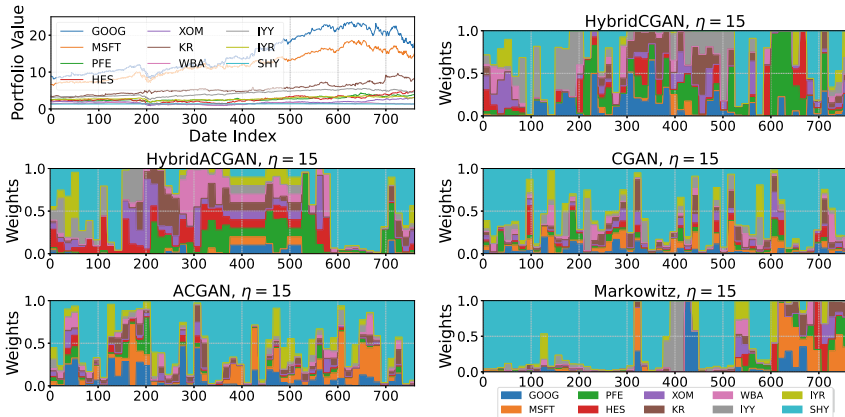


Fig. 13 Weights distribution over time on the test period for HybridCGAN, HybridACGAN, CGAN, ACGAN, and Markowitz models for the US region with rebalancing every 15 days

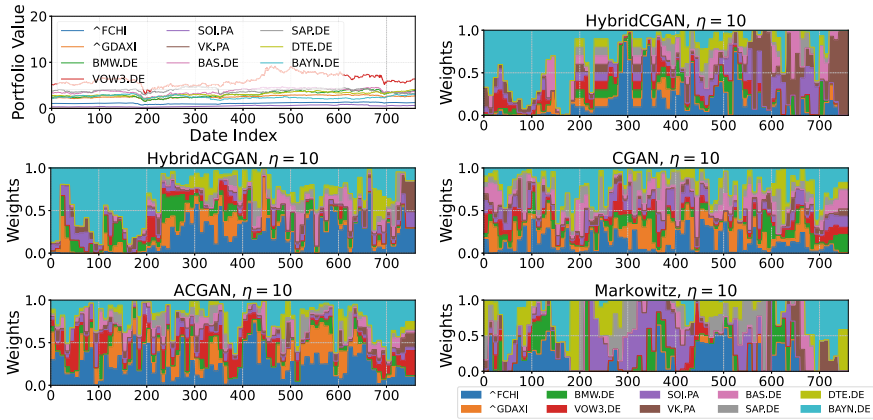


Fig. 14 Weights distribution over time on the test period for HybridCGAN, HybridACGAN, CGAN, ACGAN, and Markowitz models for the EU region with rebalancing every 10 days

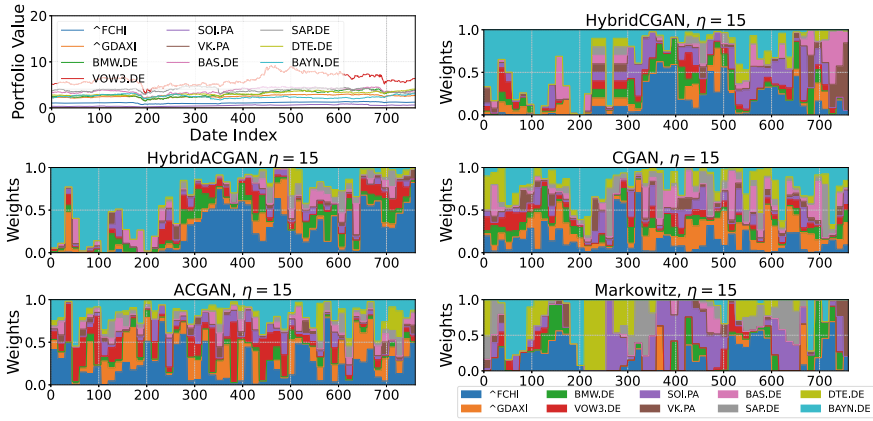


Fig. 15 Weights distribution over time on the test period for HybridCGAN, HybridACGAN, CGAN, ACGAN, and Markowitz models for the EU region with rebalancing every 15 days

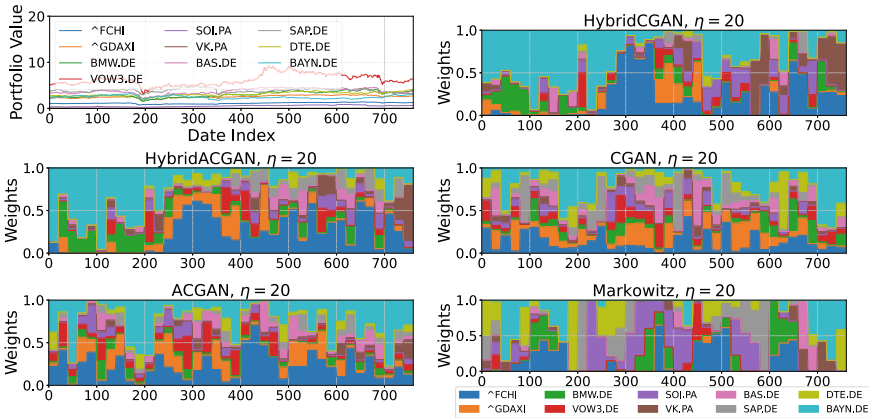


Fig. 16 Weights distribution over time on the test period for HybridCGAN, HybridACGAN, CGAN, ACGAN, and Markowitz models for the EU region with rebalancing every 20 days

References

1. Best MJ, Grauer RR (1991) On the sensitivity of mean-variance-efficient portfolios to changes in asset means: some analytical and computational results. *Rev Financial Stud* 4(2):315–342
2. Bollerslev T (1986) Generalized autoregressive conditional heteroskedasticity. *J Econom* 31(3):307–327
3. Eckerli F, Osterrieder J (2021) Generative adversarial networks in finance: an overview. arXiv preprint [arXiv:2106.06364](https://arxiv.org/abs/2106.06364)
4. Engle RF (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica J Econom Soc* 987–1007
5. Esteban C, Hyland SL, Rättsch G (2017) Real-valued (medical) time series generation with recurrent conditional GANs. arXiv preprint [arXiv:1706.02633](https://arxiv.org/abs/1706.02633)
6. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. *Adv Neural Inf Process Syst* 27
7. Green RC, Hollifield B (1992) When will mean-variance efficient portfolios be well diversified? *J Finance* 47(5):1785–1809
8. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville AC (2017) Improved training of Wasserstein GANs. *Adv Neural Inf Process Syst* 30
9. Huang W, Nakamori Y, Wang S-Y (2005) Forecasting stock market movement direction with support vector machine. *Comput Oper Res* 32(10):2513–2522
10. Kallberg JG, Ziemba WT (1981) Remarks on optimal portfolio selection. *Methods Oper Res* 44:507–520
11. Kallberg JG, Ziemba WT (1984) Mis-specifications in portfolio selection problems. In: *Risk and capital*. Springer, pp 74–87
12. Kara Y, Boyacioglu MA, Baykan ÖK (2011) Predicting direction of stock price index movement using artificial neural networks and support vector machines: the sample of the istanbul stock exchange. *Expert Syst Appl* 38(5):5311–5319
13. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. *Nature* 401(6755):788–791
14. Lu J (2021) Numerical matrix decomposition and its modern applications: a rigorous first course. arXiv preprint [arXiv:2107.02579](https://arxiv.org/abs/2107.02579)
15. Lu J (2022) Matrix decomposition and applications. arXiv preprint [arXiv:2201.00145](https://arxiv.org/abs/2201.00145)

16. Lu J, Ye X (2022) Flexible and hierarchical prior for Bayesian nonnegative matrix factorization. arXiv preprint [arXiv:2205.11025](https://arxiv.org/abs/2205.11025)
17. Lu J, Yi S (2022a) Autoencoding conditional GAN for portfolio allocation diversification. arXiv preprint arXiv
18. Lu J, Yi S (2022) Reducing overestimating and underestimating volatility via the augmented blending-ARCH model. *Appl Econ Finance* 9(2):48–59
19. Mariani G, Zhu Y, Li J, Scheidegger F, Istrate R, Bekas C, Malossi ACI (2019) PAGAN: Portfolio analysis with generative adversarial networks. arXiv preprint [arXiv:1909.10578](https://arxiv.org/abs/1909.10578)
20. Markowitz HM (1968) Portfolio selection. In: *Portfolio selection*. Yale University Press,
21. Markowitz HM (1976) Markowitz revisited. *Fin Anal J* 32(5):47–52
22. Sharpe WF (1966) Mutual fund performance. *J Bus* 39(1):119–138
23. Takahashi S, Chen Y, Tanaka-Ishii K (2019) Modeling financial time-series with generative adversarial networks. *Physica A Stat Mech Appl* 527:121261
24. Timmermann A, Granger CWJ (2004) Efficient market hypothesis and forecasting. *Int J Forecast* 20(1):15–27
25. Tsay RS (2005) *Analysis of financial time series*. Wiley, Hoboken

An Innovative Solar Power Can Satellite Model Prototype to Perceive the Environmental Data



Md. Nahidul Alam, Md. Zahid Hasan Buiyan, Md. Abdullah Al Hasan Anik, Atikur Rahman, and Nazifa Tahsin Lamisa

Abstract This paper presents an overview of the workings of CANSAT, a portable satellite capable of observing the environment. Nowadays, it is of great importance to learn about the weather conditions beforehand. Precise weather pattern data acquired in a short span of time can save many lives and property damage. The aim of this paper is to present a miniature model satellite that can measure temperature, humidity, pressure, gas detection, etc. in a given environment. This paper focuses on the results and analysis of weather data in Bangladesh. Our prototype model is able to communicate with the base station at a 300 m distance and transmit all the necessary survey data to the base station wirelessly. This proposed model also took careful consideration of the structural design of the device and data transmission speed to and from the base station for its augmentation. Different survey data, like acceleration, rotation, latitude, longitude, and altitude, were successfully collected and analyzed.

Keywords Acceleration · CANSAT · Temperature · Gas · Rotation · Altitude

1 Introduction

CANSAT is actually a device where all the circuits are inserted into a can-shaped object. The dimensions of a regular CANSAT will be 115 mm in height and 66 mm in diameter [1]. The CANSAT concept was first introduced by Stanford University professor Bob Twiggs in 1998. Bob Twiggs proposed a nano-satellite, which is a satellite in the shape of a soda can. The proposed model was started in 1999 under the name of ARLISS [2]. The concept was finally realized in 2010 by the European Space Agency (ESA) [3]. CanSat may be a standard frame calculation of a nano-satellite with a volume of around 350 ml and a mass of almost 500 g, that has the shape of a cylindrical rack structure [4–6].

Md. N. Alam (✉) · Md. Z. H. Buiyan · Md. A. Al Hasan Anik · A. Rahman · N. T. Lamisa
Bangladesh Army International University of Science and Technology, Cumilla Cantonment,
Cumilla, Bangladesh
e-mail: nahidul@baiust.edu.bd

It is possible for a CanSat to survey various data like altitude and longitude, capture images, and also provide various sensor data to the base station [7, 8]. The main track that a CanSat must have is.

- Light weight
- Affordability
- Design of a fruitful can and parachute
- Strong communication to the base station
- Battery backup management
- Sharp image capturing

Most of the researchers discussed in their papers how well it is possible to adjust all the related circuits into the compact size of a can. The size of the can is too compact, so the battery backup system is very poor. It is very difficult to insert a high-rate battery into a CanSat because of its size and weight. In our proposed system, we have kept all this under consideration. To backup our battery, we have implemented a solar cell on top of our CanSat. Because the size of our can is too compact, we have attached the parachute outside of our can so that it is possible to freely move whenever we free our can from the top.

