



**PANIPAT INSTITUTE OF
ENGINEERING & TECHNOLOGY**

Approved by AICTE, New Delhi & Affiliated to Kurukshetra University, Kurukshetra)



IRCET2021

PROCEEDINGS OF
**INTERNATIONAL
CONFERENCE ON
INNOVATION, RESEARCH
AND CHALLENGES IN
EMERGING TECHNOLOGIES**



**November
19-20, 2021**

Organised by:
**Department of Information Technology and
Department of Electronics & Communication Engineering**

**Proceeding of International Conference on
Innovation, Research and Challenges in
Emerging Technologies**

IRCET – 2021

November 19-20, 2021

Chief Editor

Dr. Shakti Kumar

Editors

Dr. Mukesh Chawla

Dr. Vaishali M. Wadhwa

Associate Editors

Sorabh Gupta

Ashima Arya

Organised by

Department of Information Technology and

Department of Electronics & Communication Engineering

Panipat Institute of Engineering and Technology

70 Milestone, GT Road, Samalkha, Panipat 132102 Haryana, India

Copyrights @ 2021 by IRCET and PIET All rights reserved.

ISBN : 978-81-955434-4-1

Published by : ABC Computers Printers

No part of the material protected by this copyright may be reproduced or utilized in any form or by any means, without the prior written permission of the copyright owners, unless the use is a fair dealing for the purpose of private study, research or review. The authors reserve the right that their material can be used for purely educational and research purposes.

Copyright and Reprint Permissions: Abstracting is permitted with credit to the source.

The papers in this book comprise the proceedings of the conference mentioned on the cover and title page. They reflect the authors' opinions and in the interests of timely dissemination, are published as presented, without change. Their inclusion in this publication is solely the responsibility of the author and does not necessarily constitute endorsement by the editors.

CONTENTS

About PIET

Vision and Mission

Messages

Organizing Committee

Advisory Committee

List of Accepted papers

Full Papers

ABOUT THE PIET

Panipat Institute of Engineering & Technology (PIET) was established by the Vidhyapeeth Education Trust and Located in the National Capital Region (NCR). The founders have a clear vision to provide world class education and to promote research. Among some of the best Engineering Institutes in NCR, PIET stands out as a clear leader. A multidisciplinary institute affiliated to Kurukshetra University, Kurukshetra and approved by AICTE, New Delhi. PIET strives to encourage and make the budding scholars into skilled professionals through its various programmes e.g. Diploma, Undergraduate and Postgraduate programmes in different disciplines. PIET has consistently pushed the bar of excellence to garner a niche for itself amongst the top institutes in Delhi/NCR in just a few years by being prudent and dedicated towards its aim of life ennobling engineering education. Due to the emerging scenario of entrepreneurship, PIET has launched its start-up and incubation center as well as entrepreneur development cell. PIET is one in only 49 institutions in the country to be awarded with a grant of 55 Lacs by AICTE for setting up the Idea Lab. The institute provides complete infrastructure for effective functioning of the Training & Placement Cell. Training activities are organized throughout the year in an effort towards preparing the prospective students for the campus selection program.



VISION

To be known and recognized as an institute for Science, Technology, Management and Research globally which has a transformative impact on the society.

MISSION

PIET's mission is to provide widely recognized leadership in the improvement of teaching, and learning, thereby, educating our students with the knowledge, skills and creativity to manage our nation's enterprise competently and confidently.

MESSAGE



Sh. Hari Om Tayal
Founder and Chairman, Panipat Institute of Engineering and Technology

It is a moment of immense pleasure to announce that Panipat Institute of Engineering and Technology is organizing an International Conference on Innovation, Research and Challenges in Emerging Technologies (IRCET-2021). This International conference is a giant forum of intellectuals, which aims at bringing the best of professionals/researchers, visionaries, scholars and technocrats to a common stage to discuss the latest emerging technologies. In this era of smart sciences, it is very essential for all the professors/researchers to dive into the emerging trends of the 21st century. In fact, in the current scenario we cannot imagine world without artificial intelligence, cloud computing, data sciences and mathematical computing. We are blessed with these technological boons to simplify our life at every single step. We hope the upcoming conference would prove itself to be the ultimate platform to showcase new perspectives and insight.

MESSAGE



Sh. Suresh Tayal
Secretary, Panipat Institute of Engineering and Technology

With a pledge to achieve a significant position in the highly competitive world of Engineering and Technology, Panipat Institute of Engineering and Technology is organizing an International Conference on Innovation, Research and Challenges in Emerging Technologies (IRCET-2021). In order to face various emerging challenges across different fronts of Engineering and Technology, it has become indispensable to explore multifarious integrated and inter-disciplinary engineering approaches. This conference focuses primarily on encouraging novel ideas and emerging technologies and as to how this shared knowledge can be spread globally. I wish our efforts to promote new ideas in Intelligent Systems by facilitating intellectuals to showcase their innovation and findings builds a great place in society.

MESSAGE



Sh. Rakesh Tayal
Vice Chairman, Panipat Institute of Engineering and Technology

With a significant rise in technological advancements, Panipat Institute of Engineering and Technology has taken a constructive step by organizing an International Conference on Innovation, Research and Challenges in Emerging Technologies (IRCET-2021). Our prime objective is to meet expectations of the invited delegates and create networking opportunities through quality discussions and deliberations. Carrying an endeavour to achieve excellence in this one-of-a-kind conference, our efforts have gone into making this conference an enriching experience for the delegates by making it a perfect knowledge-sharing platform. I am sure that the learning from this conference coupled with the fresh ideas, enthusiasm and vision, will definitely make this conference a sublime experience for all the attendees.

MESSAGE



Sh. Shubham Tayal

Member BOG, Panipat Institute of Engineering and Technology

The technological progress of our country witnessed in the last few years is far more promising than that over the last few decades, which reflects how these advancements have set up a new benchmark. The faculty members of Panipat Institute of Engineering and Technology having sensed this need of the hour and created a platform for researchers all around the world to bring about their thoughts and views to help the society at large. This indeed is a great platform for the exchange of technical know-how from the experts in the industry which will surely lead to an even better and enhanced learning opportunity for all concerned and actively involved in making it a great success. With immense gratitude I extend warm wishes to the Conveners, Co-coordinators, and Organizing Committee members for initiating the "International Conference on Innovation, Research and Challenges in Emerging Technologies (IRCET-2021)".

MESSAGE



Dr. Shakti Kumar
Director, Panipat Institute of Engineering and Technology

Twenty first century is the century of great technological and industrial advancement. A competent, confident and vast knowledge force is the most sought-after propeller needed for the social growth. Keeping the need of the hour in mind, Panipat Institute of Engineering and Technology is organizing an International Conference on Innovation, Research and Challenges in Emerging Technologies (IRCET-2021). A conference is one of the finest medium of new knowledge creation and its dissemination. In present scenario where the world has shrunk to a small global village, meeting the needs of present times has become a massive challenge. Conferences play a vital role in meeting such challenges. I positively anticipate that this conference would serve as a significant platform to connect various academicians, researchers, and scholars through a common motive to go beyond the borders in search of new frontiers in the field of intelligent systems, research and development.

MESSAGE



Dr. Vijay Anant Athavale

Conference Chair

It a great pleasure for me to extend a welcome to all of you who are participating in IRCET 2021, the International Conference on Innovation, Research and Challenges in Emerging Technologies, which is being held in PIET, Panipat, India.

The conference theme is a message of empowered optimism as we come out of pandemic and face 2022 dedicated to creating our future. Through the collaboration and connection of IRCET 2021 we continue to push the Innovative technologies forward.

I am looking forward to meet you during IRCET 2021 and accord most pleasant, interesting and fruitful conference.

MESSAGE



Dr. S. C. Gupta



Dr. Anju Gandhi

Conference Co-chairs

Welcome to the International Conference on an International Conference on Innovation, Research and Challenges in Emerging Technologies (IRCET-2021).

This conference is being organized to provide an opportunity to academic researchers, industry professionals, government delegates and students to interact and share their experiences and knowledge on cutting-edge developments in the fields of Artificial Intelligence, Cloud Computing, Internet-of-Things, Networked Systems, Data Communications, Computer Security, Image Processing, Computer Vision, Intelligent Computing, Robotics, Machine Vision, Signal Processing, Physical Sciences and Mathematical Sciences in Education.

We hope this conference will result into very fruitful scientific outcomes.

MESSAGE



Dr. Mukesh K. Chawla



Dr. Vaishali M. Wadhwa

Conveners

It is our privilege to welcome all the delegates to the International Conference on Innovation, Research and Challenges in Emerging Technologies (IRCET-2021). The theme of the conference indeed attains importance in the millennium, as it will help academicians, researchers and scholars to demonstrate the key issues prevalent in technological advancements happening worldwide in industrial and manufacturing sector. The International Conference is inspired by the need for a platform to address various emerging issues in engineering. We believe this conference is essential to all researchers who will get real opportunities to showcase their research. We hope this conference will result into very fruitful Scientific and Engineering outcomes; the outcomes which shall benefit the living beings in general and human race in particular.

Organizing Committee

Conference Chair	
Dr. Vijay Anant Athawale	
Conference Co-Chair(s)	
Dr. S. C. Gupta	Dr. Anju Bhandari Gandhi
Convener(s)	
Dr. Mukesh K. Chawla	Dr. Vaishali M. Wadhwa
Co-Convener(s)	
Dr. Bhawna Singla	Dr. Deepak Wadhwa
Dr. Monika Gambhir	Sorabh Gupta
Organizing Secretary	
Rattan Deep Aneja	Swati Gupta
Technical Programme Chair(s)	
Dr. Dinesh Verma	Dr. Vinay Khatri
Publication Chair(s)	
Neha Bhatia	Naveen Saini
Bhawna Kukreja	Sachin Dhawan
Ashima Arya	Payal Chhabra
Session Organizing Committee	
Dr. Shakti Arora	Sandeep Goel
Arun Rana	Harminder Kaur
Dr. Nitisha Aggarwal	Karuna
Jyati Mahajan	Mitu Sehgal
Internal Advisory Board	
Dr. Manoj Arora	Dr. Deepak Raj
Dr. Deepak Bhagwat	Dr. Akhilesh Mishra
Dr. Sunil Dhull	Dr. Rajat Kumar Baldua
Dr. Tanvi Singh	Amit Dubey

Prof. Atul Gautam

Advisory Board

Dr. M. N. Hoda , BVICAM, New Delhi, India	Dr. Rohit Kumar , University of Southern California Los Angeles, California, USA.
Dr. B. B. Gupta , National Institute of Technology, Kurukshetra, India	Dr. Deepak Kumar Jain , Congqing University of Posts and Telecommunications, Chongqings, China
Dr. L.K. Awasthi , National Institute of Technology , Jalandhar, India	Dr. Saad Subair , College of Computing Studies, International University of Africa
Dr. Nonita Sharma , National Institute of Technology, Jalandhar, Punjab, India.	Dr. Dharam Singh Jat , Namibia University of Science and Technology, Namibia
Dr. Dinesh Khanduja , National Institute of Technology, Kurukshetra, India	Dr. Ayodeji Salau , Afe Babalola University Ado-Ekiti, Nigeria.

Dr. Pardeep Singh , National Institute of Technology, Hamirpur, India	Dr. Shahnawaz , University College of Bahrain, Bahrain
Dr. N.P. Singh , National Institute of Technology, Kurukshetra, India	Dr. Ruchika , Indian Institute of Technology, Delhi
Dr. Kanika Sharma , NITTTR, Chandigarh	Dr. Jitender Chhabra , National Institute of Technology, Kurukshetra, India
Dr. Richa Bhatia , NSUT, East Campus, New Delhi	Dr. Kamlesh Dutta , National Institute of Technology, Hamirpur, India
Dr. Ravi Panwar , IIITDM, Jabalpur, MP	Dr. Deepak Garg , Bennett University, Greater Noida, India
Dr. Victor Gambhir , Sh. Guru Gobind Singh Tricentenary University, Gurugram	Dr. Achyuth Sarkar , National Institute of Technology, Arunachal Pradesh, India
Dr. Rajesh Mehra , NITTTR, Chandigarh	Dr. Judhistir Mahapatro , National Institute of Technology, Delhi
Dr. Rashmi Gupta , NSUT, East Campus, New Delhi	Dr. P. C. Tiwari , National Institute of Technology, Kurukshetra, India
Dr. Anil Kumar Verma , Thapar University, Patiala	Dr. Dilbagh Panchal , National Institute of Technology, Hamirpur, India
Dr. Sachi Nandan Mohanty , College of Engineering Pune, Pune	

Expert Committee

Dr. Rakesh Kumar , Kurukshetra University, Kurukshetra, India	Dr. Sachin Wadhwa , Karnal Institute of Engineering and Technology, Karnal
Dr. Shailender Singh , Punjab Engineering College, Punjab, India	Prof. (Dr.) Rajender Nath , Kurukshetra University, Kurukshetra, India
Dr. Vineet Goyal , Bhagwan Mahavir University, Surat	Prof. (Dr.) Suchita Upadhayaya , Kurukshetra University, Kurukshetra, India
Dr. Sunita Dhingra , Maharishi Dayanand University, Rohtak, India	Dr. Chander Kant , Kurukshetra University, Kurukshetra, India
Dr. Nikhil Marriwala , UIET, KUK	Dr. Pooja Sapra , North Cap Uni., Ggm
Dr. Sandeep Goyel , Maharishi Markandeshwar University, Mullana, India	Dr. Dinesh Singh , DCRUST, Murthal, Sonapat, India
Dr. Sunil Dhingra , Kurukshetra University, Kurukshetra, India	Dr. Amit Kumar Bindal , Maharishi Markandeshwar University, Mullana, India
Dr. Amit Kumar , Jaypee University, Solan, Himachal Pradesh	Dr. Mrityunjay Singh , Jaypee University of Information Technology, Solan, Himachal Pradesh, India
Dr. Sanjay Tyagi , Kurukshetra University, Kurukshetra, India	Dr. Sunil Gupta , University of Petroleum Energy and Studies, Dehradun, India
Dr. Sukhvinder Singh , Maharishi Dayanand University, Rohtak, India	Dr. Shelly Garg , GNIOT, Greater Noida
Dr. Deepali Gupta , Chitkara University, Punjab, India	Dr. Suman Lata , LPU, Jalandhar
Dr. Manoj Sharma , Maharaja Ranjit Singh PTU, Bhatinda, India	Dr. Gaurav Sharma , Jai Parkash Mukand Lal Institute of Technology, Radaur, India

Dr. Saurabh Gupta, SGT, Ggn

Dr. Shubham Tayal, SR University,
Warangal, Telangana

Dr. Vikrant Bhateja, SRMGPC, Lucknow

Dr. Deveder Prasad, Chitkara University,
Punjab, India

Dr. Sunita Satpathy, Sri Sri University,
Bhuvneshwar

Mr. Ramanjeet, Om University, Hisar

Dr. Poonam Mittal, YMCA, Faridabad

Dr. Dolly Sharma, Amity University,
Noida, India

Dr. Sarabjeet Singh, Punjabi University,
Patiala

Dr. Sunil Kumar, University of Petroleum
Energy and Studies, Dehradun, India

Dr. Tapas Kumar, Manav Rachna Inst. of
Research and Studies, Faridabad, India

Dr. Rakesh K. Yadav, KCC Institute of
Technology and Management, G. Noida

Dr. Monika Mangla, LTCOE, Navi
Mumbai, India

Dr. Subhash K. Shinde, LTCOE, Navi
Mumbai, India

Dr. Nitin Goyal, Chitkara University,
Mohali

Dr. Dinesh Chutani, Delhi Technical
University, New Delhi, India

Dr. Suman Deswal, DCRUST, Murthal,
Sonapat, India

Dr. Megha Bhushan, DIT University,
Dehradun, H.P.