2 Literature Review

In [9], Celebi et al. 2011, discuss a basic CANSAT concept which consists of a micro-controller, pressure and temperature sensors, a structure, and a parachute. In advanced level CANSAT, the mission is to launch and land the CANSAT at a predefined target point. CANSAT is able to send telemetry data to the ground station. This data consists of GPS data, acceleration, velocities, angular velocities, temperature, and pressure. The key features of a CANSAT design may be expressed as lightweight, affordable, and high technology [2] in Aydemir et al. 2013. In [10], Umit et al. discussed the launcher which is used to send the CANSAT to a certain altitude. Generally, a rocket is used for competitions or main missions. It is a prototype of real space satellite [11] Bautista-Linares et al. 2015. The internal process of CANSAT is divided into four different stages as shown in Fig. 1. In the first stages, sensors collect all of the environmental conditions. A second stage is the processing unit, it processes data. In the third stage, transmitters catch this data and send it to a receiver, which is the last stage [12] Hasan Raian et al. The CANSAT structure is made using 3D printing technology. The construction of CANSAT describes the design and manufacturing process. The newly developed CANSAT was composed of electronic and aeronautic parts. The descent control system is a subpart of aeronautic design, which consists of two parts, the parachute and the glider system [13] Kizilkaya et al. 2017.

In [14], researchers describe the folding wing glider, energy harvesting system, and real time environmental data collection and transmission, but fail to discuss the energy backup system. In [15], the author discussed the air pollution of the environment. The air is polluted due to the high amount of CO and CO₂ released

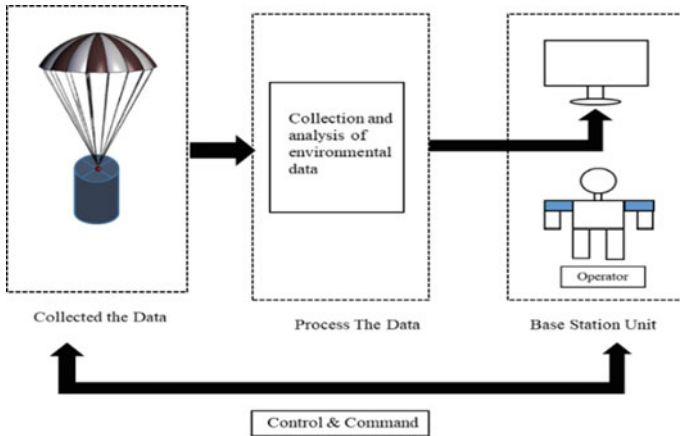


Fig. 1 A block diagram of the total CanSat system

from vehicles and other industries. The drawback of their research is that they did not discuss the communication range as well as the battery backup system of their system. In [16], the author discussed a cansat prototype model, but the main drawback is that they did not analyze enough graphical representation. As the main power supply is solar, the voltage distribution of the entire cansat device is followed via [17]. Researchers discussed weather and environmental issues in [18]. The proposal from them is to use the STEM to STEAM (Science, Technology, Engineering, Arts, and Mathematics) method in cansat. In [19], researcher focused more on the power supply of a cansat rather than the body design and sustainability of the device. In [20], authors discussed the rising issue of UV rays. The main drawback of their paper is they did not provide any data transmission and receive method clearly.

3 Methodology

The entire process is subdivided into three major parts.

- Collecting the data.
- Process the data.
- Send the data to base station.

The object is first raised to a height of 300 m. Then we initiate a free fall from 300 m and the parachute is deployed on the ground. The main task is that while landing on the ground, it has to take the necessary data before landing on the ground and send it to the base station. All the data that is surveyed by cansat is going to be showcased on the base station monitor via the nRF24I01 module. To use different sensors, first have to collect all the environmental data. That data is then transmitted to the base station monitor via the nRF24L01 module.

3.1 Data Transmission and Base Station Algorithm

At very first, the solar power gets power from the sun or any other light source, then it converts the 6 to 8 v using the boost converter module. If solar power is activated, then the battery will start storing charge through the TP4056 battery charger, and if not, then it will again search for solar power to get power from light sources. After getting power from the battery, it's time to provide power to the microcontroller, but before that, a buck module converts the voltage from 8 to 5 v so as not to burn the sensors, as most of the sensor's operating voltage is 3.3 v. After converting the voltage, the arduino get's power as well as the sensors. All the sensors are connected to the arduino, so all the sensors are activated and start capturing data from the environment as in Fig. 2. After getting data, it is now time to provide that data to the base station. A transmitter will transmit all the data to the base station.

As data is transmitted, a reception is designed at the base station unit as in Fig. 3 where the operator can easily monitor all the transmitted data through a computer. If any of the data being sent is interrupted, the operator can easily notice and fix the problem.

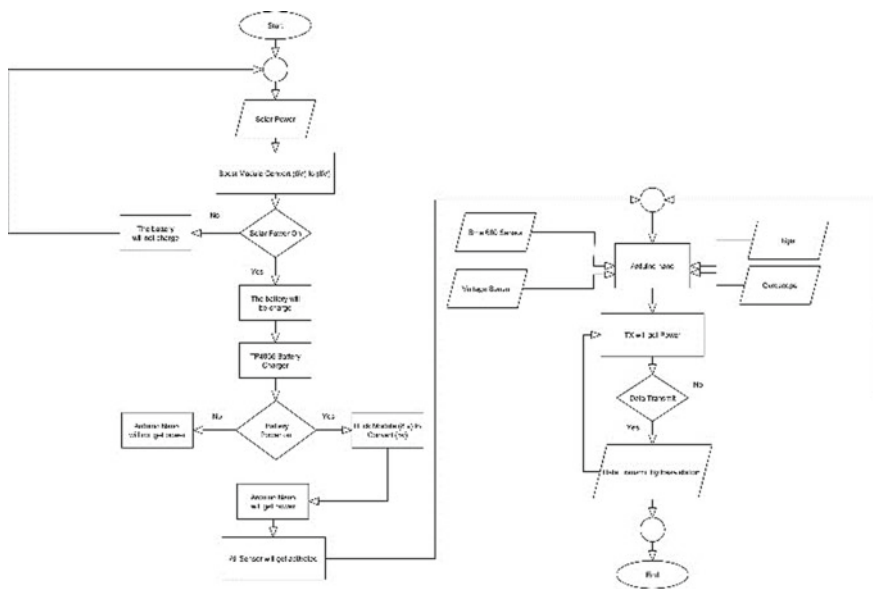


Fig. 2 Data transmission and sensor activation system block diagram

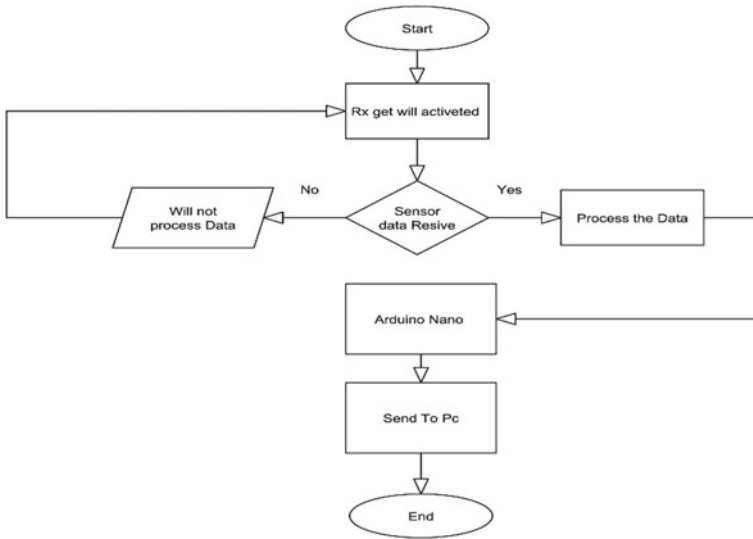


Fig. 3 A block diagram of the base station unit

3.2 Parachute Design Algorithm

As the parachute will take all the payload of CanSat, so by keeping it in mind, a parachute has been designed for the safe landing of our device as in Fig. 4. Most of the circuits are installed on a can-shaped tool, so the parachute is attached on the outside. To launch the CANSAT from the top, we developed a parachute by altitude for drop testing. Keeping a variety of factors in mind, such as size and CANSAT weight, aerodynamic factor, material parachutes and so on, the determined area is calculated by adding the following equations.

$$F_g = F_d \tag{1}$$

$$A = mg / (0.5\omega V^2 \rho) \tag{2}$$

Here,

F_g = Force of Gravity.

F_d = Drag Force.

A = Area of a parachute.

V = Descent velocity.

m = Mass of the CANSAT.

ω = Drag Coefficient of parachute.

ρ = Local density of air, assumed to be Constant = 1.225 Kg m^{-3} .



Fig. 4 Implementation of parachute

g = The acceleration of gravity = 9.8 gm^{-3} .

Area of a parachute,

$$A = mg/0.5 * c \omega \rho v^2$$

$$A = (350 * 9.81)/(0.5 * 0.8 * 1.225 * (4.5)^2).$$

$$A = 346 \text{ cm}.$$

The parachute is designed with the help of the above equation at descent rate of 4.5 ms^{-1} .

4 Useable Components and Circuit Diagram

4.1 Useable Components

XL6009 Step Up (Boost) Module. In our prototype model, a DC-DC boost module is used to increase the voltage that we are getting from the solar panel, and we are going to feed it to the battery charger. XL6009 boosts the voltage from 6 to 8 v.

Six Volt Mini Solar. In our proposed method, we are using this solar cell as the supply of the entire system. This solar cell produces a maximum of 5.8 v in output. We are placing this cell on top of our can device.

Battery Charger. A Lithium Polymer (Li Po) battery charger is used to charge our battery.

DC-DC Buck Converter LM2596. The buck converter is used to reduce the input voltage to an acceptable range to supply the microcontroller and all other sensors.

GY-521 Accelerometer and Gyroscope. It is used to get the altitude, longitude, and latitude values. The sensor works both as an accelerometer and a gyroscope.

GPS. It is used to determine the exact position of our CanSat. The cansat device is tested in various places, and it is tested at 300 m from the ground. So, it is necessary to get the exact position of the device.

BME 680. The BME680 is an environmental sensor that can detect gas, pressure, humidity, and temperature sensors. An environmental sensor connected with an Arduino to track air pollution and air quality, measure temperature, pressure, humidity, barometric pressure, and VOC (Volatile Organic Compounds).

Transmitter and Receiver (nRF24L01). To transmit the data, we have used the nRF24L01 module. The range of its communication bandwidth is 800 m.

4.2 Schematic Diagram

The diagram of the entire CanSat system is given in Fig. 5 where all the sensors as well as the transmitter are also attached to a single microcontroller. We have designed a PCB diagram for our system in Eagle software. We have used a boost module to boost the voltage that we are getting from the solar panel, as we have to feed the boosted voltage to the battery charger.