Dr. Anurag Jain, University of Petroleum
Energy and Studies, Dehradun, India

Technical Program Committee

Vishal Jain, PIET

Rajesh Kumar, PIET

Rajiv Dhanda, PIET

Rajeev Saini, PIET

Gaurav Gambhir, PIET

Akanksha Mahajan, PIET

Shiraz Khurana, PIET

Alankrita Aggarwal, PIET

Kunal Uppal, PIET

Tannu Sharma, PIET

Karun Handa, PIET

Dr. Rajender Kumar, PIET

Sudhir Mahajan, PIET

Sameeksha Khare, PIET

Sunny Kuhar, PIET

Shivani Gaba, PIET

Richa Grover, PIET

Satyawan, PIET

Vikas Goyal, PIET

Reema Lalit, PIET

Rohit, PIET

Sapna Arora, PIET

Pankaj Batra, PIET

Yeeshu Rehlan, PIET

Deepti Dhingra, PIET

Navneet Verma, PIET

Sandeep Singh Bindra, PIET

Anoop Arya, PIET

Stuti Mehla, PIET

Deepak Singla, PIET

Shekhar Singh, PIET

Neeraj Tyagi, PIET

Vikas Sethi, PIET

Puneet Sharma, PIET

Sumit Kumar, PIET

Vipin Tomar, PIET

Mohit Lalit, PIET

Pawan Kumar, PIET

Rajan Saluja, PIET

Nisha, PIET

Mandeep Kaur, PIET

Tarun Miglani, PIET

Table of Contents

Paper Title	Authors	Page No.
Automated Four-Wheeler Vehicle Collision Avoidance System	Syed Ali Fathima S J, Asfaq Moideen A, Pavitra Ravi, Shaik Mehafus, Vinishavathani S, Yuvanesh M	1-7
Comparative Study on Generative Adversarial Networks in Text to Image Generation	Syed Ali Fathima S J, Surya Kangeyan K G, Lokeshwaran N G, Kavın Pradeep R	8-14
Exploration of Artificial Intelligence's Application in Software Engineering: Current and Future Roadmap	Isha Sharma, Pardeep Singh, Rahul Kumar Singh, Alankrita Aggarwal	15-21
Landmark Recognition with Artificial Intelligence: A State of the Art	Kanishk Bansal, Amar Singh	22-37
Maintenance Directives of Analysis and Design for Complex Embedded System through Reverse Engineering	Surinder Pal Singh, Ankur Mangla, Vishal Gupta, Charanjit Singh	38-42
Blockchain for IOT Security & Privacy	Chaitali Choudhary, Dr. Inder Singh	43-47
Gradient Integrated Regression Sustainable Approach with Machine Learning towards Software Quality Assurance	Alankrita Aggarwal, Kanwalvir Singh Dhindsa, P. K. Suri, Pardeep Singh	48-52
COVID-19 Diagnosis: An Eye of Machine Learning Techniques	Anshul Raju Kumar	53-60
Software reliability model with FSM: A Study	Rahul Kumar Singh, Pardeep Singh, Gourav Bathla, Alankrita Aggrawal	61-66
A Survey on Jumping finite automata	Rahul Kumar Singh, Pardeep Singh, Alankrita Aggrawal, Isha Sharma, Shivali Gaba	67-76
Privacy Concerns in Big Data Security – Survey	Aashita Chhabra, Manu Narula, Aakanksha, Vedant Sahrawat	77-91
An Indirect Approach to Audit Structural Health of Bridges	Rakhi Akhare, Monika Mangla, Nonita Sharma, Vaishali Mehta	92-98
Recognition of Pneumonia from the X-Ray image pattern using Convolutional Neural Networks	Sanjay Santhanam, Dr. Selvi S, Sathya S, Thyagaraj T, Yanamadala Vasu Deva Sai, Kalaivani P	99-104
Search for Alzheimer's Type Dementia Pattern from Spontaneous Speech: Signal Processing Approach	Akta Rani, Deepak Kumar, Suresh Chand Gupta	105-110
An Empirical Study on the Impact of IT Services on End-User Satisfaction for University of Petroleum and Energy Studies	Jatendar Sharma, Gagan Deep Singh, Sunil Kumar, Sorabh Gupta	111-119
Machine Learning Techniques for Heart Disease Prediction	Ashima Arya, Mitu Sehgal, Neha Bhatia	120-126
IoT enabled obstruction evasion Robots for enhancing the security of the Systems	Ashima Arya, Shivam Mudgil, Dr. Sapna Juneja, Dr. Abhinav Juneja	127-131
Readiness towards Cashless Payments: India	Atul Gautam, Sanjeev Kumar, Kunal Aggarwal	132-140
Emotion Based Music Recommender System In Cloud	Pavan Kumar, Pradeep Arya, Atul Mishra, Sunil Gupta	141-147
Implementation and Comparative Analysis of Various Energy Efficient Clustering Schemes in AODV	S. Dhanabal, K. Vengatesan, P. William, Udayabhanu N P G Raju, S. Sangeetha	148-162
Plant image Classification: State of Art	Sukanta Ghosh, Amar Singh	163-169
Study on Mushroom Image Classification Using Different Machine Learning Methods	Janvi Shetty, Anushree Rahate, Suvarna Pansambal	170-177
Artificially Intelligent System for Tourism Industry	Ashima Arya, Rattan Deep Aneja,	178-185

	Mitu Sehgal, Amit Kumar bindal, Shakti Kumar	
Wireless sensor network communication protocol: A Survey	Swati Juneja, Dr Karamjit Kaur, Dr Hardeep Singh	186-191
Role of IoT Based Self Automated Agriculture System	Dr. Sangeeta, Preeti Dahiya, Dr. Annu Dahiya, Sakshi Wadhwa, Ananya*	192-195
OTT Streaming: Real Time Engagement	Dr. Annu Dahiya, Vikas Nain, Dr. Sangeeta, Riya Arora	196-200
Integration of AI and IoT in Supply Chain Management of Indian SMEs	Siddharth Sahu Anchal Gupta	201-216
Smart Breast Cancer Diagnosis using Machine Learning Techniques	Amit Kumar Jakhar, Mrityunjay Singh, Shivam Pandey	217-222
Progressive Strategies For Identification Of Straggler Node In Hadoop Mapreduce Environment	Ajay Kumar Bansal, Ashu Gupta, Manmohan Sharma,	223-228
Security and Privacy preservation in blockchain based IoT systems	Priyanka, Dr. Hemant Verma, Dr. Ritu Makani	229-234
An Automated Video-Password based Authentication System using Electrocardiogram	Shahina Anwarul	235-240
A Novel Intuitionistic Fuzzy Inference System for Sentiment Analysis	Makrand Dhyani, Sanjay Kumar, Govind Singh Kushwaha	241-248
Leveraging Machine Learning for Ransomware Detection	Nanda Rani, Sunita Vikrant Dhavale	249-256
Identification and removal of Code Smell	Jaswinder Singh, Kanwalvir Singh Dhindsa, Jaiteg Singh	257-261
Data Security in Cloud Environment: A Comprehensive Survey & Open Challenges	Sartaj Singh, Dr. Amar Singh, Dr. Ashok Sharma	262-267
Virtualization Attacks and Prevention Solutions in Cloud Environment	Vaneeta, Sandeep Kumar	268-273
Impact of Dataset Properties on Spam Email Classification Using Machine Learning	Pramod Prakash Ghogare, Manoj P. Patil	274-283
Load Forecasting Based Smart Grid Using Machine Learning	Pushpa, Sanjeev Indora, Mukhtiar Singh	284-294
Fertilizer Recommendation and Prediction using Machine learning	Sonu Sharma, Poonam Mittal, Payal Gulati	295-306
TCP/IP Benefits over ATM Architecture with Special Reference to 5G Network	Deepty Dubey, Thipendra Pal Singh, Arup Bhattacharjee	307-316
Design and Implementation of Augmented Reality for Smart Driver Assistance System	K. Baskar, P. William, S. Yuvaraj, K. Vengatesan, Rahul Bibave	317-329
Hybrid Machine Learning Technique for Personality Classification from Online Text using HEXACO Model	P. William, Dr. Abhishek Badholia, Dr. Anurag Sharma, Dr. Vijayant Verma	330-340
Implementing Variability Techniques in Software Systems	Ashima Arya, Bhawna, Payal Chhabra, Neha Bhatia, Karuna	341-346
A Modified approach for Accuracy Enhancement in Intruder Detection with optimally Certain Features	Shivani Gaba, Shally Nagpal, Alankrita Aggarwal, Dr. Deepak Kumar	347-353
Designing Next Generation Recommender Systems	Chaitali Choudhary, Dr. Inder Singh	354-362
Protection of Transformer using Accelerated Convolution Neural Network	Harish Balaga, Marrapu Deepthi, Venkata Sravani A	363-367
Performance and Comparison of Various Machine Learning Classifier for Sentiment Analysis on Amazon Dataset	S.Madhusudhanan, Jyothi N. M, M. Ramesh, N. Arun Vignesh	368-373
The Perceived Benefits and Risks of Cloud Computing in Terms of Organizational Cost and Security: An Exploratory	Veshala Lokesh Kumar, Veshala Likhitha, Spuritha M	374-382

Hyper-parameter Optimization of Image Classifiers using Nature-Inspired Algorithms	Rahul Sharma and Amar Singh	383-390
BERLin: Bidirectional Encoder Representations from Linformer	Sandeep Kumar, Jerin Verghese, Ritesh Dutta	391-394
An IoT based 4-Wheeled Safety Monitoring System	Harsh Yadav, Gunjan Chhabra	395-402
Blockchain based identity management system for food supply chain	Dr. Anuradha Konidena, Neha Bhatia	403-408
Crop Yield Prediction & Fertilizer Recommendation System	Suresh Chand Gupta, Shekhar Singh, Deepak Kumar	409-413
Clinical and Epidemiological Factors Analysis for COVID-19 by Mamdani-Fuzzy Inference System	Poonam, Monika Mangla, Nonita Sharma, Reena, Vaishali Mehta Wadhwa	414-427
Parametric Optimization of Friction Stir Welding for Electrical Conductivity of Aluminium Joints Using ANN Approach	Maneesh Tewari, Devaki Nandan	428-435
Effect of Thickness on Laser Induced Nonlinear Optical Properties of Erbium Doped ZnO Thin Films	Vinay Kumari, Mukesh Kumar	436-441

Automated Four-Wheeler Vehicle Collision Avoidance System

¹Syed Ali Fathima S J*, ²Asfaq Moideen A*, ³Pavitra Ravi*, ⁴Shaik Mehafus*, ⁵Vinishavathani S* and ⁶Yuvanesh M*

¹Assistant Professor, Computer Science & Engineering,
^{2,4,6}UG Student, Electronics and Communication Engineering,
³UG Student, Bio-technology Engineering
⁵UG Student, Civil Engineering

^{1,2,3,4,5,6}Kumaraguru College of Technology, Coimbatore , India

¹syedalifathima.sj.cse@kct.ac.in, ²asfaqmoideen.20ec@kct.ac.in, ³pavitraravi.20bt@kct.ac.in,
⁴shaikmehafus.20ec@kct.ac.in, ⁵vinishavathani.20ce@kct.ac.in, ⁶yuvanesh.20ec@kct.ac.in

Abstract: Every year, the road traffic accidents cause the death of approximately 1.3 million population. A lot of factors contribute to accidents, some of which are non-use of seatbelt, over speeding, bad weather, bad roads, distracted driving, driving under the influence of psychoactive substances like alcohol and many more. While many of the factors can be avoided, one main factor which is key to such tragic accidents is speed. Controlling the speed of a car avoids most of the fatal car accidents from happening. The increase in speed is directly proportional to the chances of an accident. To prevent collisions, many automotive companies have equipped their vehicles with advanced technologies. Electric cars nowadays have advanced systems to predict and mitigate collisions. With current conditions in mind, the advanced features like automatic drive and cruise control are exclusive to expensive vehicles. The objective of this paper is to propose an efficient collision avoidance system which can be installed in cars with no such fancy technology. The proposed technology will basically detect an object from afar, measure the distance to know the probability of a collision. If the distance between the car and the obstacle is less, the system will produce a series of audio and visual cues. If the driver doesn't respond to the cues, the car slows itself at first. If the driver is still inactive, the car will be stopped automatically.

Keywords: Microcontrollers, Actuators, Collision, Object detection , Alert system, Distance sensors.

1 Introduction

Accidents are the uncertain events that happen in our daily life. Among the accidents, road accidents are the major threat and have always been a major cause of concern all over the world. Road accidents leads to major trauma and loss of lives. In 2019, more than 151 thousand fatalities have been reported in India due to road accidents [1]. According to the World records, India accounts for 11% of the accidents which is highest in the world but contributes to only 1% of the vehicles. This is a bad situation and share of road accidents death in India is shown in Fig.1. About 4.5 lakhs of accidents happen in a year and about 1.5 lakh people die because of these accidents. Considering the road crash injuries and deaths, a report indicates that India is the topper in the world. "It has 1 percent of the world's vehicles but accounts for 11 percent of all road crash deaths, witnessing 53 road crashes every hour; killing 1 person every 4 minutes. About 13 lakh people lost their lives and 50 lakhs got injured in the last decade".

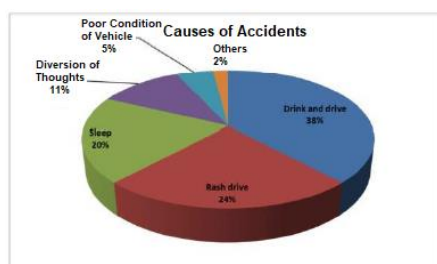


Fig. 1. Share of road Accident Deaths in India

In the year 2019, 449,102 accidents happened in India causing 451,375 injuries and 161,113 deaths recorded in Road Accident Report 2019. When compared to previous year 2018, the accident numbers reduced by 3.86%, the persons injured reduced by 3.86%. and the accident-related deaths reduced by 0.20 % . These

decrease in killings and injury due to road traffic crashes are reported in 2019 appear to be a consequence of the upgrading the vehicles focusing on road safety. This included severe hike in penalties for violating traffic rules, inter-alia, and stiff monitoring through electronic enforcement. There is a decrease in the accident rate because the safety of the vehicle have been improved a bit. The accident rates can be completely decreased when we focus on the safety of the vehicles and improve some of its features. On the other hand, looking into the reports of collision and the crash ratios from the Ministry of Road Transport and Highways, the crash costs at a staggering rate of 3.14 percent of the country's GDP (or) estimated to value of Rs 5.96 lakh crore. Globally too, road traffic injuries are the chief cause of death and stand as the 10th leading reason of all collapses. It makes up an astonishingly significant part of the universal problem of bad health issues. Every year, the road traffic accidents cause the death of about 1.3 million population and 50 million are injured that takes thirty to seventy percentage of population in hospitals which is now becoming a major issue of any developing country. Continuing the similar trend in future may result the road accidents to be the 3rd foremost contributor to the global account of injury and disease soon.

Unexpected crashes and accidents are predictable and preventable, but suitably advanced technology is important to understand the ways in which vehicles can be made safer for people, inclusive of the driver, passengers, and pedestrians. Incorporation of safety features in public transport vehicles and other common vehicles in developing countries is the need of the hour. [3]. An important goal in the field of saving lives can be achieved by providing driving aids with the aim of preventing traffic accidents and reducing the number of traffic victims.

Assistance systems that alert the driver of impending collisions through audible or visual signals have been incorporated in commercial cars nowadays. While these techniques are vital and provide a warning to the driver, the dependence on the driver's reaction time vis-à-vis a decision to be made continues to be a problem. The most frequent road accidents in urban areas are the result of over speeding and consequent sudden, abrupt braking that demands a prompt response from the driver's side. Considering this, our system proposes direct actuation on the brake and clutch pedals. Risky situations like blind spots on curved roads, roads having poor or bad visibility, or unanticipated braking of the car ahead suggest that a quick response system involved with automatic action on the brakes is needed to avoid accidents.

The proposed solution aims to design an automated collision avoidance system that can be installed in existing manual vehicles which controls the vehicle during emergencies to prevent or reduce the severity of a potential collision. To demonstrate the achievement of the stated aim, this project will monitor the speed of the vehicle, calculate the braking distance with respect to speed, alert the driver on impending danger, determine how to express constraints on an alternative algorithm, and prevent a fatal accident.

2 Literature Review

Juergen Dickmann et al., defined an outline of using radar in rising automobile electronics. Dense factor cloud generation, co illustration of static and dynamic environment, velocity detection to enhance vehicle safety has been proposed [4]. Amir Mukhtar et al. Made an evaluation of sensors and queuing techniques. Analysis of verification techniques, tracking, and nighttime detection the use of sensors which do not use machine learning. Vehicle detection mostly relies upon machine learning. These kinds of sensors are capable of tracking different types of vehicles. This paper analyzes state-of-the-art automobile detection techniques. It begins with an overall performance assessment of sensors which infers that active sensor works properly under harsh climate conditions, however face interference issues. 3D laser scanners can efficaciously classify vehicles. [5].

Lovely Gaur, Imdad Rizvi. suggested that by using simple non-colored cameras, identifying the brake(danger) lights the object detected and gives signal to the PU. Number plate detection for vehicle detection. Comparing six sample videos, graphs have been plotted representing and 80 to 95 is evaluated as their success ratio [6]. Messaoud Doudou et al. defined the overall structure of the bio-sensor system which consists of signal collection, data collection, and control modules. Analyzing latest development of biosensors in the market which can be used to deal with driver drowsiness problems and be more efficient. [7].

Jianqiang Wang et al. Implied based on the statistic analysis, identified the preferred braking stress by the means of imitating a driver via the usage of models. Collision detection algorithms and automated braking

are developed primarily considering the driver's pedal action and they proposed an algorithm identifying the driver characteristics which is used as parameters of driver model may be recognized from the information in the manual phase which used the models, and the end result is carried out at some stage in the automated phase. [10].

Pragyaditya Das et al. Proposed a system that includes Infrared transmitter and receiver. IR rays are transmitted into the eye via the means of the transmitter. The output is high when the eyes are shut (or) closed for certain fraction of second and the output is low in case the eyes are open. This output is transferred to an alarm system fixed to outside and inside the vehicle triggering a respective action. The same device can be linked to the braking system to automate the process which may be used to slow down a vehicle[11].

3 Existing Solutions

The developments in collision avoidance are huge and can be realized when the data from multiple sensors and systems are integrated in proper way. As an advancement, the goals set by automakers for increased safety is evident. Collision avoidance systems are widespread in some countries, such as in the EU, United States and in research projects including some devices which are specific to certain manufacturers. Some vehicles have collision avoidance systems fitted as a standard which are predominantly present in 20% of the new and advanced cars as Automated Braking System, Autonomous Emergency Braking, etc. The remaining 80% of the vehicles are old and do not have the Collision avoidance system. Collisions often occur in buses, lorries and in common manual cars. To prevent this, all vehicles need to be fitted with this collision avoidance system which is easily installable in any type of vehicle.

4 Proposed System

4.1 Objective

- To avoid serious car accidents, injuries and fatalities.
- To detect an imminent crash.
- To provide a safer driving experience.
- To improve decision strategies for active safety driving.

4.2 Workflow

When a four-wheeler approaches an obstacle, the Millimeter-wave radar sensor detects it and sends a signal to the microcontroller which will go on to produce a series of audio and visual cues. If the driver doesn't respond to it, the microcontroller will send a signal to the actuators. The actuators will apply full clutch and half of the brakes first to slow down the vehicle. Then if the danger of a crash is imminent, the brake will be applied completely, and the vehicle will be stopped.(Note: the extent of braking completely depends on the braking distance. If an obstacle suddenly crops up in a short distance, both brakes and clutch will be applied completely).

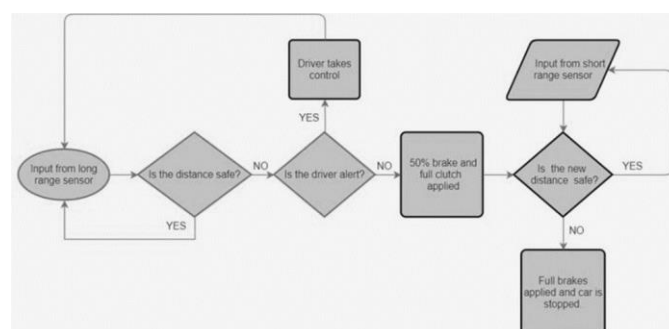


Fig. 2. Workflow of the system

5 Components

5.1 Microcontroller

Microcontroller acts as the central unit of the system. The microcontroller we intend to use is Arduino mega. The reason for the usage of this microcontroller is the fact that it is easy to use and relatively cheaper than other microcontrollers. It has more memory space and a higher processing power. This microcontroller takes the input from the sensors and processes it to give a signal to the actuators.

5.2 Millimeter Wave Radar

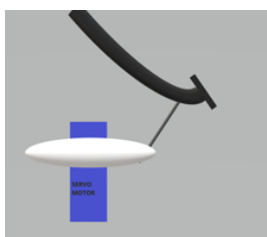
Millimeter-wave radar sensor is a sensor which is used to determine the range, velocity and distance of an obstacle. The reason for picking millimeter-wave radar as the long range sensor is because it is efficient in extreme weather conditions like rain, glare, fog and so on.

5.3 Alert System

This system gives a series of audio and visual cues to the driver before actuating the brakes. The main objective of this system is to alert an inactive driver so that he can avoid the crash by himself. If he doesn't respond, the system will proceed to actuate the brakes. The alert system includes an LCD display, a buzzer and seatbelt as a stimulating device which uses vibration to alert the driver.

5.4 Actuators

The use of actuators in this system is to apply the brakes and clutches automatically in case the driver fails to respond. Two custom actuators are fixed to the brake and clutch pedals. When the sensors detect an obstacle which is unavoidable, the microcontroller gets input from the sensors and sends a signal to the actuators to apply the brake and clutch.



(a)

Fig. 3(a). clutch at normal position



(b)

Fig. 3(b). clutch at actuated position

5.5 Embedded Coding

The coding is used for the functioning of embedded systems. The components used in the program are given unique variables. No functions are created. The input and output pins for the sensors and actuators are determined. The main part of the coding is done using switch and else if cases. The switch case is used to determine the Analog input given by the driver. The else if case is used to determine the speed and distance given by the ultrasonic sensors.

6 Results and Inference

This collision avoidance system is a good start towards the development of a more intelligent and capable collision warning system. This system has been installed in all the automated cars nowadays, but this device has been made applicable even to normal cars and vehicles. This device is suitable for all types of cars. Usually, the collision avoidance system is applicable only to expensive cars but this device is cost efficient and can be installed easily. In this device sensors have been used to detect the objects which are in the road as an obstacle. Millimeter wave radars have been used to detect the objects which are in longer distance.

Ultrasonic sensors have been used to detect the objects which are in shorter distance. Arduino mega has been used in order to make the process easy and liable. Therefore, this device is much simpler, cost efficient, and is compatible with all the cars and vehicles.

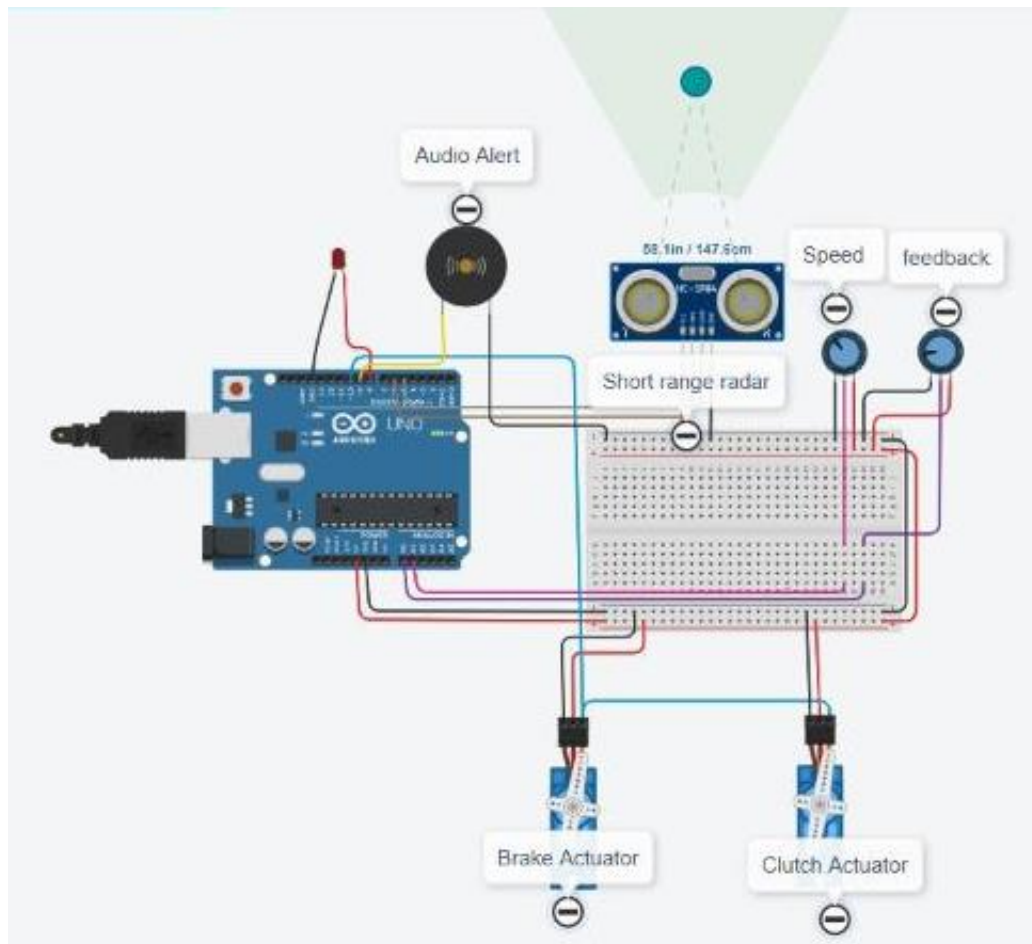


Fig. 4. Virtual simulation done using Tinkercad

The output from the Arduino simulator:

```

Serial Monitor
Vala: 53; CarSpeed: 21; Distance: 166 cm
Vala: 53; CarSpeed: 21; Distance: 166 cm
Vala: 53; CarSpeed: 21; Distance: 166 cm
Vala: 53; CarSpeed: 21; Distance: 164 cm
Vala: 53; CarSpeed: 21; Distance: 164 cm
Vala: 53; CarSpeed: 21; Distance: 165 cm
Vala: 53; CarSpeed: 21; Distance: 166 cm
Vala: 53; CarSpeed:

```

Fig. 5. Output from serial monitor

Table 1. Vala, CarSpeed and Distance obtained as output

Vala	CarSpeed	Distance
171	78	298 cm
171	95	302 cm

7 Discussions

The device "Automated collision Avoidance system " is intended to be installed on any four-wheeler vehicles in cost efficient way. The installation is simple and cost effective. This system consists of basic components like arduino, mmWave radar and ultrasonic sensors, which will give most efficient sensing, detecting and expected outcome. This solution to collision can even be installed in older vehicles. It has been made simpler and easy to install. Most importantly it avoids the collision very efficiently and this device will overcome the serious situation. In this system we can map the object due to continuous wave patterns on the object, but it cannot tell what the exact object is. By using machine learning and twin cameras as they use in autonomous cars, we can know what exactly the obstacle is.