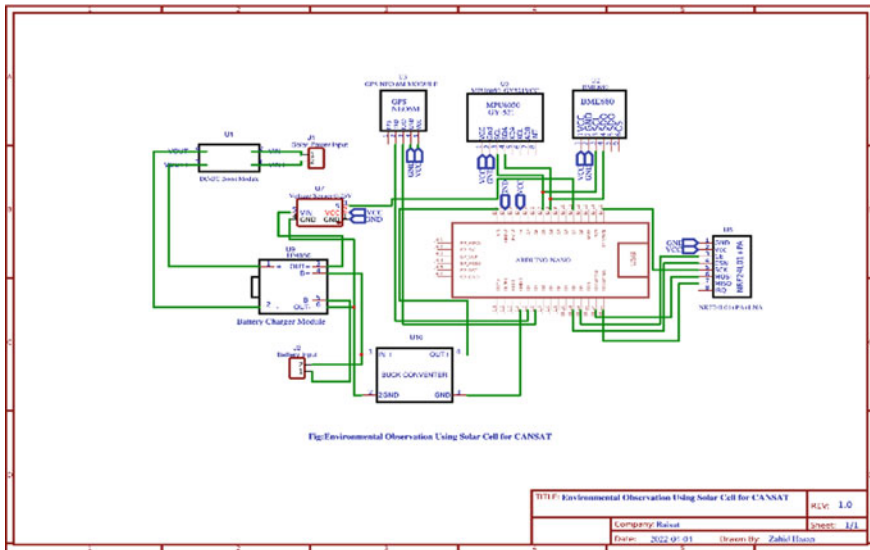


Fig. 5 Schematic circuit diagram of the transmitter

The PCB design of the transmitter circuit is designed in Eagle software by keeping in mind that this board has to be inserted into our can. Figure 6 represents the PCB design of the transmitter circuit.

Figure 7 is the circuit diagram of the receiver portion. The nRF modules are connected to an Arduino Nano that works as a receiver. This receiver part will be kept at the base station where the operator can easily access all the data.

In Fig. 8 representing the PCB of the receiver portion. This part will be kept in the base station as the operator has to handle different data.

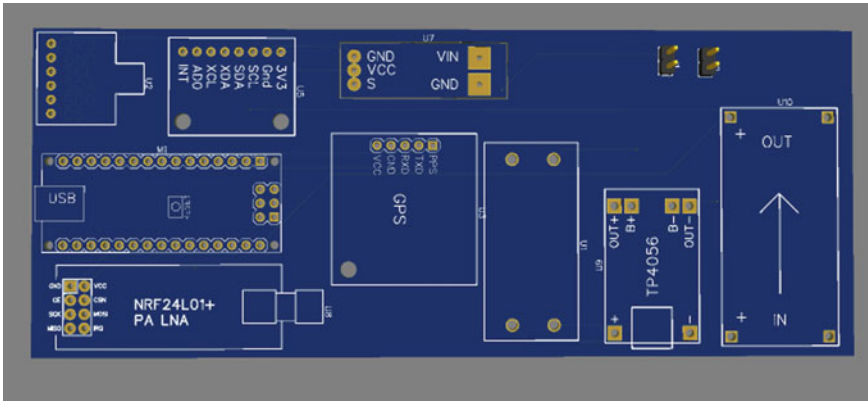


Fig. 6 PCB design for transmitter

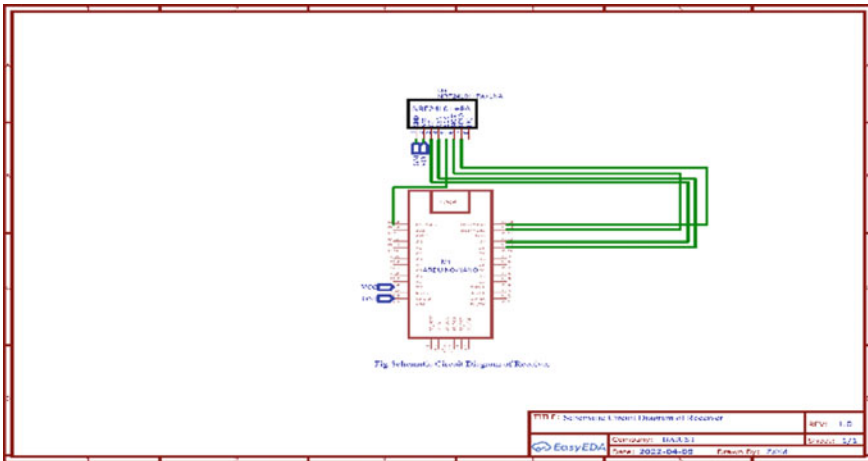


Fig. 7 Schematic circuit diagram for receiver

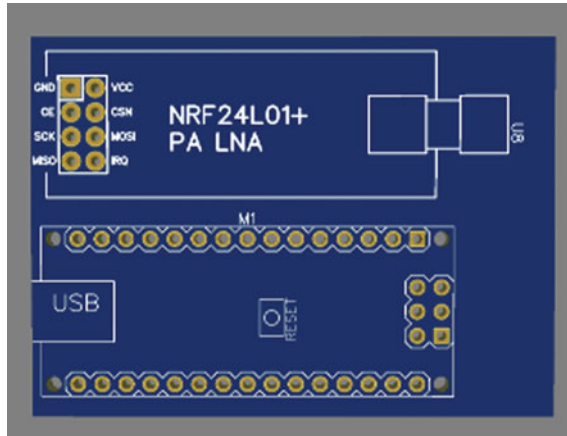


Fig. 8 PCB design for receiver

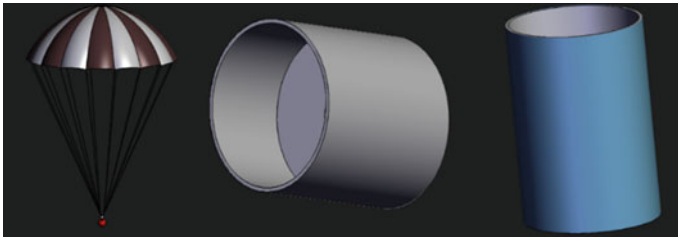


Fig. 9 3D design of CanSat and parachute

5 3D Design and Practical Experiment

5.1 3D Design

A 3D model is illustrated both for the parachute and CanSat in solidworks as in Fig. 9. To avoid complexity, we must keep the payload and light weight in mind when designing the can and parachute, we kept our 3D design very simple.

5.2 Practical Experiment

All the sensors and transmitter are inserted into a single PCB board so that the length does not get too high. Only solar power will be attached to the device's top, and a battery will be inserted into the can. We open a small part of our can to make sure the transmitter as in Fig. 10 and GPS can talk to each other GPS.

In Fig. 11 the total set of parachute and CanSat is shown. For free movement, we have attached the parachute outside of the can.

For the base station section, the receiver module is connected to a computer where all the data is accessible on a serial monitor display. All the data is continuously updated as at different altitudes we are getting different data. In Fig. 12 a base station receiver part is shown.



Fig. 10 Implemented circuit diagram of the transmitter



Fig. 11 Implementation of CanSat prototype

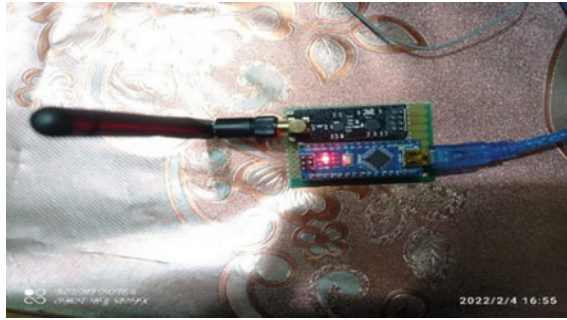


Fig. 12 Implemented circuit diagram of Receiver

6 Result and Cost Analysis

6.1 Result

As our proposed method can detect temperature, humidity, pressure, gas, acceleration, altitude, longitude, and latitude, all this information is sent to the base station for the purpose of next stage analysis. The output of the sensors are shown in Fig. 13.

In Fig. 14, the outputs for BME 680 are given where temperature, humidity, gas, and voltage are visible. In the graph, the 'X' axis represents time and the 'Y' axis represents voltage.

In Fig. 15 different acceleration positions are shown.

In Fig. 16 different rotational graphial representations are given.

In Fig. 17 represents the altitude of our system, where the x axis represents time and the y axis represents altitude.

```
Temperature : 25.50 °C
Pressure : 1009.69 hPa
Humidity : 61.88 %
Gas : 217 PPM
Acceleration X: 0.10, Y: -0.33, Z: 9.31 m/s^2
Rotation X: -0.05, Y: 0.02, Z: 0.01 rad/s
INPUT Voltage: 8.08V
Latitude: 23.461380
Longitude: 91.172882
Altitude: 30.18 m
Date: 2/5/2022
Time: 08:17:32.00

Temperature : 25.48 °C
Pressure : 1009.69 hPa
Humidity : 61.99 %
Gas : 237 PPM
```