8 Conclusion

The main reasons for the vehicle collision and the effective ways to reduce that is by working on active and passive safety systems of vehicles has been elaborated in the present study. The research problem is significant because the project outcome addresses the strategies of avoiding collision in public transports like buses and trucks which so far never had the facility for the eternal systems to avoid collision. Till now other collision avoidance systems will give only visual and audio alerts but this system will stop the vehicle by applying clutches and brakes at the right time. Since the braking system works by actuators this may have some errors on continuous driving and moreover this system is externally installed. This project has been virtually prototyped using tinkercad.

References

- [1] Number of deaths due to road accidents across India from 2005 to 2019 <https://www.statista.com/statistics/746887/india-number-of-fatalities-in-road-accidents/>, last accessed 2021/11/10.
- [2] Traffic Crash Injuries and Disabilities : The Burden on Indian Society (English), <http://documents.worldbank.org/curated/en/761181612392067411/Traffic-Crash-Injuries-and-Disabilities-The-Burden-on-Indian-Society>, last accessed 2021/11/10.
- [3] Road Traffic Accidents Increase Dramatically Worldwide <https://www.prb.org/resources/road-traffic-accidents-increase-dramatically-worldwide/>, last accessed 2021/11/10.
- [4] Dickmann, Juergen. :Automotive Radar the Key Technology for Autonomous Driving: From Detection and Ranging to Environmental Understanding, IEEE Radar Conference (RadarConf) (2016).
- [5] Mukhtar, A., Xia, L., Tang, T. B.: Vehicle Detection Techniques for Collision Avoidance Systems: A Review, IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 5, pp. 2318-2338 (2015).
- [6] Gaur, L., Rizvi, I.: Improved Vehicle Collision Avoidance System, Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), pp. 1071-1075 (2018).
- [7] Messaoud Doudou, Abdelmadjid Bouabdallah, Veronique Cherfaoui.: A Light on Physiological Sensors for Efficient Driver Drowsiness Detection System. Sensors & Transducers Journal, International Frequency Sensor Association (IFSA), vol. 224 (8), pp.39-50 (2018).
- [8] Oken, Barry, Salinsky, M.: Alertness and attention: Basic science and electrophysiologic correlates. Journal of clinical neurophysiology : official publication of the American Electroencephalographic Society. Vol. 9. pp. 480-94 (1992).
- [9] Dr.Jayaprakash ,Dr.Mirza Muhammad Naseer,Muhammad Ajmal khaN, - .Road Traffic Accidents Research In India : A Scientometric Study From 1977 To 2020, <https://digitalcommons.unl.edu/libphilprac/4325/>, last accessed 2021/11/10.
- [10] Wang, J., Zhang, L., Zhang, D., Li, K.: "An Adaptive Longitudinal Driving Assistance System Based on Driver Characteristics," in IEEE Transactions on Intelligent Transportation Systems, vol. 14, no. 1, pp. 1-12 (2013).
- [11] Das, Pragyaditya, Pragadeesh, S.: A Microcontroller Based Car-Safety System: Implementing Drowsiness Detection and Vehicle- Vehicle Distance Detection In Parallel. International Journal of Scientific & Technology Research (2015).
- [12] Gopal P. Gawande, Shruti V. Gavhale, Irshad, A. Zariye, Sagar P., Ritpurkar.: Review of Speed Control and Automatic Braking System, International Journal of Engineering Research & Technology (IJERT) vol. 03(02) (2014)
- [13] Brooker, Graham.: Millimetre Wave Radar for Tracking and Imaging applications, 1st International Conference on Sensing Technology, Palmerston North, New Zealand (2005).

- [14] Merete Östmann, Lotta Jakobsson.: An Examination of Pre-crash Braking Influence on Occupant Crash Response using an Active Human Body Model , IRCOBI Conference (2016).
- [15] Richard R. Mudge, Alain Kornhauser, Matt Hardison.: Market Framework and Outlook for Automated Vehicle Systems, Society of Actuaries (2018).
- [16] Syed Ali Fathima, S. J., Ajith Kumar, M., Bhuvaneshneshwar, Arunsunai Bala Kumaresan, P.: IoT Based Horn Detection System for Safety Four Wheel Driving, International Research Journal of Engineering and Technology, Vol. 5(5), pp 2359 – 2361 (2018).

Comparative Study on Generative Adversarial Networks in Text to Image Generation

¹Syed Ali Fathima S J, ²Surya Kangeyan K G, ³Lokeshwaran N G, ⁴Kavin Pradeep R

¹ Assistant Professor, Department of CSE, Kumaraguru College of Technology, Coimbatore
^{2,3,4} UG Student, Department of CSE, Kumaraguru College of Technology, Coimbatore
¹syedalifathima.sj.cse@kct.ac.in, ²suryakangeyan.18cs@kct.ac.in, ³lokeshwaran.18cs@kct.ac.in,
⁴kavinpradeep.18cs@kct.ac.in

Abstract. Viewing an item gives us more knowledge about that item compared to listening to someone saying what that item looks like. For example, seeing an image makes us easily understand what that image is, rather than hearing someone describing the image. Many have this problem of visualizing things from the explanations they hear about. Thus, the text to image generation problem is necessary because it has significant applications in CAD, art generation etc and a challenging one since the image should be realistic and consistent with the text. Bridging the semantic gap between text and image domain is a difficult and challenging one. Thanks to the success of Generative Adversarial Network (GAN) Systems, which achieved high accuracy in generation of superior quality images that are meaningfully in consistent with text. The goal of this survey is to review some of the existing GAN models such as VanillaGAN, StackGAN, AttnGAN, MirrorGAN, LeicaGAN and DF-GAN and their approaches and technique followed to tackle the challenge and evaluate their performance using complexity metric such as Inception Score which tells us how realistic the output is produced by the GAN. It is evident that inception score and human evaluation coincide with each other.

Keywords: GAN, StackGAN, MirrorGAN, AttnGAN, LeicaGAN, DF-GAN, Inception Score, Generator, Discriminator.

1 Introduction

With the upsurge of AI and deep learning, NLP and CV have high accuracy in results and become good exploratory fields. The text to image generation is a basic problem which seeks to produce a visually convincing image that is meaningfully consistent with the text [24], for example refer Fig. 1[14]. This problem requires processing fuzzy and incomplete information in natural language processing. This problem has many significant applications such as editing a photo, generation of art and CAD etc. [18]



Fig. 1. Bird with black wings and white belly [14]

Existing methods address this issue by combination of search and supervised learning process. To link explanations to pictures, the similarity among keywords and pictures that defined descriptive and text units is used. These units then look for most expected text-conditioned images pieces ultimately improving the text-conditioned image structure. The main drawback in this conventional approach is that it lacks the capacity to produce new picture content. Recently Text to image generation has remarkable progress by Generative Adversarial Network (GAN) system which succeed in generating photo realistic images that is semantically consistent with the text aligned [19]. However, there are many issues like generating high quality images, loss of fine-grained information of text, and so on. GAN uses various approaches and techniques to solve these problems. The goal of this paper is to review the existing models with the technique used to solve the challenges and drawbacks of those approaches.

1.1 Why GAN ?

At first, text to image generation was performed usually by a supervised learning process. The main disadvantage of the conventional text-to-image generation method is that they were lacking the capacity of producing new picture content. The variational autoencoders have the capacity of both compressing the data and synthesizing them. But when compared to GAN [22], images produced by VAEs are more blurred. But GAN has the capacity of generating images similar to the original image.[21] GAN looks more into the details of the data and so it can easily interpret various versions of data.[16] By using GAN, we can generate images which are more clear and almost accurate.[20]

2 Literature Review

2.1 Vanilla GAN

Vanilla GAN [24] is a simple GAN which consists of both Generator and Discriminator. In our articulation instead of giving noise vector as input, noise vector is given in concatenation with text embedding. The given text is converted to numbers because every single machine learning and deep learning models deals with numbers and word embedding (i.e., the vector representation of particular word) that is one of the most popular representation of document vocabulary. There are various algorithms like Word2Vec(W2V) [27], Average W2V, Average TF-IDF [28] available to convert word to word embedding.

How does the word is converted to vector?

There are three big families to convert words to vectors. They are statistical methods, predictive methods and combining both statistical and predictive methods.

Statistical

In the statistical method, a co-occurrence matrix is created. A K-sized window is set. Then look through the text with count i.e., the number of times pair of words occur together.

Example: Consider two sentences “The fast cat wears no hat” and “The cat in the hat ran fast”

Table 2. Co-occurrence Matrix

	cat	fast	hat	in	no	ran	the	wears
cat	0							
fast	2	0						
hat	2	2	0					
in	1	1	1	0				
no	1	1	1	0	0			
ran	1	1	1	1	1	0		
the	3	3	3	2	1	1	1	
wears	1	1	1	0	1	0	1	0

The co-occurrence number solely is not a good measure to calculate the co-occurrence probability as it doesn't take the number of times each pair occur together into account.

Consider the words ‘the’ and ‘cat’, their co-occurrence is high, but that is mainly because of a greater number of ‘the’ in the sentence. To solve this problem, Pointwise Mutual Information(PMI) [28] is used to estimate the probabilities from co-occurrences.

$$PMI(w,c)=\log \hat{p}(w,c)/\hat{p}(w) \hat{p}(c) \quad (1)$$

Predictive

In predictive methods, we will be training machine learning algorithms so that the predictions can be made using the words and their context. Sometimes we will be using neural networks which follow two techniques namely Continuous Bag of Words (CBOW) and Skip Gram.

If two methods provide good results the combination of them will always provide a better result, that means training a model and getting the word vectors from weights of it, but instead of using a sliding window to get contexts, train using the co-occurrence matrix. Finally, the text is converted into 256-dimensional vector and concatenated with 100-dimensional noise vector. This expression will be helpful for generators to produce images that are consistent with given text semantically.

On other hand, the discriminator will distinguish real images. In our articulation, instead of having real and fake image as input, a pair of images with text embedding will be given as input since the discriminator has additional responsibility, besides finding whether the generated image is real or fake, the discriminator should predict the likelihood of whether the generated image is semantically consistent with the given text.

2.2 Stack GAN

In semblance with painters, how they draw, Stacked Generative Adversarial Networks break down the complication of text to image synthesis into sub-problems. Stack GAN consists of two separate stages in which the stage I produces low resolution image with the input text embedding. By conditioning stage 1 result and text embedding to stage 2, it gains knowledge of how to seize the text information which is excluded by stage 1 [23]. The important reason why the second stage GAN is able to produce high resolution image is because the support for the model distribution produced from the coarsely aligned minimal resolution image has a cut above probability of intersecting with the calculated support of image resolution[30].

Stage-I:

According to [7], the given text is converted to text embeddings, and they are conditionally augmented which is given as an input to a fully connected layer in Generator to produce N_1 dimensional text conditioned vector c_0 which later sampled from Gaussian distribution and this N_1 dimensional conditioned vector is chained with N_2 dimensional noise vector for generation of image by an up-sampling block series. And for discriminator the text vector is squeezed into N dimension with help of fully connected layer and is reproduced to the form $M \times M \times N$, In the meantime the pair of real and generated images are down sampled with a down sampling block sequence till it has $M \times M$ dimension. the image is then chained with word embedding and finally the fully connected layer with one node would produce a decision score.

Stage-II:

Same as Stage I, Stage II network will be fed with word embedding and stage I resulted image. Stage II GAN will spatially replicate the word embedding to the form $M \times M \times N$ dimension matrix. In the meantime stage I resulted image is down sampled to the form $M \times M$ with series of down sampling blocks which then both image and text embedding will be coupled and then features of the encoded image with the features of the text are given to residual blocks, which is helpful in understanding multi-modal representation of the features of the input text and image. At the end, with sequence of up sampling blocks the image will be produced. The discriminator role will be similar to stage I.

2.3 AttnGAN

In the most recently proposed Generative Adversarial Network systems for T2I problem. The approach commonly used is to convert the whole sentence into text vector. Though these models give very good results, conditioning on whole text vector alone lacks important fine grained word level information. To address this issue,[3] propose an Attentional GAN which is multistage refinement for text to image generation.

Rather than producing image with global vector, AttnGAN generates low resolution image with global sentence vector and Attn model in every stage word level vector is combined with global vector forming multi model conditioned vector. New image generation relies upon the multi modal conditioned vector. Thus, AttnGAN begins with low resolution image and then subsequently upgrades it over various steps to output the final image without lack of fine-grained information hidden in the individual words. AttnGAN contains another novel component known DAMSM (Deep Attentional Multimodal Similarity Model) [17].

DAMSM consists of 2 deep neural-network models that estimate the similarity between image and sentence at word level. The text encoder which is a bi-directional LSTM (Long Short-Term Memory) [4] gives out the vectors from whole sentence description. $w = R^{N_1 \times N_2}$ is the indication of the feature matrix of the words where N_1 - the dimension of word vector, N_2 - number of words and w_i represents the vector of the i^{th} word. The last layer of LSTM is concatenated with a global sentence vector. The next one is image encoder which is a CNN model [15] that depicts images to its feature vectors. The features of different sub regions of the image are learnt by intermediate layers of CNN, while whole sentence features are learnt from the last layer [19]. The model built this image encoder upon Inception-v3 model [5] already trained on Image Net [6]. Then the local feature matrix is get extracted which consists of local features of all subregions of the image. In the meantime, the last layer of inception-v3(Average pooling layer) gives whole sentence feature vector.

Then the image features are transformed to regular semantic space of text features.

The matching score of the image-text pair is driven by attention is modelled to evaluate the correlation of the image sentence pair which also depends on the attention model among image and text. Thus, DAMSM play vital role in AttnGAN by providing DAMSM loss which is finally grained text to image matching loss.

2.4 Mirror GAN

Modelling semantic consistency within the image and text is made difficult by the domain gap between them. However [3] the attention mechanism addresses this problem by guiding generators with word level features to focus on generating different sub regions in each stage, using features for each word alone does not ensure the consistency of the whole sentence due to the heterogeneity between images and text. Text to image generation (T2I) can be regarded as a converse challenge of captioning of image [8,9] which is image to text generation(I2T) that generates text description of the given image. According to [10], Modelling and aligning the underlying semantics in both image and text is required to address each task. I2T redescription should have the same semantics if the image generated from the text is consistent with the text. Thus, the author proposes a novel text-to-image- and-image-to-text model called MirrorGAN which makes use of the objective of learning text to image generation by its redescription.

MirrorGAN comprises of three novel components STEM, GLAM and STREAM. STEM (Semantic Text Embedding Module) creates word-level features for each word and sentence level features for the given text which fed to GLAM Recurrent Neural Network (RNN) [11,18] is used to extract semantic embedding from the sentence which consists of word level embeddings and sentence embedding. GLAM (Global-Local Collaborative Attentive Module) is a multistage cascaded architecture which generates target images from the text embedding. Finally, STREAM tries to reproduce the text description from the image generated, which basically use encoder-decoder based image captioning framework [8,9].

2.5 Leica Gan

T2I is a highly difficult problem because bridging the gap between the text and image domain is a highly challenging one. Humans cleverly face this issue by learning from objects to form a clear idea about meaning, appearance, layout, colours, and shapes, from the given description. We generally create a mental image of the picture that is highly consistent with the given textual data and the picture is drawn by adding more details as we progress.

According to [12], a creative Text to Image model called Leica GAN was proposed which combines the three phases which are the Learning phase, Imagination phase, Creating/Drawing phase in a combined framework. Firstly, the learning phase of multiple priors is formulated from (TVE) textual-visual co-embedding which is comprised of a text image encoder which is used for learning the priors such as semantics, textures, colors in conjunction with a text-mask encoder for learning the priors such as shapes and layouts. Secondly, the imagination phase is formulated from multiple priors' aggregation where the previous priors are combined and also added with noise for diversity in generation of images. A module which generated attention features is designed which produces two attentive word- and sentence-context features which fuses the two pairs of textual features, with the visual features of the image which is generated previously. At last, the creation or drawing phase is formulated using the (CAG) cascaded attentive generator where the image is produced from coarse/rough to a fine manner.

2.6 DFGAN

Although the previous models have produced notable and impressive results, there are some drawbacks that come along with them. Firstly, in the stacked architecture of the previous models there exists entanglements between the generators which make the final refined image. It Looks like mixture a of simple fuzzy shapes with a few subtle details. Secondly, for ensuring text-image semantic consistency extra networks have been introduced which increases the training complexity and increases the computational cost. Thirdly, simple concatenation of text and image features doesn't completely use the textual information and the cross-modal attention in the previous models can be applied only twice on 64x64 and 128x128 image features because of large computational cost. This leads to the ineffective fusion of textual and image features.

Solutions Proposed in DFGAN [13]:

1. To solve the entanglements of generators in the stacked architecture.
The idea of replacing stacked backbone architecture with one-staged backbone which is composed of hinge loss and residual networks.
2. The second issue (text-image semantic consistency) is solved by a novel target aware discriminator which is composed of MA-GP and a single way output. By pushing the original image and matching in input pairs to a stage at which the discriminator's loss function gradient are zero, the generator will later be elevated to produce more naturalistic and semantically consistent images. MA-GP can also stabilize the training process and simultaneously ensures that text and image are consistent with each other. The previous model's two-way output is not suitable for maintaining the text image consistency as it slows down the intersection process of the generator. It is also solved by replacing it with a One-way output. It can improve the efficacy of MA-GP and speeds up the intersection process of the generator. Without the introduction of extra networks, DFGAN can synthesize high quality images which are proved to be image-text meaningfully consistent.
3. For simple text-image feature concatenation problems a deep text-image fusion block (DFBlock) is used which fuses the text information into image features. The DFBlock is made up of several Affine networks which is a lightweight module that only manipulates the image/visual features by channel-wise scaling and finally shifting operations. Thus, stacking of various DFBlocks on all image scales intensifies the text-image fusion process and elevates the performance of the generator.

Then the concatenation of the noise vector with the sentence vector and applying truncation operation on the noise converts a single discrete sentence vector to a series of sentence vectors thus providing the generator a stable and continuous text latent space.

3 Complexity Metrics

There are several ways to evaluate the Generative Adversarial Networks (GAN). Inception Score (IS), Frechet Inception Distance (FID) and so on. All the models are evaluated using inception score.

Table 2. Inception score of various models

Model	Inception Score
VanillaGAN	2.88
StackGAN	3.70
StackGAN++	3.82
AttnGAN	4.58
MirrorGAN	4.56
LeicaGAN	4.62
DFGAN	4.86

The Inception Score[29] is a standard for the evaluation of the quality and the diversity of synthetically produced images by GAN. It was proposed by Tim Salimans [1]. It involves usage of a pre-trained deep learning neural network model (inception v3 model) to separate the generated images [2,25]. A large number

of images are classified using the inception model. In this model the probability of images belonging to each class is predicted. Those predictions are finally encapsulated into a metric called inception score. Two properties are captured in the inception score quality and diversity. The lowest value of the inception score is 1.0 and the highest value is almost as same as the maximum number of classes. The model supports 1,000 classes of the ILSVRC 2012 dataset, so the maximum inception score for this dataset is 1,000. The inception score of the CIFAR-10 dataset which contains 10 classes of 50,000 images is 8.09 +/- .07 when generating synthetic images for this dataset.

4 Conclusion

In this work various Generative Adversarial Networks such as VanillaGAN, StackGAN, AttnGAN, MirrorGAN, LeicaGAN, DFGAN are discussed. Model architecture of various GANs is understood, this work gives us the basic brief understanding of how various GAN networks work and their underlying model architecture, and the accuracy of the output images generated by them. Analysing various GAN models leads us to the conclusion that DFGAN generates images which has high fidelity and is semantically consistent with the textual inputs.

References

- [1] Tim Salimans., Ian Goodfellow., Wojciech Zaremba., Vicki Cheung., Alec Radford., Xi Chen.: Improved Techniques for Training GANs (2016).
- [2] Jose Camacho-Collados., Mohammad Taher Pilehvar.: From Word To Sense Embeddings: A Survey on Vector Representations of Meaning. *Journal of Artificial Intelligence Research* 63:743-788.
- [3] Schuster, M., Paliwal, K. K.: Bidirectional recurrent neural networks, In: *IEEE Transactions on Signal Processing*, vol. 45, no. 11, pp. 2673-2681, (1997).
- [4] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: *CVPR* (2016)
- [5] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, A., Berg, A. C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. In: *IJCV*, 115(3):211-252, (2015)
- [6] Zhang, H., Xu, T., Li, H., Zhang, S., Wang, X., Huang, X., Metaxas, D.: StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks. In: *CVPR*.
- [7] Karpathy, A., Fei-Fei, F.: Deep visual-semantic alignments for generating image descriptions. In: *The IEEE Conference on Computer Vision and Pattern Recognition CVPR* (2015).
- [8] Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: *The IEEE Conference on Computer Vision and Pattern Recognition CVPR* (2015).
- [9] Qiao, T., Zhang, J., Xu, D., Tao, D.: MirrorGAN: Learning Text-to-Image Generation by Redescription. In: *CVPR*
- [10] Cho, K., van Merriënboer, B., Gulcehre, C., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using rnn encoder-decoder for statistical machine translation. In: *Proceedings of conference on Empirical Methods on Natural Language Processing EMNLP* (2014)
- [11] Qiao, T., Zhang, J., Xu, D., Tao, D.: Learn, Imagine and Create: Text-to-Image Generation from Prior Knowledge. In: *NeurIPS* (2019).
- [12] Tao, M., Tang, H., Wu, S., Sebe, N., Jing, X., Wu, F., Bao, B.: Deep Fusion Generative Adversarial Networks for Text-to-Image Synthesis. (2021)
- [13] Xu, T., Zhang, P., Huang, Q., Zhang, H., Gan, Z., Huang, X., & He, X.: AttnGAN: Fine-Grained Text to Image Generation with Attentional Generative Adversarial Networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1316-1324 (2018).
- [14] S. Albawi., T. A. Mohammed and S. Al-Zawi.: "Understanding of a convolutional neural network," 2017 International Conference on Engineering and Technology (ICET), pp. 1-6, (2017).
- [15] Alec, R., Luke, M., Soumith, C.: Unsupervised representation learning with deep Convolutional Generative adversarial networks. *arXiv:1511.06434 [cs.LG]* (2015).
- [16] Linyan, L., Yu, S., Fuyuan, H., Tao, Z., Xuefeng, X., Jinchang, R.: Text to Realistic Image Generation with Attentional Concatenation Generative Adversarial Networks, *Discrete Dynamics in Nature and Society*, vol. 2020, Article ID 6452536, 10 pages, (2020).
- [17] Daniel Jiwoong, I., Chris Dongjoo, K., Hui, J., Roland, M.: Generating images with recurrent adversarial networks. *arXiv:1602.05110 [cs.LG]* (2016)
- [18] Shakir, M., Balaji, L.: Learning in Implicit Generative Models. *arXiv:1610.03483 [stat.ML]* (2016)
- [19] Mehdi, M., Simon, O.: Conditional Generative adversarial Nets, In: *International Conference on Machine Learning*. (2014)
- [20] Tero, K., Timo, A., Samuli, L., Jaakko, L.: Progressive Growing of GANs for Improved Quality, Stability, and Variation, *International Conference on Learning Representations* (2018)

- [21] Zhengwei,W., Qi ,S., Tomas ,E. Ward.,: Generative Adversarial Networks in Computer Vision: A Survey and Taxonomy, In: arXiv:1710.10196 [cs.NE].
- [22] Diederik P ,K., Max ,W., arXiv:1312.6114 [stat.ML],(2013)
- [23] Ian J. Goodfellow, Pouget-Abadie,J., Mehdi ,M., Bing ,X., David Warde-,F., Sherjil ,O., Aaron ,C., Yoshua ,B.,: Generative Adversarial Nets” arXiv:1406.2661v1 [stat.ML] (2014)
- [24] Yu ,Z., Huchuan ,L., Ali ,B.,:Statistics of Deep Generated Images.(2019)
- [25] Tomas, M., Kai, C., Greg, C., Jeffrey, D.: Efficient Estimation of Word Representations in Vector Space, arXiv:1301.3781v3 (2013)
- [26] Sammut, C., Webb, G.I.: Encyclopedia of Machine Learning. Springer, Boston, (2011)
- [27] Soleimani, B. H., Matwin, S.: Fast PMI-Based Word Embedding with Efficient Use of Unobserved Patterns. Proceedings of the AAAI Conference on Artificial Intelligence ,33(01),7031-7038, (2019).
- [28] Barratt, S., Sharma, R.: A note on the inception score. In: ArXiv preprint arXiv:1801.01973, (2018) Han, Z., Tao, X., Hongsheng, L.: StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks (2017).

Exploration of Artificial Intelligence's Application in Software Engineering: Current and Future Roadmap

¹Isha Sharma, ²Pardeep Singh, ³Rahul Kumar Singh and ⁴Alankrita Aggarwal

¹ Computer Science Engineering, Chandigarh University, Ajitgarh, Punjab 140413, India

^{2,3} School of Computer Science, University of Petroleum and Energy Studies, Dehradun, Uttarakhand
248001, India

⁴ Department of Computer Science, PIET, Panipat, Haryana, India
rahulsinghcse25@gmail.com

Abstract: To improve the development process of software in the current scenario there is a need for applying Artificial intelligence techniques. There are number of techniques that can be used and advocated by researchers like knowledge-based systems, fuzzy logic, neural networks, and data mining. Software, because of its dynamic nature in the context of operation, needs improvement in the quality. Traditionally the expertise of the developer, with the same kind of projects and his knowledge is the parameter for improving the quality but the dynamic nature and the volatile environment today stresses upon in exploring the potential of AI for making all the improvements in the development phases of software being developed. The objective of this paper is to bring out the application of AI in software engineering and suggesting the best method of AI that will help in the overall improvement in software development phases. This paper covers the various AI methods like neural networks, genetic algorithms, fuzzy logic, and natural language processing techniques for software engineering and provides an organized survey of current and future aspects of AI in software engineering.

Keywords— Artificial intelligence, Metaheuristics Software Engineering, Software Risk Analysis .

1. Introduction

The software engineering industry is witnessing a software crisis since a half-century ago. A large number of methodologies, techniques, and environments have been unitized and implemented to solve the software crisis issue. Though the changing scenario of the market and the complex nature of the software to be developed along with the dynamic state of software had hit hard in the software industry and is halting the progress. This changing scenario is leading to more chances of software failures. Various software fails before the actual development work starts. This makes the forecasting of the estimated development time or the delivery time taken by the project a difficult step. Sometimes there might be the case that due to changing constraints about the software, the deadlines need to be shifted and thus impacting the overall cost by increasing the budget for the development [1]. This adds more risk to the risk analysis module that needs to be used before the start of the software project to predict the delivery dates.

Researchers have seen a perceptible rise in the area of Artificial Intelligence (AI) that enforces and recommends a change in software engineering [2]. The research paper is an endeavor that focuses on the applicability of AI in improving software engineering. An attempt has been made to look for the perspective in using AI methods for the development of software projects that are more prone to risks to save them for immature fallouts due to changing market scenarios.

This review article differs from the previous article in such aspects a (1) classified the current studies based on various AI applications utilized in the software development life cycle. (2) the article emphasized the use of AI applications in the different phases of SDLC. (3) summarized the available and used techniques of AI in software development.