Fig. 13 Outputs of different sensors

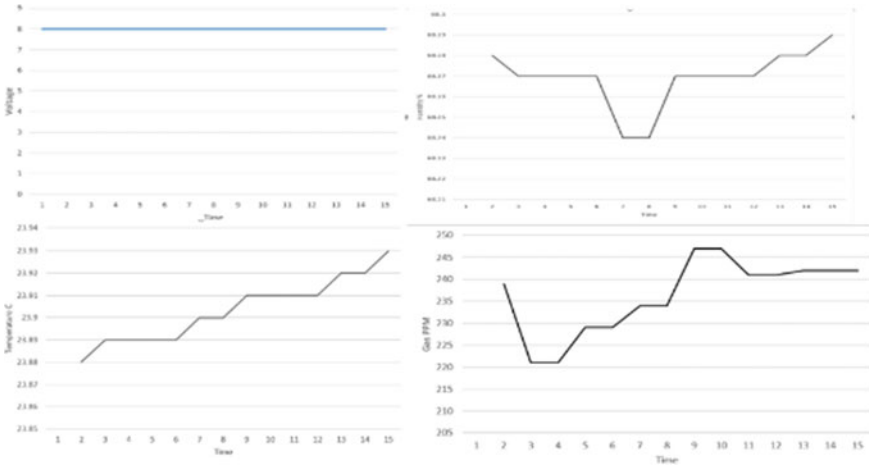


Fig. 14 Graphical representation of temperature, gas, humidity, voltage

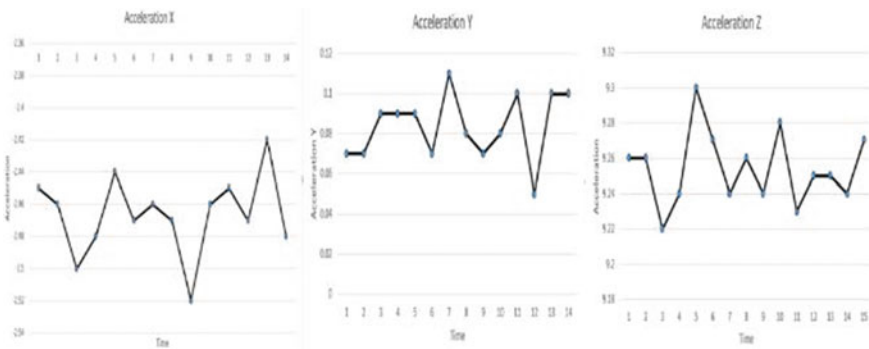


Fig. 15 Graphical representation of acceleration

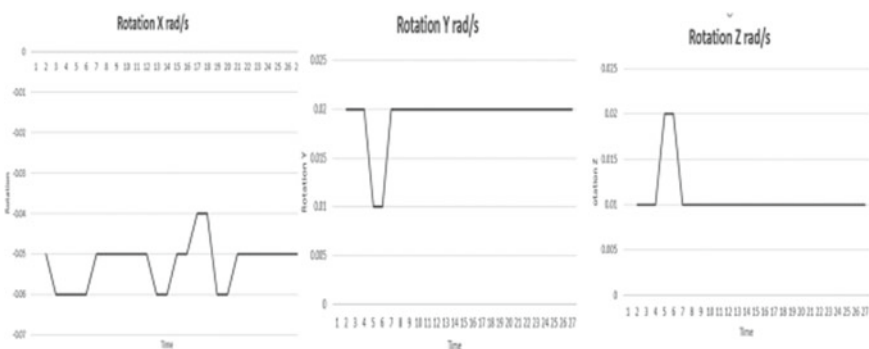


Fig. 16 Graphical representation of rotation

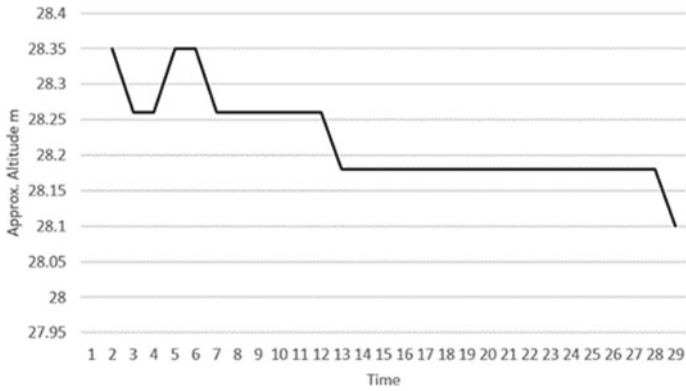


Fig. 17 Graphical representation of altitude

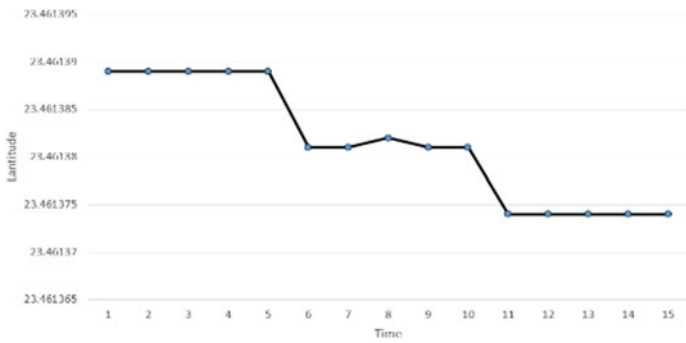


Fig. 18 Graphical representation of latitude

In Fig. 18 represents the latitude of our system, where the x axis represents time and the y axis represents latitude.

In Fig. 19 represents the longitude of our system, where the x axis represents time and the y axis represents longitude.

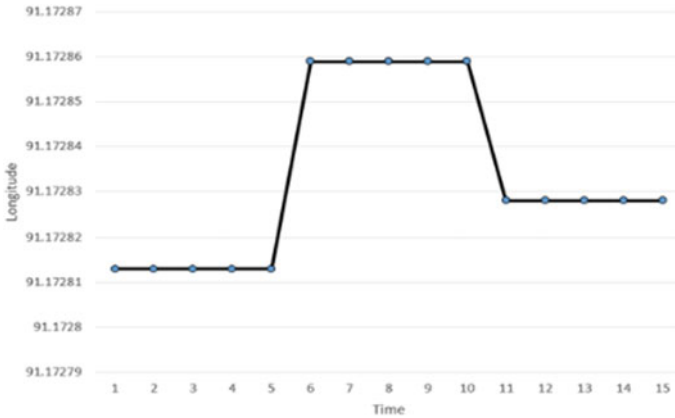


Fig. 19 Graphical representation of longitude

6.2 Cost Analysis

The below Table 2 lists the cost estimation of different components in USD.

7 Conclusion

In our proposed model, we discuss the construction and design of a compact-sized CanSat. As it is difficult to observe the data at different moments, our device is providing data accurately, which we have discussed in “Table 1.” We have discussed different moments of data graphically. The device we are proposing is less expensive than any other option currently available on the market. Although in the future, we are going to reconstruct our can design with carbon fiber and give extra focus on data communication. While collecting the environmental data, we faced some errors. These type of errors are basically for the variation of environmental factors. These environmental factors are very necessary for the precise data collection of cansat.

Table 1 Data observation model at different moments

Data observation	1 st moment	2 nd moment	3 rd moment	4 th moment	5 th moment
Temperature (°C)	23.88	23.89	24.51	25.45	23.94
Pressure (hPa)	1013.21	1012.2	1009.21	1014.22	1014.25
Humidity (%)	68.28	68.27	67.29	67.21	68.29
Gas (PPM)	221	229	229	242	235
Altitude (m)	28.35	28.26	28.26	28.26	28.18
Latitude (m)	23.46	23.46	23.46	23.46	23.46
Longitude (m)	91.17	91.17	91.17	91.17	91.17
Battery input voltage	8.1	8.1	8.1	8.1	8
Acceleration (X, Y, Z) position	-2.45, 0.07, 9.26	-2.46, 0.07, 9.26	-2.5, 0.09, 9.22	-2.48, 0.09, 9.24	-2.44, 0.09, 9.3
Rotation (X, Y, Z) position	-0.05, -0.01, 0.01	-0.06, 0.05, 0.01	-0.05, 0.03, 0.01	-0.06, 0.02, 0.01	-0.05, 0.02, 0.01
GPS date and time	2.5.2022 08:13:46	2.5.2022 08:13:51	2.5.2022 08:13:54	2.5.2022 08:13:57	2.5.2022 08:13:58

Table 2 Cost estimation of different components in USD

Sl no	Components name	Quantity	USD
01	6 V mini solar	1	4.65
02	LM2577S boost converter	1	0.70
03	LM2596 buck converter	1	0.70
04	TP4056 battery charger	1	0.58
05	Transmitter and receiver	2	5.35
06	GPS module	1	3.72
07	Arduino Nano	2	6.98
08	Accelerometer and gyroscope	1	1.40
09	PCB Printing	1	4.65
10	3D print of CANSAT body	1	2.33
11	BME 680 sensor	1	6.98
12	Voltage sensor	1	0.58
13	Wire	2 sets	0.70
Total			39.33

References

1. Ostaszewski M, Dzierzek K, Magnuszewski Ł (2018) Analysis of data collected while CanSat mission. In: 2018 19th international carpathian control conference (ICCC), May 2018, pp 1–4

2. Aydemir ME, Dursun RC, Pehlevan M (2013) Ground station design procedures for CANSAT. In: 2013 6th international conference on recent advances in space technologies (RAST), June 2013, pp 909–912
3. Unknown (2015) Rosetta CanSat team: history of CanSat, Rosetta CanSat Team, 26 February 2015
4. Soyer S (2011) Small space can: CanSat. In: Proceedings of 5th international conference on recent advances in space technologies – RAST 2011, June 2011, pp 789–793
5. Yarce A, Sebastián Rodríguez J, Galvez J, Gómez A, García MJ (2017) Simple-1: development stage of the data transmission system for a solid propellant mid-power rocket model. *J Phys Conf Ser* 850:012019
6. Çabuloğlu C et al (2011) Mission analysis and planning of a CANSAT. In: Proceedings of 5th international conference on recent advances in space technologies – RAST 2011, June 2011, pp 794–799
7. Development of a meteorology and remote sensing experimental platform: The LAICAnSat-1 | IEEE Conference Publication | IEEE Xplore
8. Botero AY, Rodríguez JS, Serna JG, Gómez A, García MJ (2017) Design, construction and testing of a data transmission system for a mid-power rocket model. In: 2017 IEEE aerospace conference, March 2017, pp 1–14
9. Çelebi M et al (2011) Design and navigation control of an advanced level CANSAT. In: Proceedings of 5th international conference on recent advances in space technologies - RAST2011, June 2011, pp 752–757
10. Ümit ME, Cabañas W, Tetlow M, Akiyama H, Yamaura S, Olaleye S (2011) Development of a fly-back CANSAT in 3 weeks. In: Proceedings of 5th international conference on recent advances in space technologies – RAST 2011, June 2011, pp 804–807
11. Bautista-Linares E, Morales-Gonzales EA, Herrera-Cortez M, Narvaez-Martinez EA, Martínez-Castillo J (2015) Design of an advanced telemetry mission using CanSat. In: 2015 international conference on computing systems and telematics (ICCSAT), October 2015, pp 1–4
12. Hasan Raian FMT, Islam HMJ, Islam MdS, Azam R, Islam HMJ, Debnath S (2020) An affordable cansat design and implementation to study space science for Bangladeshi students. In: 2020 IEEE region 10 symposium (TENSYP), June 2020, pp 1205–1208
13. Kizilkaya MÖ, Oğuz AE, Soyer S (2017) CanSat descent control system design and implementation. In: 2017 8th international conference on recent advances in space technologies (RAST), June 2017, pp 241–245
14. Aliyev I et al (2017) Design of solar powered subscale glider for CanSat competition. In: 2017 8th international conference on recent advances in space technologies (RAST), June 2017, pp 453–457
15. Faroukh YM et al (2019) Environmental monitoring using CanSat. In: 2019 6th international conference on space science and communication (IconSpace), July 2019, pp 239–244
16. Sharath Kumar S, Adithya K, Ranjith Srinivas AB, Rao S (2016) Development of CanSat. In: INCOSE international symposium, vol 26, no s1, pp 291–301, November 2016
17. Karuppusamy DP Synchronization of reactive power in solar based DG and voltage regulated elements using stochastic optimization technique. *J Electr Eng Autom*
18. Chancharoen W, Witoon S, Pataranutaporn P, Ngamarunchot B, Theanthong P (2018) The national cansat competition: lessons, challenges, and outcomes of the first cansat competition in THAILAND. <https://doi.org/10.13140/RG.2.2.35374.77127>
19. Marubin D, Yi S (2021) Development of appropriate power distribution design for can-sized satellite (canSAT). *J Adv Ind Technol Appl* 02. <https://doi.org/10.30880/jaita.2021.02.02.001>
20. Bhad B, Akant K (2019) Experimental Cansat for measurement of UV radiation, November 2019, pp 1–4. <https://doi.org/10.1109/ICETET-SIP-1946815.2019.9092213>

Artificial Neural Network Based Fault Diagnosing System



M. Brindha, P. Nabisal Afrine, R. Priyadarshini, and P. S. Manoharan

Abstract The use of an Artificial Neural Network to track and detect a faulty Photovoltaic string accurately at the time of fault occurrence is addressed in this research study. Solar panels are becoming the most efficient renewable energy source. In a hybrid system, any malfunction in the solar panel would cause considerable harm to the entire system, including residential appliances and industrial machinery. Fault detection in photovoltaic (PV) systems is critical for increasing the power output and extending the life of a PV system. The primary objective of this research work is to incorporate an accurate fault detection method. In the proposed model, input data is trained by using an Artificial Neural Network (ANN) technique. The simulation was carried out by using the Matlab/Simulink software.

Keywords PV system · Fault detection · Artificial Neural Network · Matlab/simulink

1 Introduction

Fault detection in a PV system using ANN [1] predicts only the fault and normal condition. However, classification of fault is equally important to correct the system and operate in a normal condition. Deep learning and Convolutional Neural Network [2–4] techniques are used for detecting the fault in PV system by leveraging less accuracy. Fault detection in solar panel is as important as the fault detection in other power generation systems. Since any fault in solar panel can affect the total system, it may also cause damage to the appliances. The most important parameter for generating the power efficiently is by attaining higher efficiency and reduced loss. To improve the efficiency, solar panel should overcome all types of fault. Incorporating an appropriate fault detection mechanism will reduce the overall system damage. Moreover, installing solar panels require more investment and also the damages in solar panel

M. Brindha · P. Nabisal Afrine · R. Priyadarshini · P. S. Manoharan (✉)
Department of Electrical and Electronics Engineering, Thiagarajar College of Engineering,
Madurai, India
e-mail: psmee@tce.edu

will lead to investment loss. In solar power plants, it is highly difficult to monitor the performance and occurrence of faulty conditions in PV panel. The mean time to repair is also high. Thus, the Artificial Neural Network (ANN) [5–8] is selected for detecting and identifying the faults efficiently. ANN is the supervised learning, which trains the data effectively to execute it in any application. For simulation, we have used Matlab, since it implements and test the algorithm easily and perform extensive data analysis.