The major objective of this survey paper is to understand the use of AI methods in the field of software development life cycle and how can we enhance the current system. The paper also brought out the future prospects of using AI in software development. The structure of the paper as follows: Section 2 defines the project planning based on AI. Requirement specification and software design based on AI is described in section 3. Software coding and testing phrase based AI techniques are defined in section 4. Section 5 summarizes the AI techniques used in SDLC. At last, conclude the article in section 6.

2. Artificial Intelligence Based Project Planning

Before the actual development of the project, there are number of factors that need to be planned off before time. Planning these in advance is required for the success in development of software projects. These factors not only includes planning the project completion time well in advance and total cost for the project, but it also involves the staffing requirements and dependencies between various activities in the project. These are mere estimates that will be guiding the correct path to be followed in the project development.

The research is now focusing on the application of Artificial intelligence in the area of software engineering. Different AI approaches have been proposed that allow the use of AI in various phase of software development life cycle model. Fig.1. demonstrates the various AI techniques like genetic algorithms, case based reasoning, knowledge based systems, and neural networks in the planning of the software project and its usefulness.

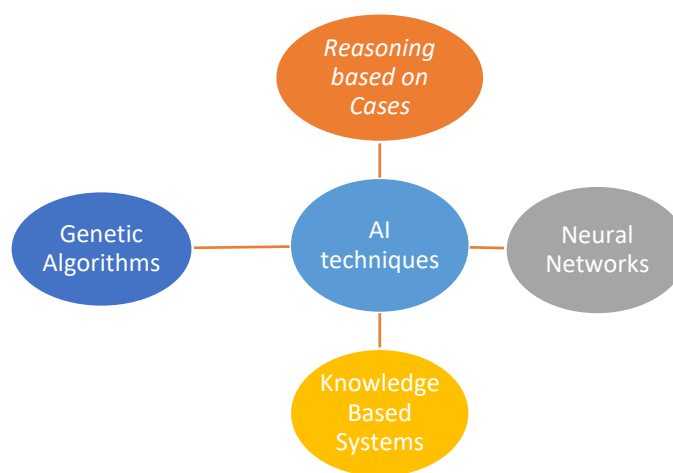


Fig. 1. AI techniques in the planning of the software project

2.1. Genetic Algorithms

The researches in the application of Genetic algorithms (GA) to scheduling the project activities are very large in number in various areas [3] [4] [5] [6]. An investigation of their appliance in manufacturing and operations management can be found in [7] [8]. In GA terminology the project plan is viewed as a constraint satisfaction problem with an objective. The objective is to optimized and transformed into the optimal problem with the aid of GA.

In [9], the author proposed evolving functions using a Genetic Algorithm for calculating the total effort required to develop the software project well in advance time. A wide range of mathematical functions and control statements are used that are applied on target grammars. As per the statistics, more than 400 software projects are used as training set data that is tested for more than 30 features. The major features are being tested using the approach for example the requirements for the projects, changing market scenario, type of application. The output is collected and compared with the other approaches and was found to be better. The GA approach uses non-linear team size as compared to traditional estimates used for calculating total effort for the software project.

Assigning the staff to activities is crucial for any software project that had a direct impact on the schedule of the activities too. In [10], the author also uses GA to produce optimal time and effort estimates. The research proposes chromosomal representation for assigning people to the activities and maintaining the precedence relationship among the activities so that the child produced after cross-over operation ensures a valid timeline. The outcome of this approach validates the usage of GA in the field of software development and more in software planning. A future proposal of finding the usage of GA when the project deadlines got shifted due to uncertain or certain reasons is still required.

It has been observed that while developing the software projects, quality's focussed more as a result improving the quality of project might result in increased cost of the project and also changing the completion time. This further increases user satisfaction. It has been also observed that higher quality is also achievable by adding more resources to the project, thereby helping in decreased overall cost though the activity level cost is high and early development of the project.

In [11], the author advocated the application of GA to deal with these variations. In their experiment, the crossover and mutation operations are used to represent the schedule of the project activity. A fitness function is there that considers both cost and time factors. The experiment is implemented on a large number of projects having numbers of activities not more than 30. The experiment favored the use of GA. In [12], the author suggested a Meta-Heuristic Approach for Software Project Risk Schedule Analysis using GA. The research illustrates a technique to shorten the time duration using the structured method.

2.2 Reasoning based on Cases

Experiences from the past of the development of projects are one of the factors that affect the success of the software projects. The more bad experiences the developer has gained, the more chances is there that there is less probability of repeating the same experience in developing similar projects.

A less number of researches are there that support the use of CASE-BASED REASONING (CBR) for planning the development of the software. In [13], the author proposed the use of CBR in planning the software project. HCA(Hierarchical Criteria Architecture) is used for representing the projects in terms of cases where the case is describing the project in terms of requirements of the client, resources that the project needs. HCA uses varying weights allowing variations to be tested and provide more flexibility to the project team leader. With the new project, similar cases are extracted using data mining. This aids project planning. The author [14] used more than 40 projects as an input set to show the applicability of CBR and data mining to help in finding out optimal total completion time of the project with more probability percentage.

2.3 Knowledge Based Systems

Experience in working with similar projects in past helps in planning the new projects optimally. Several studies have been made that support this analogy and use this experience in Knowledge based Systems (KBS). In [15], the author supported the use of KBS in the planning of the projects by providing a well-defined representation for the same. They proposed new semantics to support successful project planning. A frame-based language is used for developing this representation. The research supports the use of the language for characterizing the project milestones, targets, tasks, and duration. Other researches point out the usage of associative networks and production rules supporting KBS approach to be implemented [16]. However, the adoption of KBS for representing knowledge is a costly affair. In [17], the author focused on using an automated Software Modelling information software tool that will assist in project planning in easy manner using the concept of expert systems.

2.4 Neural Networks

Neural networks have been extensively and effectively used for problems that require classification. Neural networks are found to be perfect for the software engineering domain thereby dealing with the situations where there is a requirement of making a prediction for example about the risks in the maintenance module [18], analyzing the risks [19] and for fault prediction using matrices [20]. In [21], the authors proposed new research in this domain. They used the approach of identifying the risk assessment features based on old classifications by taking the input by interviewing the project managers. A total of 39 risk factors are found in the study that is further classified into 5 risk categories. The area of software engineering, project management, teamwork, cooperation, and project complexity is found to be more risky in nature and the classification of risk is made according to these parameters. Further using the PCA (Principal components analysis) approach, the 39 risk factors are dependent on each other are reduced to 19 independent factors. As per the study made, and experiments are performed, the Neural network approach was found to be the best approach for forecasting the risk. A total of 50 examples are used for this study. 35 examples out of the total 50 examples are used for training and 15 examples are used for testing purposes. A genetic algorithm trained Neural Network was used for the prediction and the result are grouped under successful projects, partially failed projects, and failed projects. Although the confidence in the research that Neural Network is better the approach was made on the small data set and more number of projects needed to be tested to have permanent confidence in the research made that calls for future extension in the research work made.

3. Requirements Engineering And Software Design

Requirements engineering is the first phase of any software development life cycle model. This phase is part of every life cycle model and is the most important phase too. The phase is actually not a single step of

execution but has multiple areas defined in it that seek attention [22]. Requirement engineering helps the companies in defining the requirements of the clients but due to different perceptions of different people involved in software development like company members and clients, there is need to express requirements in such a language that is natural in understanding and need to be documented [23]. As we already know the requirement engineering phase is not single activity rather it involves sub-activities like gathering of the needs for the project and analysis of the same. The valid document achieved after requirement engineering is the valid set of requirements that have been added in the contract between the client and the company as the requirements to be achieved. The requirement specification document is then further formulated to be base for designing the software. Though it sounds so easy that gathering the requirements is not a tedious task but it is actually the most important and deciding task in software development that form the basis for effective design. Therefore it is essential to look for the gaps in the requirement gathering and analysis step do that before actual implementation of the system in form of designing it , the same will be resolved and not carried forward thereby making a loop of problems and errors in the software process. The following gaps are identified during this phase:

- The natural language expression of the requirements are often found to be ambiguous thereby making the requirements invalid and inconsistent. [33]
- Requirements are mostly the confusing needs of the customers that is maximum times not sure about what they demand from the software process being developed d by the organization .this incompleteness and fuzzy nature of the customer need calls for another gap in requirement engineering.
- Requirement are dynamic in nature and hence keeps on changing and improving over the time and time need to be devoted to incorporate them in the software system
- The most exigent task in this phase is perceptive the requirements of the customer.
- The management of requirements is also a difficult task. The success of this task will further decide the success of other recurring tasks that follows requirement engineering.

Artificial intelligence is involved in this phase of engineering requirements in software projects as follows:

- By developing the tools that are able to extract and transform natural language requirements in no ambiguous form.
- By providing knowledge-based systems and ontologies.
- Combining intelligence at computation level to deal with problems coming up with the requirement engineering phase

4. Software Coding, Testing And Artificial Intelligence

4.1 Coding

A number of AI techniques are there that will help in easing the task of coding for the real world software projects. With the help of AI it is possible to mechanize the entire process of software coding. The basic idea behind this mechanized task is to involve the programmer in developing the algorithmic solution for the problem, and then to use AI techniques that will generate automatic functions or the complete program as per the programmer specification. There are various AI methods that can be applied. The below is the list of such methods contributing computational intelligence in software domain:

- Object oriented method [24]
- Case based Reasoning analogies
- Constraints prior programming approach
- Genetic Algorithms
- Heuristic software engineering based on search optimization

In [25], the author devised Artificial Intelligence for semantic-based code searching in which the required code is extracted by querying the software system with the application of NLP. The AI based software supports a huge number of functions that will ease the querying based on natural language from the software

4.2 Testing

The important phase of developing the software is to check for the effective working by means of testing the software against anomalies. The task of testing the software being developed is a lengthy procedure. It will be taking more time if done manually by team of testers, therefore, is requiring the need of computerizing the

whole task of testing and there AI is an important component that will help in this domain. In 1991, DeMasie and Muratore has drafted the work in this direction by using the concept of constraint-based testing [26]. In [27], the authors suggested the use of knowledge based expert systems to automate the process of testing. In the mid of 1990, the concept of planning testing using AI has blossomed by the development of the planner supporting the generation of test cases and executing the same to find the loopholes in the software being developed. In [28], the author supported the use of genetic algorithms for simulating the task of software testing and supports the generation of optimal cases for testing.

In 2019-2020 [29] [30] recommended the use of Deep neural networks that will help in finding out the complex anomalies software code. Deep Neural Networks have the advantage of adaptation that will help the software development team in identifying the rare errors and thus improves the overall software quality. AI-based fuzzy systems are found to be far accurate than the other available hybrid or manual versions [31].

AI eases the process of software based testing and because of the automation profile, it is helping in saving the manpower hours in executing the tests and documentation overheads as well. The software development cost and the Time to market costs are also found to be reduced significantly [32].

5. Artificial Intelligence Techniques Used In Software Engineering

This section focuses on the available and used techniques from the AI domain that are applicable in software development. Fig. 2. demonstrates the AI techniques in SDLC.

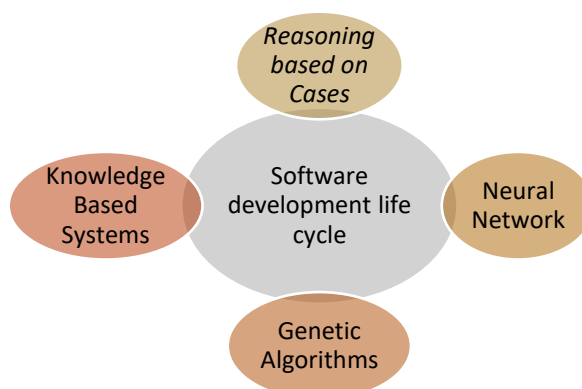


Fig.2. AI technique used in SDLC

Table 1 describes the available and used techniques of AI in the SDLC.

Table 1: AI techniques for software development

Techniques of AI	Where to apply in software engineering
Neural networks	Risk analysis, software Maintenance
Genetic Algorithms	Software testing
Fuzzy systems	Software requirements and engineering, software testing
Knowledge-based systems	Software design, planning
Natural Language processing	Requirement engineering
Expert system	Software risk management
Case based Reasoning analogies	Software coding
Software engineering based on search optimization	Software coding

6. Conclusion And Future Scope

This paper aims to find the application of artificial intelligence and software engineering. The introduction of AI was targeted in the area of software engineering to solve the real-time problems software development life cycle for developing the projects faced. The application of computational intelligence in the software domain has seen major decrease in the development time of the software development and decrease in cost for the production of the same and hence resulted in better quality projects.

References

- [1] T. L. Fox and J. W. Spence, "The effect of decision style on the use of a project management tool: an empirical laboratory study," *ACM SIGMIS Database: the DATABASE for Advances in Information Systems*, vol. 36, no. 2, pp. 28-42, 2005.
- [2] G. Booch, "Object-oriented development," *IEEE transactions on Software Engineering*, vol. 2, pp. 211-221, 1986.
- [3] J. Tewari, S. Arya and P. N. Singh, "Approach of intelligent software agents in future development," *Int. J. Adv. Res. Comput. Sci. Softw. Eng.* http://ijarcse.com/Before_August_2017/docs/papers/Volume_3/5_May2013/V3I4-0319.pdf, 2013.
- [4] H. H. Ammar, A. W and M. S. Hamdi, "Software engineering using artificial intelligence techniques: Current state and open problems," in *In Proceedings of the First Taibah University International Conference on Computing and Information Technology (ICCIT 2012), Al-Madinah Al-Munawwarah, Saudi Arabia (Vol. 52)*, 2012.
- [5] M. Harman, "The role of artificial intelligence in software engineering," in *In 2012 First International Workshop on Realizing AI Synergies in Software Engineering (RAISE) (pp. 1-6)*. IEEE., 2012.
- [6] P. Jain, "Interaction between software engineering and artificial intelligence-a review," *International Journal on Computer Science and Engineering*, vol. 3, no. 12, p. 3774, 2011.
- [7] F. Meziane and S. Vadera, "Artificial Intelligence Applications for Improved Software Engineering Development: New Prospects," *IGI Global*, 2009.
- [8] F. N. Raza, "Artificial intelligence techniques in software engineering (AITSE)," in *In International MultiConference of Engineers and Computer Scientists (IMECS 2009)*, 2009.
- [9] P. R. Srivastava and T. H. Kim, "Application of genetic algorithm in software testing," *International Journal of software Engineering and its Applications*, vol. 3, no. 4, pp. 87-96, 2009.
- [10] B. Hooshyar, A. Tahmani and M. Shenasa, "A Genetic Algorithm to Time-Cost Trade off in project scheduling," in *In 2008 IEEE Congress on Evolutionary Computation (IEEE World Congress on Computational Intelligence) (pp. 3081-3086)*. IEEE., 2008.
- [11] A. Bertolino, "Software testing research: Achievements, challenges, dreams," in *In Future of Software Engineering (FOSE'07) (pp. 85-103)*. IEEE., 2007.
- [12] I. Sharma and D. Chhabra, "Meta-Heuristic Approach for Software Project Risk Schedule Analysis," *In Analyzing the Role of Risk Mitigation and Monitoring in Software Development IGI Global*, pp. 136-149, 2018.
- [13] K. A. Kobbacy, S. Vadera and M. H. Rasmy, "AI and OR in management of operations: history and trends," *Journal of the Operational Research Society*, vol. 58, no. 1, pp. 10-28, 2007.
- [14] H. L. Yang and C. S. Wang, "Recommender system for software project planning one application of revised CBR algorithm," *Expert Systems with Applications*, vol. 36, no. 5, p. 8938-8945, 2009.
- [15] Y. R. Reddy, M. S. Fox, N. Husain and M. McRoberts, "KBS: a Knowledge Based Simulation System," *IEEE software*, vol. 3, no. 2, p. 26, 1985.
- [16] A. K. Shradhanand and D. S. Jain, "Use of fuzzy logic in software development," *Issues in information systems*, vol. 8, no. 2, pp. 238-244, 2007.
- [17] K. M. Aljebory and M. QaisIssam, "Developing AI Based Scheme for Project Planning by Expert Merging Revit and Primavera Software," in *In 2019 16th International Multi-Conference on Systems, Signals & Devices (SSD) (pp. 404-412)*. IEEE., 2019.
- [18] H. Yong, C. Juhua, R. Zhenbang, M. Liu and X. Kang, "A neural networks approach for software risk analysis," in *In Sixth IEEE International Conference on Data Mining-Workshops (ICDMW'06) (pp. 722-725)*. IEEE., 2006.
- [19] S. Wappler and J. Wegener, "Evolutionary unit testing of object-oriented software using strongly-typed genetic programming," in *In Proceedings of the 8th annual conference on Genetic and evolutionary computation (pp. 1925-1932)*, 2006.
- [20] Y. Ge and C. Chang, "Capability-based project scheduling with genetic algorithms," in *In 2006 International Conference on Computational Intelligence for Modelling Control and Automation and International Conference on Intelligent Agents web Technologies and International Commerce (CIMCA'06) (pp. 161-161)*. IEEE., 2006.
- [21] Y. Hu, J. Chen, Z. Rong, L. Mei and K. Xie, "A Neural Networks Approach for Software Risk Analysis," in *In Proceedings of the Sixth IEEE International Conference on Data Mining Workshops (pp. 722-725)*, Hong Kong. Washington DC: IEEE Computer Society., 2006.
- [22] L. C. Briand, Y. Labiche and M. Shousha, "Stress testing real-time systems with genetic algorithms," in *In*

- Proceedings of the 7th annual conference on Genetic and evolutionary computation (pp. 1021-1028).*, 2005.
- [23] A. Baresel, D. Binkley, M. Harman and B. Korel, "Evolutionary testing in the presence of loop-assigned flags: A testability transformation approach," *ACM SIGSOFT software engineering notes*, vol. 29, no. 4, pp. 108-118, 2004.
 - [24] M. Gupta, F. Bastani, L. Khan and I. L. Yen, "Automated test data generation using MEA-graph planning," in *In 16th IEEE International Conference on Tools with Artificial Intelligence (pp. 174-182). IEEE.*, 2004.
 - [25] H. Husain, H. H. Wu, T. Gazit, M. Allamanis and M. Brockschmidt, "Codesearchnet challenge: Evaluating the state of semantic code search," *arXiv preprint arXiv:1909.09436.*, 2019.
 - [26] M. P. DeMasie and J. F. Muratore, "Artificial intelligence and expert systems in-flight software testing," in *In Proceedings of the Tenth IEEE Conference on Digital Avionics Systems Conference (pp. 416-419), Los Angeles, CA. Washington DC: IEEE Compute*, 1991.
 - [27] C. A. Bering and M. W. Crawford, "Using an expert system to test a logistics information system," in *In Proceedings of the IEEE National Aerospace and Electronics Conference (pp. 1363-1368), Dayton, OH. Washington DC: IEEE Computer Society.*, 1988.
 - [28] L. Wallace and K. M, "Software project risks and their effect on outcomes," *Communications of the ACM*, vol. 47, no. 4, pp. 68-73, 2004.
 - [29] T. M. King, J. Arbon, D. Santiago, D. Adamo, W. Chin and R. Shanmugam, "AI for testing today and tomorrow: industry perspectives," in *In 2019 IEEE International Conference On Artificial Intelligence Testing (AITest) (pp. 81-88). IEEE.*, 2019.
 - [30] H. Liang, L. Jiang, L. Ai and J. Wei, "Sequence Directed Hybrid Fuzzing," in *In 2020 IEEE 27th International Conference on Software Analysis, Evolution and Reengineering (SANER) (pp. 127-137). IEEE.*, 2020.
 - [31] A. Takanen, J. D. Demott, C. Miller and A. Kettunen, "Fuzzing for software security testing and quality assurance," Artech House, 2018.
 - [32] H. Hourani, A. Hammad and M. Lafi, "The impact of artificial intelligence on software testing," in *In 2019 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT) (pp. 565-570). IEEE.*, 2019.

Landmark Recognition with Artificial Intelligence: A State of the Art

Kanishk Bansal¹, Amar Singh¹

¹ Department of Computer Applications, Lovely Professional University, Phagwara, Punjab, India - 144411

Abstract: Landmark Recognition from images is a challenging issue in the field of Computer Vision due to the large-sized and complex available datasets. In this article, a summary is presented for different AI (Artificial Intelligence) based approaches for Landmark Recognition. To understand the scope of improvement, a comparative analysis has been done of the different approaches and datasets available in the State of the Art. From this study, it is observed that the sizes and complexities of Landmark Recognition datasets are increasing with time. As the sizes are increasing, the accuracies of the available approaches are decreasing. In this paper, it was observed that up to now, Google Landmarks Dataset V2 is the largest available dataset and the SE ResNetXT152 approach achieved the best training and validation accuracy on this dataset as compared to the other approaches. In the Future Scope section of the article, it is proposed that one can optimize the performance of different AI-based approaches for Landmark Recognition by using Nature Inspired Computing (NIC) based Search and Optimization Approaches.

Keywords. Landmark Recognition, Computer Vision, Robotic Technologies, NIC, Optimization

1. Introduction

Artificial Intelligence is the branch of computer science which deals with making computers efficient so that they can make good, well-informed decisions on their own. An artificially intelligent computer system learns from the existing knowledge very fast and can take decisions (sometimes) better than humans even [1]. AI learns from the existing knowledge, builds models for a better understanding of a scenario, and delivers intelligent decisions. AI systems mostly work on statistical procedures to learn from the data provided to it in the form of knowledge, build models for a better understanding, and then deliver good, well-informed decisions based on the models formed from the data provided. For a good AI model, not only high-tech hardware is mostly required but better AI models too would have to be created. This requires both, good machinery and a good human brain for building the AI system. If developed aptly, AI models can outperform humans in delivering insightful answers and decisions.

Out of many uses of AI, one of the uses includes processing media also. Recently, AI has been used to identify objects from an image [2]. This area is known as *Image Processing by AI*. This area in AI is extremely important in the long run if we want machines to learn and understand the process of Visioning and comprehending what exists in the environment or even in a 2D image. In this, we can present an image to a computer system and ask it to identify what all exists in that image [3]. This field of *Intelligence Systems* is known as *Computer Vision*. An intelligent computer system can identify most of the objects in the image. A lot of research has been done and is ongoing in this field. AI systems can identify many things in photos, images, and even videos. These things include plants, fruits and vegetables, crops, animals, colors, and even human emotions.

Today, there exist many application software that can even process images and videos to an extent that the software recognizes any face in an image and uses it on some other face that existed in a video. This concept is known as *Video Morphing*. However, there are things which still most computer systems are not able to figure out. One of these is the important *Landmarks* in a picture which may include historical monuments, buildings, wonders of the world as well as simple immovable objects, etc. An AI system has been trained for single *Landmarks* but as dataset size increases, the AI systems also tend to suffer and fail.

However, *Landmark Recognition* is an important task in the field of *Computer Vision* because it is one of the first steps in enabling a full-fledged Computer Vision. A properly enabled AI system will be able to identify almost everything with a human-like precision or more. Work is ongoing to identify such *Landmarks* with high accuracy by a computer system [4]. However, we still have a great deal to accomplish. A dataset of *landmarks* can be extremely large. Therefore, training a computer system to learn and identify these

landmarks is a difficult task. A great deal of work has been already done in this area and a great deal more is required. Training accuracies of 90% and more have been achieved on small datasets but on larger ones, training accuracies don't cross 70% [5].

Fig. 1. Examples of Landmarks from Weyand, Tobias, et al. (2015) [9]



Google routinely announces new competitions to improve accuracies in this field [6]. However, as the datasets become more exhaustive, the out-of-the-sample accuracies are no more than 40%. We need much better techniques to handle such large datasets and improving *validation accuracy* i.e. accuracy for data outside training data, is needed to a very large extent [7]. The work in this field of Computer Vision is ongoing and a lot more is required yet. We, still, have a long way to reach an era when AI systems or robots will be able to identify everything in our surroundings just by an *Artificial Computer Vision*.

As we have seen, the first section of this paper introduces us to the subject of *Landmark Recognition by Artificial Intelligence*. In section 2, we have presented the literature related to *Landmark Recognition* and have compared various research articles based on the databases used in them, approaches used, achievements made, and how this subject has been used for robotic technologies. In section 3, we discuss the implementation of some classical *Machine Learning* approaches on some part of the Google Landmarks Dataset V2. In section 4, we describe how *Nature Inspired Computing* is being used to optimize classical approaches and how *Nature-Inspired Algorithms* have a lot of scope in *Artificial Intelligence*. In the last section, we conclude the article describing our findings and the future scope of the subject.

2. Survey of Literature

Work on Landmark Recognition is progressing, and a lot of success has been accomplished. The research articles like Ozaki, Kohei, and Shuhei Yokoo. (2019), have tried to better the accuracies for large datasets available for *Landmark Recognition* [8]. The dataset used in the article is noisy and diverse. The approach in this research article is based on *Deep Convolutional Neural Networks with Metric Learning*, trained using *cosine-softmax-based losses*. These methods are sensitive to noise, and this could hinder in developing a reliable metric. To address the issue, the authors have developed an *automated data cleaning system*.

Similarly, more works that are ongoing, and which are accomplished have been discussed in our article. The articles have been surveyed in terms of their datasets, approaches used, accomplishments, and the technologies for which when they arrived.

2.1. Survey by Trends in Datasets

Table 1. The table shows some research papers along with the datasets they used for landmark recognition and their accuracies.