In this paper, the PV panel is simulated to collect data and feed ANN for detecting four different types of conditions, such as open circuit (OC), short circuit (SC), normal shading and partial shading. Different blocks are simulated in Simulink for each conditions. In this work our main objective is.

- i. To track and identify the faulty PV string accurately at the time of fault occurrence.
- ii. To increase the fault detection accuracy.

2 Methodology

2.1 Artificial Neural Network

Artificial Neural Network (ANN) is a machine learning technology in which inputs are directed through numerous layers of neurons to achieve the desired output. The number of neurons in the input equals the number of samples, and the number of neurons in the output equals the number of neurons in the input. The developer must select the number of hidden layers since it will execute all computational work and actions. If the hidden layer is small, the output will not match the real output, resulting in a huge percent of inaccuracy between the actual obtained output. If the number of hidden layers is increased, it leads to overfitting, which means that it will learn alongside the noise present in the data, producing poor results and taking longer to train. There are two types of learning processes used to train the model:

- i. Supervised learning-output predicted using the neurons trained with labeled data
- ii. Unsupervised learning-output predicted using the neurons trained with unlabeled data

2.2 Photovoltaic System

Nowadays, PV systems are used worldwide. Since it is renewable and clean source of energy. PV system composed of more solar panel array with an inverter and electrical hardware. Each panel produces some amount of energy when it is exposed to sunlight. All the panels linked with each other to produce large amount of energy as Solar Array. The electricity produced from the panel is direct current (DC). For making it useful we are using inverter to convert direct current (DC) to alternating current (AC). This AC electricity from inverter can be used locally or send to the grid for future use. If some faults occur in single solar panel, there is a chance of the whole system getting damaged. So that we implemented this work to rescue the system before it gets damaged and finding out exactly in which panel the fault is occurred.

2.3 Types of Fault

In photovoltaic system, there are three types of fault, they are,

- i. **Open circuit fault**-It is simulated by connecting a resistor with resistance equal to infinite ohm in series with the last panel of the particular string to indicate the OC fault. In OC, the fault current will be zero and voltage across the terminal will be high.
- ii. **Short circuit fault**-It is simulated by short circuiting the positive and negative terminal of the panel using a resistor with zero resistance.
- iii. **Partial shading condition**-Under partial shading condition, insolation over the panel differs. To indicate this each string is insulated with different insolation, where it varies from standard insolation 1000 W/m^2 to reduced level.

3 Problem Formulation

In this paper, the roof top PV system in Department of EEE has been considered. The specification of each module of TP310 LBZ 145 under electrical rating at 1000 W/m^2 of insolation and $25 \text{ }^\circ\text{C}$ of panel temperature is a short circuit current of 8.88 A, an open-circuit voltage of 44.9 V. The overall power rating of this PV system is 25 kW. The various parameters of the PV module considered are shown in Table 1.

The current generated from each string is used for diagnosing the fault in particular string. The considered system consists of six strings of panel connected in parallel is shown in Fig. 3. Each string consists of 20 panel connected in series is shown in Fig. 4. The block diagram of overall system is shown in Fig. 1.

Table 1. Panel datasheet

Parameter	Value
Peak power (Pmax)	310 W
Peak voltage (Pmax)	36.9 V
Peak current (Pmax)	8.41 A
Open circuit voltage	44.9 V
Short circuit current	8.88 A
No. of cells in a panel	72 cells

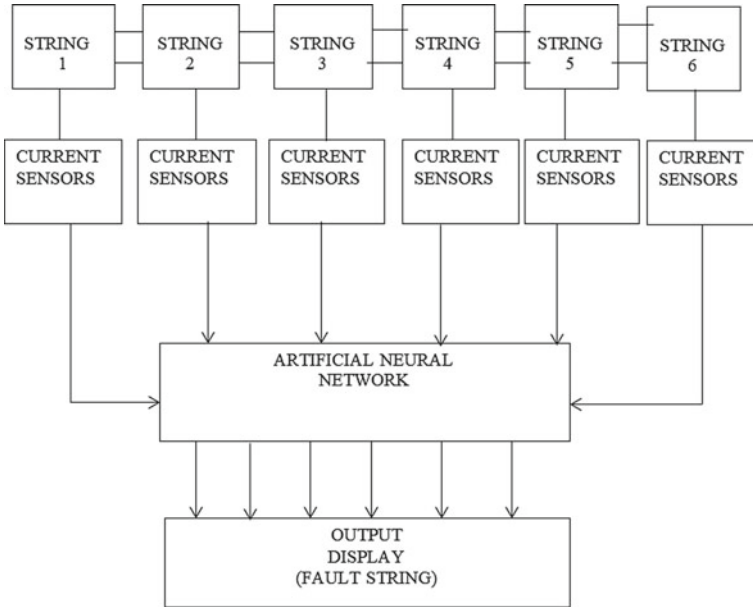


Fig. 1 Block diagram of overall system

PV system is simulated using MATLAB Software with respect to panel configuration and panel datasheet. Each solar cell is configured as 18 cells to form a panel shown in Fig. 2.

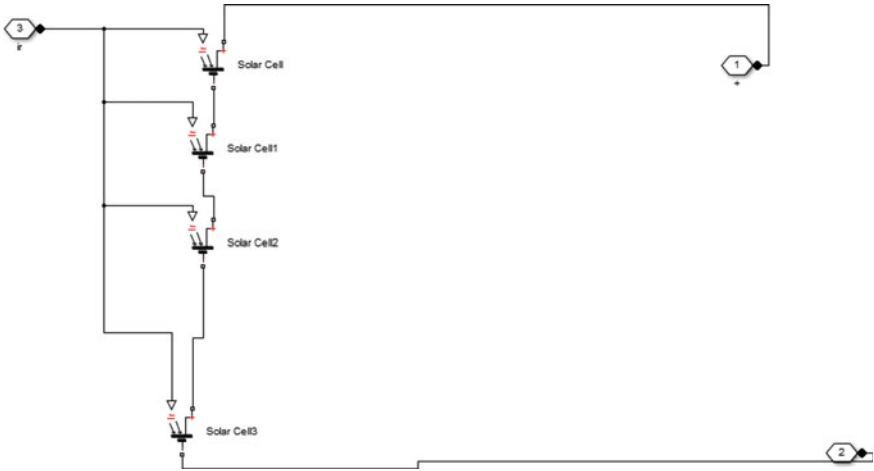


Fig. 2 PV panel

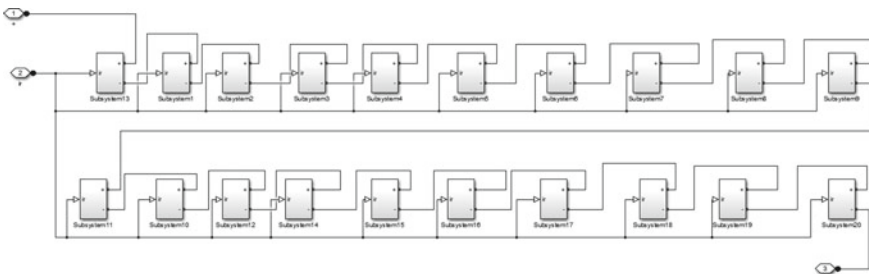


Fig. 3 Panel connection inside each string

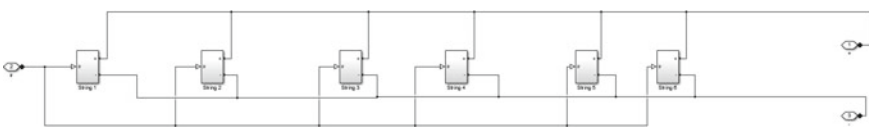


Fig. 4 String connection

4 Results and Discussion

The overall simulation is shown in Fig. 5. The output voltage and current from the overall PV system are measured using a voltmeter and ammeter. The current measured is given as the input to the ANN algorithm for classifying PV faults. In PV system there 20 subsystem shown in Fig. 6 and in each subsystem there are 6 strings as shown in Fig. 7.

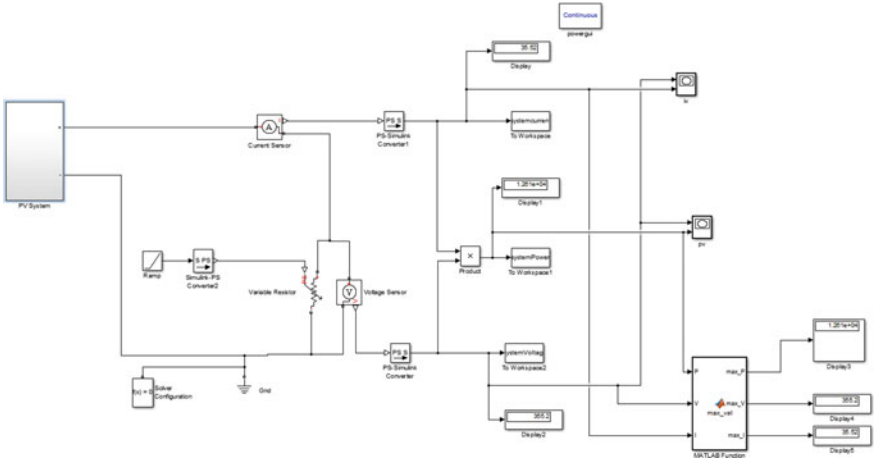


Fig. 5 The overall simulation block

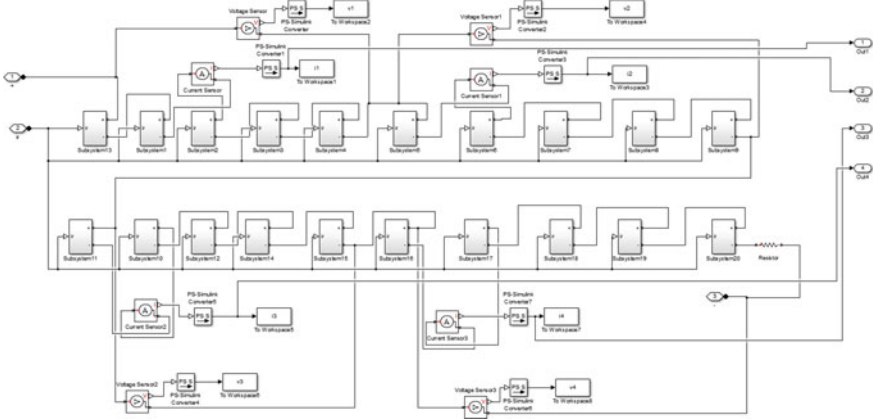


Fig. 6 Subsystem connection

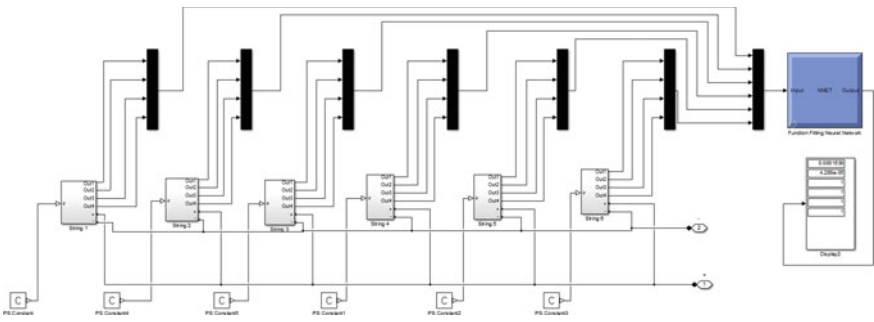
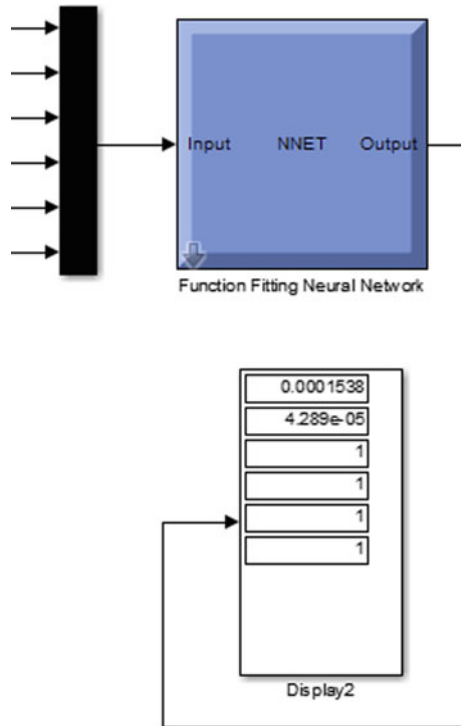


Fig. 7 String connection in each system

Fig. 8 Detection of open circuit fault

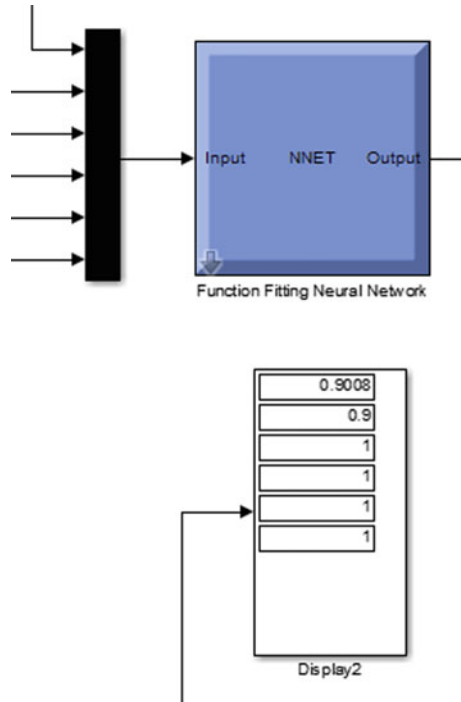


In Fig. 8 Open circuit fault takes place at string 1 and string 2 since current reading shown in output is nearly zero for string 1 and string 2. It was done by disconnecting the particular string. Strings other than 1 and 2 reading current value as 1, which indicates the normal condition.

In Fig. 9 Short circuit fault takes place at string 1 and string 2 since current reading shown in output is 0.9 for string 1 and string 2. Short Circuit fault is implemented by opening a circuit breaker between the panel and the grid system where the supply from the panel is disconnected. String other than 1 and 2 reading current value as 1, which indicate the normal condition.

In Fig. 10 Partial shading takes place at string 1, string 2, string 4, string 5 and string 6. Since current reading shown in output is slightly less than 1. It was by giving insolation less than standard condition that is 1000 W/m^2 . Here insolation for string 1, string 2, string 3, string 4, string 5 and string 6 as 200 W/m^2 , 500 W/m^2 , 1000 W/m^2 , 700 W/m^2 , 700 W/m^2 and 800 W/m^2 respectively.

Fig. 9 Detection of short circuit fault



Data has been imported from workspace in which data has been collected from the solar panel WSM-145 aware energies. The data has two parameters namely current and voltage and these two are predictors. Response has the type of conditions namely normal, open circuit, short circuit and partial shading. Using this we should train, validate and test the data using neural fitting tool. Here, we had collected 2807 samples of voltage and current data in real time. From the collected data.

- 70% of data is used for training
- 20% of data is used for validation
- 10% of data is used for testing

In neural network, there will be three layers namely input, hidden and output layer. Input has respective outputs. Figure 11 shows the structure of neural layers for training the model. In the hidden layers there will be weights in each layer that will give the target output. The number of hidden layers can be set according to the target output.

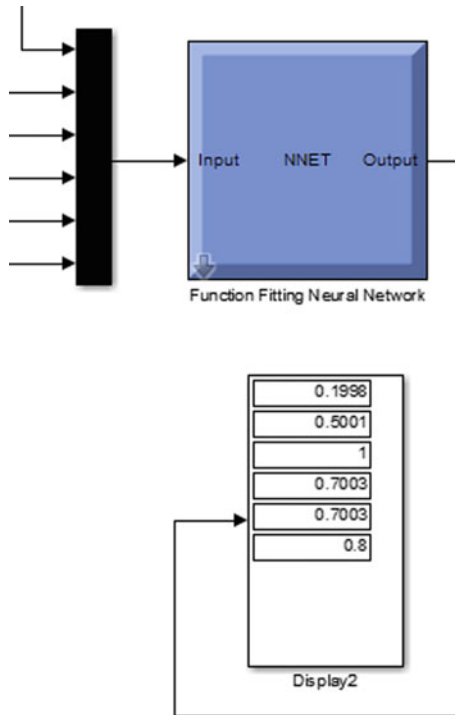


Fig. 10 Detection of partial shading condition

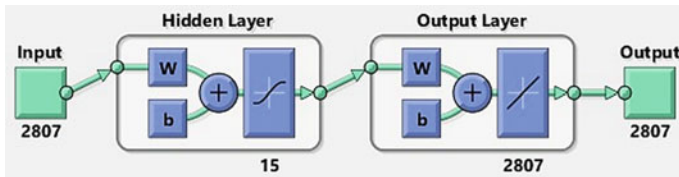


Fig. 11 Neural layers

In neural fitting app, some amount of the data given is taken for training, validation and testing. So from the given data, 75% of data is taken for training, 20% of data is taken for validation and 10% of data is taken for testing. Figure 12 shows the accuracy for training, validation, testing and overall accuracy. The accuracy is nearly 0.9 for training, validation and testing. It takes less time to train the data. Hence ANN is the best method in classifying the fault in PV system.

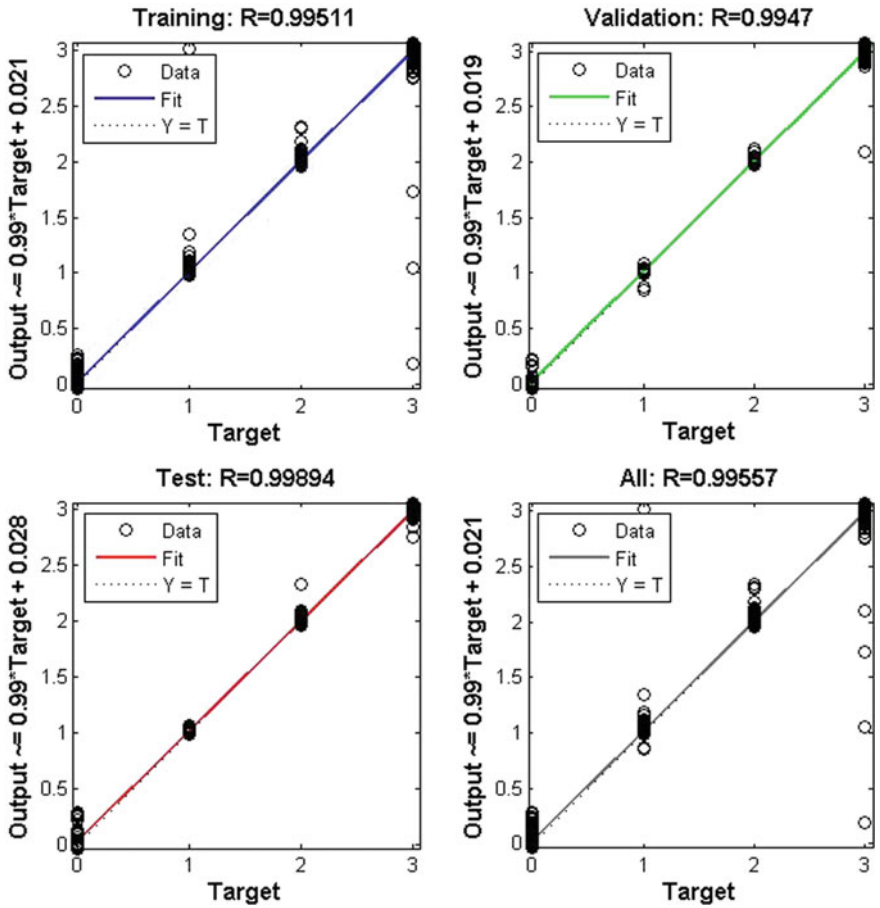


Fig. 12 Accuracy for input data

5 Conclusion

Fault prediction in solar panel is very important. If the grid connected solar panel has any fault, it will lead to the failure of total system and also it will cause damage to the appliances. Here, ANN is selected as a machine learning algorithm to perform fault prediction. In future, an algorithm with best accuracy, less training and testing time can be proposed and also train the model with large number of data than ANN. In future, we may test the model with the Internet of Things [IoT] techniques to perform advanced fault diagnosis.

References

1. Soffiah K, Manoharan PS, Deepamangai P (2021) Fault detection in grid connected PV system using artificial neural network. In: 7th international conference on electrical energy systems (ICEES), pp 420–424
2. Mansouri M, Trabelsi M, Nounou H, Nounou M (2021) Deep learning based fault diagnosis of photovoltaic systems: a comprehensive review and enhancement prospects. *IEEE Access*
3. Aziz F, Ul Haq A, Ahmad S, Mahmoud Y, Jalal M, Ali U (2020) A novel convolutional neural network-based approach for fault classification in photovoltaic arrays. *IEEE Access* 8:41889–41904
4. Fazai R, Abodayeh K, Mansouri M, Trabelsi M, Nounou H, Nounou M, Georghiou GE (2019) Machine learning-based statistical testing hypothesis for fault detection in photovoltaic systems. *Sol Energy* 190:405–413
5. Gielen D, Boshell F, Saygin D, Bazillian MD, Wager N, Gorini R (2019) The role of renewable energy in the global energy transformation. *Energy Strategy Rev* 24:38–50
6. Sun Y, Li S, Lin B, Fu X, Ramezani M, Jaithwa I (2017) Artificial neural network for control and grid integration of residential solar photovoltaic systems. *IEEE Trans Sustain Energy* 8
7. Rizzo SA, Scelba G (2015) ANN based MPPT method for rapidly variable shading conditions. *Appl Energy* 145:124–132
8. Jiang LL, Maskell DL (2015) Automatic fault detection and diagnosis for photovoltaic systems using combined artificial neural network and analytical based methods. In: International joint conference on neural networks (IJCNN)

Smart Application for Voice Over Control on Electronic Devices Using NodeMCU



S. Florence, Lakshmi Narayanan, and K. Meena

Abstract In recent years, automated device control has been a key demand in developing the future electronic devices. The electronic devices are made smart and autonomous by incorporating sensors and other related materials. However, cost is a crucial consideration when adopting the automation initiatives. If automation can be done at a low cost, it will benefit everyone. The proposed system automates the electronic devices such as fans, lights, and refrigerators based on voice. This has been made possible by using Dialogflow, NodeMCU, and Firebase frameworks. Further, the Hub RED is linked to the IoT-NodeMCU device. The NodeMCU can be integrated in any ordinary home appliances. The Hub Red broadcast will be transferred to DialogFlow and further it will be connected with Amazon Alexa or Google Assistant. When the commands are sent, a cloud configuration interface gets installed on a standard NodeMCU. Further, the information is transformed into a signal to switch ON or OFF the household appliances.

Keywords Smart control · Sensors · Automated device control · Node MCU · Hub RED

S. Florence (✉)

Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute Science and Technology, Chennai, India
e-mail: florences@veltech.edu.in

L. Narayanan

Department of Computer Science and Engineering, Department of Networking and Communications, SRM Institute of Science and Technology, Kattankulathur Campus, Chennai, India
e-mail: lakshmir4@srmist.edu.in

K. Meena

Department of Computer Science and Engineering, GITAM School of Technology, GITAM, Bengaluru, India



Fig. 1 Device control through mobile

1 Introduction

In this computer age, people's desire for home automation is growing at an unprecedented rate. However, it needs to be implemented at reduced cost. With the recent technological innovations, voice commands or text input can be used to turn ON or OFF the household items. Moreover, the home appliances can be controlled from anywhere, and the electronic device can be programmed to work for a specific amount of time.

Nowadays automated device control is considered as the basic need especially for the elderly and disable population. Automating the electronic devices will lead to a smart life as shown in Fig. 1. In any situation, developing an automated framework will consume more cost, that is a major setback for the implementation of home automation. The concept can be applied with the help of the available technologies without utilizing any sophisticated methods. The idea of capturing the commands, identifying and modifying the commands into symbols will perform the function according to the specified instructions, and finally maintains the commands in the database to eliminate the signal conversion of the same commands Fig. 2.

2 Literature Survey

Home automation by utilizing the Internet of Things (IoT) technology to operate fans, lighting, and other domestic equipment—is as popular as the latest. Messages transmitted through chatbots are designed and developed to employ common language management systems. By using a customized web application, a chatbot can manage household electronics [1, 2].

The sensor may be knowledgeable about sensing temperature, humidity, light, and movement. The automation framework also improves the order of voice of



Fig. 2 Automated home

consumers. It muffles the customer's voice and concentrates on the special relevance of the order. The system is built over an autonomous Arduino BT board, and apparatus is linked to it through Transfers. Developing a home robot with artificial intelligence approaches can accomplish a wide range of controlled equipment. As a result, there is an issue with energy conservation [3].

The framework has been implemented by using Android objects and Firebase by Google, the latest in the IoT (Objects Web) domain. This function helps to achieve remote computing using the existing devices [4].

The proposed study plans to develop a model for utilizing a key framework for home appliances based on a voice recognition framework. The framework is used to control a variety of household electronics [5].

The effective framework for attaining such a deployment is determined by the IoT technology. The devices included in this framework are linked to Raspberry Pi.

Further, the proposed framework provides a offline control on all local household equipment via a fixed location [6].

As a transmission device, the WSN architecture of home automation framework employs the Atmega328p microcontroller, PC, GUI, and ZigBee [7]. Different sensors are provided by the frameworks for automating the transportation devices [8].

A low-level and user-friendly home automation model is based on a tiny Arduino controller and various sensors. It focuses on the dependability, affordability, and efficiency demands of the user [9]. A home automation system uses an IR remote, Bluetooth and GSM technology to control AC devices by using the Android app is discussed to utilize the traditional switch [10, 15].

The study provides the overall implementation of a low cost appliance control system with a wireless (WiFi) architecture. The system provides permission to access the system at the same time and change the priority. It provides the GUI with low cost embedded systems and user friendly installation of application [11]. Further, a

general architecture has been designed and implemented. Wireless Home Automation System (WHAS) with the help of voice recognition application makes the installation and implementation to be cheaper and easier [12, 14]. The sensors are then used to operate the health related devices [13, 16]. Though the existing home automation system works well, it requires higher implementation cost.

3 Proposed System

Existing research efforts to control household appliances are effective, but at the same time they are more expensive and complex. The proposed solution is less expensive and easier to deploy. Google Assistant, Firebase, and NodeMCU were used to create the ability to control home appliances. Figure 3 depicts the general architecture of the proposed system.

The hardware and software requirements are detailed in the Table 1.

3.1 Modules of Proposed System

The proposed system has been implemented with 3 Modules.

- Voice Input
- Training (DialogFlow)
- Output using NodeMCU

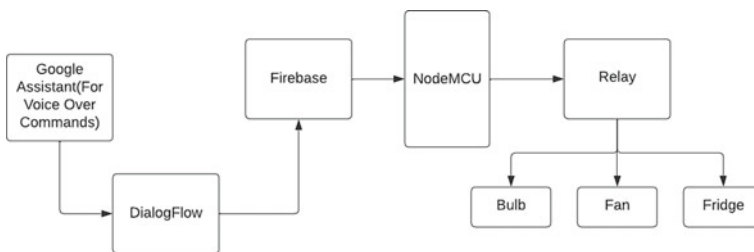
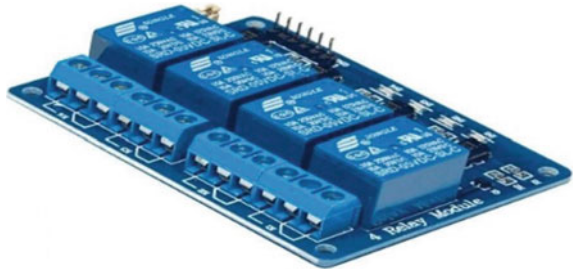


Fig. 3 General architecture of automated home

Table 1 Hardware and software requirements

Hardware requirements	Software requirements
<ul style="list-style-type: none">• NodeMCU• Relay• Electronic Devices• Mobile Phone	<ul style="list-style-type: none">• Google Assistant• Dialog Flow• Firebase• System (i3 processor, 2 GB RAM)

Fig. 4 Relay

Module 1 is mainly focused on input. The input is nothing but the commands given by the user. The commands were captured by using Google Assistant and data will get forwarded to Google Dialog Flow for achieving a better identification of the exact commands given by the user. Identified exact commands will then be passed to the firebase for storage and also the commands will be passed to the NodeMCU for executing the appropriate operations.

All the transfers react to at least one electrical parameters like voltage or flow with an end goal to open or close the contacts or circuits. Figure 4 shows the general appearance of a relay.

The proposed system uses Google Assistant to capture the user requests for a certain task. Transmission is as important as the safety of switching the gadgets to a large control system or gear types. Figure 5 shows the complete architecture of the proposed system. Interaction with the rest of the device is carried out only for this purpose. Each electrical device is then equipped with a NodeMCU. As NodeMCU requires less power, less cost and also have an integrated Wi-Fi support, it has been selected for connecting the devices present in the proposed system.

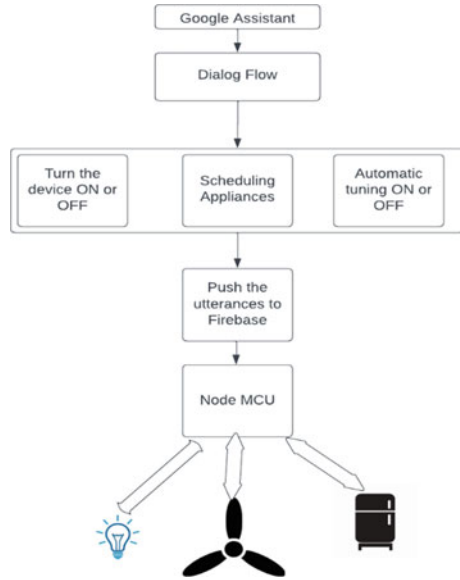
The parallel wires allow the flow of electricity to start at one point in the circuit and then move on to the next in the light of the fact that power needs a way to move. Most of the connecting wires are made of copper or aluminium.

Dialogflow is a framework created by Google for representing the real-time applications, which enables the interaction between users and computer with natural language processing abilities. It has been used in the proposed work as the bot gets associated with the application connected devices based on the user given commands. It has been integrated with the Google Assistant for making the required communication. Firebase provides real-time and backend database as a built-in program.

4 Results and Discussion

Figure 6 shows the input data that has been used to test the proposed system. The home automation framework is incorporated to control the home appliances remotely from any location. By connecting client gadgets with client Wi-Fi, or other smart gadgets,

Fig. 5 Proposed system architecture



the client’s device can operate the electronic devices for switching heating, cooling, and lighting operations remotely. It can facilitate a daily or continuous commitment of electronic devices to work automatically based on the command.

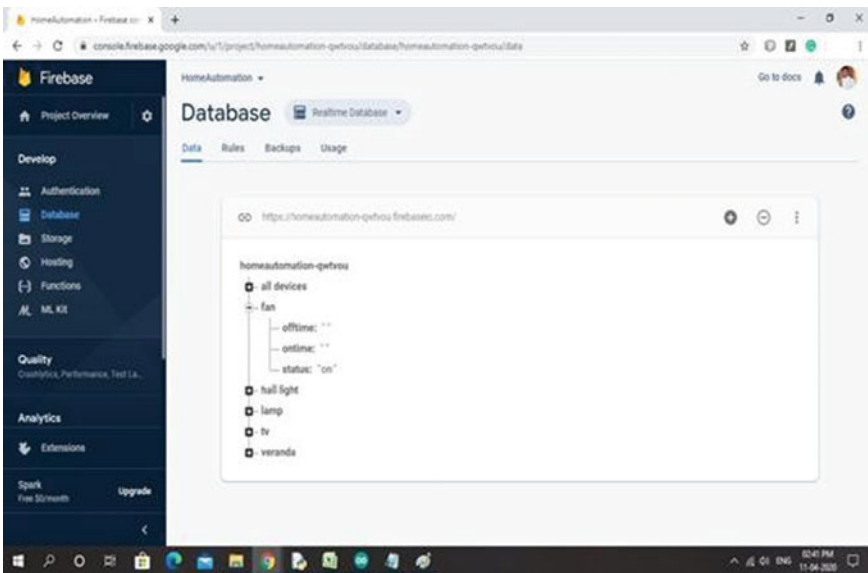
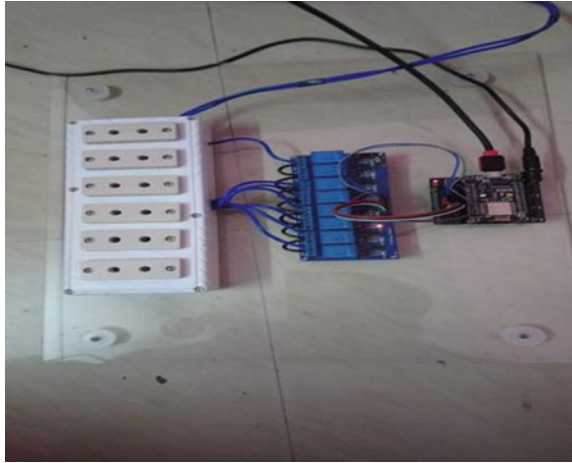


Fig. 6 Sample input data given to connected devices

Fig. 7 Sample hardware output



As manufacturers continue to improve the transparency, safety, and security of their electronic goods. The most recent household electronics and smart gadgets are loaded with intelligence and exceptional performance than earlier versions. Figure 7 depicts the hardware response to human input.

5 Conclusion

The goal of the proposed automated device control using Google Assistant is to manage the electrical and electronic devices in a person's household. The way the project was discussed was efficient, and the installation was successful. This framework is quite robust and assists the elderly and disabled population to switch electric and electronic devices smartly without relying on the traditional ON/OFF switch.

6 Future Enhancements

Various sensors can be used to improve the smart home safety and control, such as a weight sensor, which can be placed outside the home to notify when someone is about to enter the home. The automation bot can also be made more adaptable in terms of providing electricity credit. Similarly, the customer might consider the cost of influence on a daily basis and set aside money in the same manner.

References

1. Mustafa B, Iqbal MW, Saeed M, Shafqat AR, Sajjad H, Naqvi MR (2021) IOT based low-cost smart home automation system. *IEEE*, pp 1–6. <https://doi.org/10.1109/HORA52670.2021.9461276>
2. Salvi S, Geetha V, Kamath SS (2019) Jamura: a conversational smart home assistant built on telegram and google dialogflow. *IEEE*
3. Thakare S, Yadav S, Waghade A (2019) IoT and AI bases home automation system. *IJRESM*
4. Singh S, Verma S, Kumar S, Kumar S, Verma P (2019) Home automation using node MCU, Firebase IOT. *IJRESM*
5. Dorve J, Samarth MK, Jais R, Sheikh MdDS, Kumar P, Korde H (2019) A review on home automation using voice via bluetooth through Raspberry PI3. *IJRESM*
6. Bepery C, Baral S, Khashkel A, Hossain F (2019) Advanced home automation system using Raspberry-Pi and Arduino. *IJRESM*
7. Kodali RK, Soratkal S (2017) MQTT based home automation system using ESP8266. *IEEE*
8. Florence S, Shyamala Kumari C (2019) Big data and IoT in smart transportation system. *Int J Innov Technol Explor Eng (IJITEE)* 8(9). ISSN 2278-3075
9. Baby CJ, Khan FA, Swathi JN (2018) Home automation using IoT and a chat-bot using natural language processing. *IEEE*
10. Shinde A, Kanade S, Jugale N, Gurav A, Vatti RA, Patwardhan MM (2017) Smart Home automation system using IR, bluetooth, GSM and android. *IEEE*, pp 1–6. <https://doi.org/10.1109/ICIP.2017.8313770>
11. Haque ME, Islam MR, Fazle Rabbi MT, Rafiq JI (2019) IoT based home automation system with customizable GUI and low cost embedded system. *IEEE*, pp 1–5. <https://doi.org/10.1109/STI47673.2019.9068035>
12. Sravanthi G, Madhuri G, Sharma N, Tiwari A, Kashyap A, Suresh B (2018) Voice recognition application based home automation system with people counter. *IEEE*, pp 574–578. <https://doi.org/10.1109/ICACCCN.2018.8748409>
13. Florence S, Kumari CS, Priyadharshini LL (2018) Smart health monitoring system based on internet of things with big data analytics and wireless networks. *Int J Eng Technol (UAE)* 7(1.7):109–111
14. Chakraborty T, Datta SK (2017) Home automation using edge computing and Internet of Things. *IEEE*, pp 47–49. <https://doi.org/10.1109/ISCE.2017.8355544>
15. Singh A, Mehta H, Nawal A, Gnana Swathika OV (2018) Arduino based home automation control powered by photovoltaic cells. *IEEE*, pp 729–732. <https://doi.org/10.1109/ICCMC.2018.8488144>
16. Manoharan JS (2021) Capsule network algorithm for performance optimization of text classification. *J Soft Comput Paradigm (JSCP)* 3(01):1–9
17. Shakya S (2021) A self monitoring and analyzing system for solar power station using IoT and data mining algorithms. *J Soft Comput Paradigm* 3(2):96–109
18. Kottilingam D (2020) Emotional wellbeing assessment for elderly using multi-language robot interface. *J Inf Technol Digit World* 2(1):1–10
19. Gupta L, Varma N, Agrawal S, Verma V, Kalra N, Sharma S (2021) Approaches in assistive technology: a survey on existing assistive wearable technology for the visually impaired. In: *Computer networks, big data and IoT*. Springer, Singapore, pp 541–556

Author Index

A

Abinaya, M., 347
Abu-Zaideh, Saja, 803
Agarwal, Varun Niraj, 99
Agrawal, Jitendra, 139
Ahmed, M. Athiq, 603
Akshaya Krishna, N., 161
Al Hasan Anik, Md. Abdullah, 869
Al-Haija, Qasem Abu, 803
Alam, Md. Nahidul, 519, 869
Alam, Shireen Rafat, 201
Anagnostopoulos, Theodoros, 483
Anandaraj, S. P., 561
Anisha, N., 395
Anuraj, K., 171, 425
Ar-Reyouchi, El Miloud, 715

B

Banu, Sufia, 621
Baranidharan, B., 661
Belhoussine Drissi, Taoufiq, 547
Belwal, Meena, 331
Bharathi Mohan, G., 831
Bhatt, Himanshu, 55
Bhinge, Nikhil A., 265
Bongarde, Prasad, 217
Brindha, M., 885
Buiyan, Md. Zahid Hasan, 869
Butko, Igor, 573

C

Chandak, Sharda, 217
Chandran, Saravanan, 453

Chandrasekharan, Deepti, 123
Charan, Marrivada Gopala Krishna Sai,
171, 425
Chaudhary, Rahul, 781
Chavhan, Nishant, 217
Chaware, Bhushan, 217
Chowdhury, Maruf Haider, 519
Cimtay, Yucel, 743, 757

D

Deepa, T., 769
Deepak, P. V., 675
Dewangan, Deepak Kumar, 299
Ding, Danny, 849
Dongre, Swati, 139
Doriya, Rajesh, 201
Duraipandian, N., 381

E

Ezhilarasan, M., 467

F

Florence, S., 897

G

Ganesamoorthi, M., 603
Ganguli, Souvik, 781
Garg, Romy, 287
Ghoumid, Kamal, 715
Glukhov, Sergey, 573
Gupta, Bhoomi, 287

Gupta, Chekka Venkata Sai Phaneendra,
171, 425
Gurulakshmi, A. B., 265

H

Hao, Feifan, 587
Hariprasad, S., 769
Harsoor, Bharati, 649
Hegde, Nayana, 621
Hindu, K., 533
Hiremath, Manjunatha, 191

I

Injam, Sri Latha, 613
Islam, Md. Rayhanul, 519

J

Jain, Saurabh, 201
Jaishree, M., 689
Jakka, Aishwarya, 731
Jayachandran, A., 395, 561
Jayanth, S., 275
Jerusha, J. Ansulin, 497
Jeyasekar, A., 497
Jing, Shan, 587
Josephine, D. Diana, 603
Joshi, Devanshu, 703

K

Kalimulin, Temir, 573
Kalpana, T., 675
Kalpande, Shyamkumar D., 229
Kanade, Vijay A., 793
Kanshi, Avaneesh, 99
Karthikeyan, M., 675
Kaur, Narinder, 287
Kavipriya, K., 191
Kavyashree, M. K., 67
Khizhnyak, Irina, 573
Khudov, Hennadii, 573
Khurram, Mohammed, 123
Komisopoulos, Faidon, 483
Kommineni, Kiran Kumar, 27
Kumar, B. P. Pradeep, 621
Kumar, R. Ajay, 603
Kumar, Sharan, 55
Kumar, Tarun, 407
Kumar, V. S. Selva, 1
Kumari, Shweta, 407

L

Lajoie, Isabelle, 715
Lakshmi Surekha, T., 613
Lal, Anisha M., 99
Lamisa, Nazifa Tahsin, 869
Lun, Jun, 849

M

Madhu, G. C., 27
Makoveichuk, Oleksandr, 573
Manoharan, P. S., 885
Manoj, N., 265
Md., Osman Khan Zeeshan, 613
Meena, K., 815, 897
Melarkode, Navneet Vinod, 99
Meshwin, A., 603
Mishra, Pavan Kumar, 437
Mohan, Poornima, 275, 319
Mohod, Chandrashekhar D., 229

N

Nabisal Afrine, P., 885
Naik, Sukanya D., 161
Nainwal, Ankita, 113, 703
Nanda, Pranamita, 381
Nanthini, K., 407, 675
Narayana, Y. L., 533
Narayanamurthy, Rajadurai, 27
Narayanan, Lakshmi, 897
Nayak, Deekshitha S., 55, 161
Nazre Amarnath, Tejas Kumar, 123
Nsiri, Benayad, 547
Ntalianis, Klimis, 483
Nur Yilmaz, Gokce, 743

P

Pamarthi, Meghana, 613
Pant, Bhaskar, 113
Parkhi, Vrinda, 219
Patel, Sunil Kumar, 453
Patil, Milind S., 229
Pavan, M., 275
Pavithra, S., 15
Pon Suresh, Manoharan, 815
Poorna, S. S., 171, 425
Prabakeran, S., 347
Praneeth, Choragudi Sai, 171, 425
Prasad, Sanjana, 621
Prasanna Kumar, R., 831
Pravin, Sheena Christabel, 1
Priya, K. Sindhu, 1

Priyadarshini, R., 885
Pyingkodi, M., 407, 675

R

Rahman, Atikur, 869
Raj, Amrit, 367
Rajkumar, R., 689
Rani, S. Rashmi, 621
Rao, Niharika, 55
Rao, Sannidhi, 55
Rattal, Salma, 715
Rian, Shahrukh Hossain, 519
Ruban, Igor, 573
Rukmini, S., 637

S

Sabarinathan, C., 661
Sahu, Jagmohan, 437
Sahu, Satya Prakash, 299
Sahu, Sonal, 299
Sai Pradeep, K. P., 689
Sai Sumanth, P. G., 171, 425
Sakthivelan, R. G., 15
Salmon, Ioannis, 483
Sambhram, V., 161
Samimalai, Arun, 621
Santhosh, H. M., 265
Saranya, J., 1
Satish, Palavalasa Venkata, 331
Satish, S., 41
Saxena, Rohan, 287
Sefraoui, Omar, 715
Shah, Dikshan, 251
Shamrai, Nazar, 573
Sharma, Garima, 113, 703
Sharma, Sanjeev, 265
Sheela, S. Antony, 497
Shekhawat, Sanjay P., 229
Shetty, Krishang, 161
Shetty, Sahana, 161
Shreelatha, G. U., 67
Shyamala Kumari, C., 815
Singh, Gurpreet, 27
Singh, Partheesh Ranjan, 123
Sivabalaselvamani, D., 407

Snober, Mohammad Abu, 803
Sobhana, M., 533
Soma, Shridevi, 637
Sreema, M. A., 561
Sridevi, S., 41
Srikar, Kota, 171, 425
Srivastav, Pragun, 319
Sudarson Rama Perumal, T., 561
Suganthi, S., 1
Suma, S., 649
Sunandha Shri, S., 467
Sushwanth, Y., 275
Svetaa, S., 15

T

TamilSelvan, S., 689
Tejesh, N., 275
Thenmozhi, K., 675
Thillai Rani, M., 689
Toulni, Youssef, 547
Tripathi, Swathi, 123
Tripathi, Vikas, 703

U

Ullah, Mahfuz, 519
Uppalapati, Sri Sai Bharat, 319

V

Vaithiyashankar, Jayakumar, 367
Vakula Rani, J., 731
Venkatasubramanian, S., 81
Vykuntam, Praneeth, 319
Vykuntam, Venkata Rohith, 319

Y

Yahiaoui, Réda, 715
Yamini, A., 533
Yilmaz, Gokce Nur, 757
Yogesh, O. M., 265

Z

Zhao, Chuan, 587