S. No.	Year	Research Article	Authors	Dataset Used	Landmarks	Technique	Training Accuracy	Validation Accuracy
1	2019	Google Landmarks Dataset v2-A Large-Scale Benchmark for Instance-Level Recognition and Retrieval[9]	T. Weyand, A. Araujo, B. Cao, J. Sim	Google Landmarks Dataset v2	200000	ResNet101 + ArcFace	70%	40%
2	2019	2nd Place and 2nd Place Solution to Kaggle Landmark Recognition and Retrieval Competition 2019[6]	Chen, Kaibing, et al.	Google Landmarks Dataset v2	200000	SE ResNeXt152	81%	60%
3	2017	LargeScale Image Retrieval with Attentive Deep Local Features[10]	H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han	Google Landmarks	30000	DELf	81%	62%
4	2015	Visual landmark recognition from Internet photo collections: A large-scale evaluation[11]	T. Weyand and B. Leibe	Paris500k	13000	Voting	61%	43%
5	2016	Fine-tuning CNN Image Retrieval with No Human Annotation[12]	F. Radenovic, G. Tolias, and O. Chum.	Flickr-SfM	713	CNN	90%	65%
6	2016	Deep Image Retrieval: Learning Global Representations for Image Search[13]	A. Gordo, J. Almazan, J. Revaud, and D. Larus	Landmark URLs	586	Deep Nets	83%	63%
7	2018	Revisiting Oxford and Paris: Large-Scale Image Retrieval Benchmarking[14]	F. Radenovic, A. Iscen, G. Tolias, Y. Avrithis, and O. Chum	Revisited Oxford	11	HesAff+rSIFT-SMK	95%	82%
8	2015	24/7 place recognition by view synthesis[15]	A. Torii, R. Arandjelovic, J. Sivic, M. Okutomi, and T. Pajdla.	24/7 Tokyo	125	View Synthesis	85%	62%
9	2012	Worldwide Pose Estimation using 3D Point Clouds[16]	Y. Li, N. Snavely, D. Huttenlocher, and P. Fua	Landmarks-PointCloud	1k	3D Point Clouds	73%	45%
10	2011	Building Rome in a Day[17]	S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. Seitz, and R. Szeliski	San Francisco	15k	ANN SIFT	70%	44%
11	2010	City-Scale Landmark Identification on Mobile Devices[18]	D. Chen, G. Baatz, K. Koser, S. Tsai, R. Vedantham, T. Pytanainen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk	European Cities 50k	20	PFI Pipeline	95%	67%
12	2010	Feature Map Hashing: Sub-linear Indexing of Appearance and Global Geometry[19]	Y. Avrithis, G. Tolias, and Y. Kalantidis	Geotagged StreetView	---	FMH	---	---
13	2010	Avoiding Confusing Features in Place Recognition[20]	J. Knopp, J. Sivic, and T. Pajdla	Rome 16k	69	Confuser suppression + Query expansion	48%	---
14	2008	Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases[21]	J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman	Paris	11	Soft Assignment	72%	48%
15	2008	Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search[22]	H. Jegou, M. Douze, and C. Schmid	Holidays	500	HE + WGC	53%	40%
16	2015	Landmark recognition with sparse representation classification and extreme learning machine[23]	Cao, Jiwen, et al.	NTU Campus	1	Sparse Representation Classification	93%	83%
17	2019	Large scale landmark recognition via deep metric learning[24]	Boiarov, Andrei, and Eduard Tyantov	Scenes	4	Deep Metric Learning	90%	80%
18	2018	An accurate retrieval through R-MAC+ descriptors for landmark recognition[25]	Magliani, Federico, and Andrea Prati	Holidays	500	R-MAC+ with retrieval	91%	75%
19	2017	A location-aware embedding technique for accurate landmark recognition[26]	Magliani, Federico, Navid Mahmoudian Bidgoli, and Andrea Prati.	ZuBuD	460	locVLAD	86%	70%
20	2018	Efficient nearest neighbors search for large-scale landmark recognition[27]	Magliani, Federico, Tomaso Fontanini, and Andrea Prati	Holidays	500	BoI adaptive multi-probe LSH	86%	63%
21	2018	Landmark detection in 2D bioimages for geometric morphometrics: a multi-resolution tree-based approach[28]	Vandaele, Rémy, et al.	Geometric Morphometrics	20	Tree Based	82%	70%
22	2019	Fast Landmark Recognition in Photo Albums[29]	S. Pierre, and C. Oprean	Film Simulation HALDCLUT	---	---	---	---
23	2016	Bee-inspired landmark recognition in robotic navigation[30]	CA Kodi Cumbo, et al.	Robot Camera	---	---	---	---
24	2008	Multiscale Unsupervised Segmentation of SAR Imagery Using the Genetic Algorithm[31]	Xian Bin Wen, Hua Zhang and Ze Tao Jiang	---	---	Genetic Algorithm	93%	---
25	2008	Perceptual Image Segmentation Using Fuzzy-Based Hierarchical Algorithm and Its Application to Dermoscopy Images[32]	J. Maeda, A. Kawano, S. Yamauchi, Y. Suzuki A. R. S. Marcal and T. Mendonc	---	---	Fuzzy Algorithm	90%	---
26	2007	A QuantumInspired Evolutionary Algorithm for Multiobjective Image Segmentation[33]	Hichem Talbi, Mohamed Batouche and Amer Draa	---	---	---	---	---
27	1996	Neural-based color image segmentation and classification using self-organizing maps[34]	Jander Moreira and Luciano Da Fontoura Costa	---	---	---	---	---

In this section, we survey trends in features of available datasets in the field of Landmark Recognition. As we can see in Table 1, in the article, *Torii, Akihiko, et al. (2015)*, there was a dataset of just 125 landmarks and still, the accuracy was only 85% [15]. This is very less accurate considering the size of the dataset. In this

research paper, 125 landmarks were used although the images totaled up to 1,000 out of which 315 were used as test images. Though this dataset was very small, however, the dataset in it could be used at any time meaning that the dataset was stable. In contrast to this, *Weyand, Tobias, and Bastian Leibe. (2015)*, has a dataset of 501k images and landmarks amounting to 13k [11]. However, in this dataset accuracy was drastically reduced to just 61% and the dataset was unstable meaning that it could not be used for further uses. This article, published in 2015, showed that in that year neither the large datasets could be stable, nor could we get high accuracy on such dataset. Therefore, it was implicit that we required better AI models for a competent Computer Vision.

In *Gordo, Albert, et al. (2016)*, authors analyze how making a computer learn *Global Representations* could be used for image search [13]. An AI model can be trained using *Global Representations* by which they can learn how to distinguish different things from others. In this article, there were 586 landmarks and still, the accuracy was just about 83% on testing data. Then, *Radenović, Filip, Giorgos Tolias, and Ondřej Chum. (2018)* was published which came up with an accuracy of 90% on a dataset with 713 landmarks. Although the accuracy is considerable, however, in comparison to [11], the number of landmarks is too small and so was the dataset with just 120k images which included training and test images. *Noh, Hyeonwoo, et al. (2017)* is a paper published taking *Google dataset* as a training dataset [10]. This was an unstable dataset but a large one. It had about 30k landmarks and 1.2M images. This research article came up with an accuracy of approximately 81% which is not considerable but taking into consideration, the size of the dataset, this dataset was able to achieve some accomplishments at least. Then, in 2018, an article named *Radenović, Filip, et al. (2018)* was published [14]. This article had only 11 test landmarks and was able to achieve 95% accuracy. It was not a big achievement but was the article was very important because introduced new methods of benchmarking for areas like *Landmark Recognition*. This paper had 70 images and all of them were used as test images.

From Fig. 2, we observe that the trends show that the dataset sizes have, continuously, been increasing with each passing year. Though the datasets might not be robust, however, sizes are increasing which means variability in the data is increasing and thus machine learning algorithms are becoming more powerful with each passing year. The year 2019 shows a dataset of 200,000 landmarks which amounted to more than 4.1M images meaning that an average landmark class had over 200 images. Also, we observe that there were articles with very few landmarks that were not that relevant since dataset sizes have seen an increasing trend after any number of articles with small datasets.

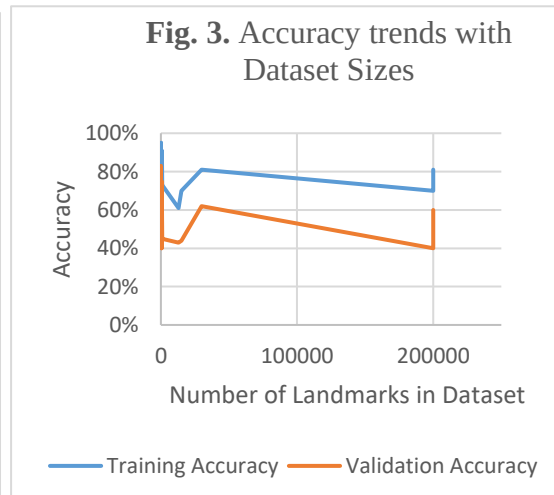
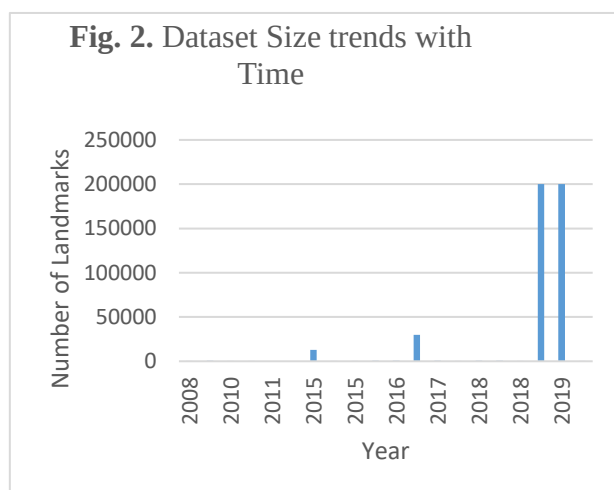


Fig. 3 shows the trends in accuracies achieved in datasets as their sizes increase. As we notice, when first, the size of datasets increases considerably, the accuracy also falls by considerable numbers. However, when years pass by and dataset size remains almost the same or does not increase a considerable level, the accuracy increases by substantial amounts. We can infer from the figure that though, accuracy decreases with drastically increasing dataset sizes, however, substantial accuracies are achieved as years pass by.

Table 2. Training and Validation Accuracies of various techniques on Google Landmarks Datasets [9]

Technique	Training Accuracy	Validation Accuracy
DELF-KD-tree	44.84	41.07
DELG global-only	32.37	32.02
DELG global+SP	56.35	55.01

Table 3. Training and Validation Accuracies on Google Landmarks Datasets V2 for various teams along with the techniques used by them. [9]

Team Name	Technique	Training	Validation
smlyaka	GF ensemble -> LF -> category filter	69.39	65.85
JL	GF ensemble -> LF -> non-landmark filter	66.53	61.86
GLRunner	GF -> non-landmark detector -> GF+classifier	53.08	52.07

Then it was the next article which introduced very large-scale and stable datasets in 2019. This article named *Weyand, Tobias, et al. (2019)* uses more than 4.1M images as training and test data [9]. Just like [10], the dataset used in this article is has been published by Google but unlike [13], the dataset was stable, meaning the data could be used for later uses. Also, the accuracy reduced considerably to 70%. This was the testing accuracy using training data itself, whereas validation accuracy was below 40% which is extremely low accuracy. Since then Google has regularly announced competitions to reward those who come up with better accuracies on such large datasets.

Fig. 4 and Fig. 5 show the training and validation accuracies achieved by various techniques and teams on *Google Landmarks Dataset* and *Google Landmarks Dataset V2* [9]. As we can see *Google Landmarks Dataset* had lower accuracies as compared to *Google Landmarks Dataset V2*, thereby meaning that, *Dataset V2* was more robust as compared to the previous version. Though the dataset size of *version 2* was more, getting better results meant that *AI techniques* bettered manifolds.

2.2. Survey by Approach

2.2.1. Survey for Landmark Recognition

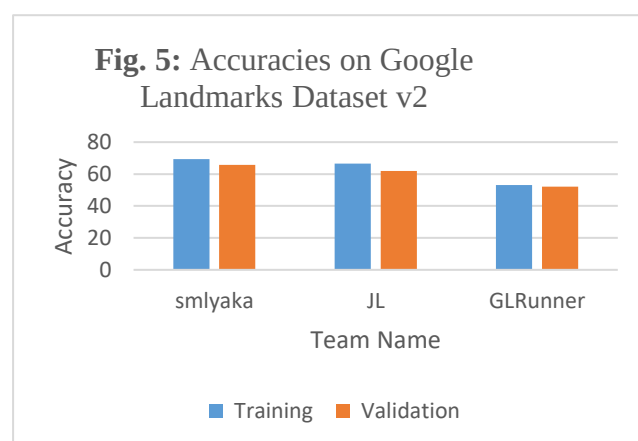
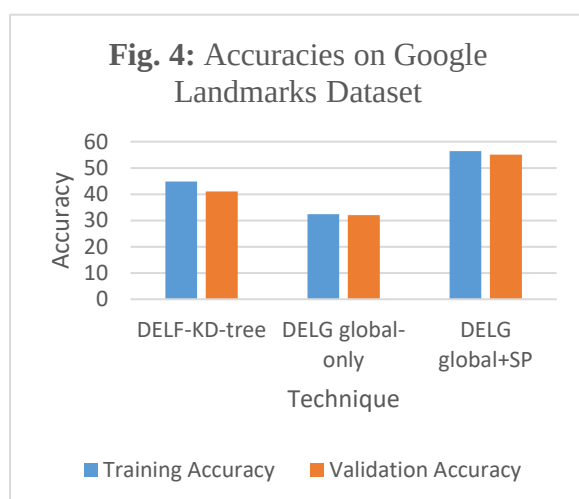


Table 4: Research articles with the need addressed in them and work accomplished

Research Article	Need Addressed	Work Accomplished
Babenko, Artem, et al. (2014) [35]	Image Recognition and Retrieval	Built Convolutional Neural Networks
Suau, Pablo. (2005) [36]	Robust Landmark Recognition	Developed Polar Histograms
Chen, Yuntao, et al. (2019) [37]	Deal Large Datasets	Design Distributed Frameworks
Crall, Jonathan P., et al. (2013) [38]	Instance Recognition	Neural Nets for Species Recognition
Scheirer, Walter J., et al. (2011) [39]	Assigning scores for Image Recognition	Theory and Practice Methods Discussed
Teichmann, Marvin, et al. (2019) [40]	Image Classification	Efficient Regional Aggregation
Zitová, Barbara, and Jan Flusser (1999) [41]	Better Landmark Recognition	Landmark Recognition using invariant features
Cao, Jiuwen, et al. (2015) [42]	Better Landmark Recognition	Extreme Learning Machines
Srinivasan, Sridhar, and Laveen Kanal. (1997) [43]	Qualitative Landmark Recognition	Visual cues for image recognition
Cao, Jiuwen, Tao Chen, and Jiayuan Fan. (2014) [44]	Feature Extraction for Landmark Recognition	SLFNs
Chen, Tao, and Kim-Hui Yap. (2013) [45]	Qualitative Landmark Recognition	Mobile Landmark Recognition
Chen, Tao, Kim-Hui Yap, and Dajiang Zhang. (2012) [46]	Qualitative Landmark Recognition	BoW Framework
Li, Xiaowei, et al. (2008) [47]	Landmark Recognition and Modelling	Iconic scene graphs

This section gives an insight into how the approaches for *Landmark Recognition* and *Instance Recognition* have been varying with time. The article *Babenko, Artem, et al. (2014)* tries to make neural networks for image recognition and retrieval [35]. It shows that simple machine learning techniques cannot work on image retrieval and we need more complex techniques. So, authors build neural networks including Convolutional Neural Networks (CNNs) which help us in image retrieval and classification. *Suau, Pablo. (2005)*, had developed polar histograms for robust *Landmark Recognition* [36].

Chen, Yuntao, et al. (2019) discusses how distributed frameworks help us in dealing with large datasets such as those for instance recognition [37]. We can design simple and versatile distributed frameworks for this matter. One of the developed frameworks: *SimpleDet*; has been discussed in the research article. According to “*Chen, Yuntao, et al.* ”, “*SimpleDet* supports up-to-date detection models with best practice. *SimpleDet* also supports distributed training with near-linear scaling out of the box.”

Crall, Jonathan P., et al. (2013) deals with one of the instance recognition frameworks namely species recognition [38]. This article discusses neural nets build for animal and plant species recognition. Though the article has little to do with *Landmark Recognition*, it was also one of the good articles in establishing procedures related to instant recognition. *HotSpotter* is a fast, accurate algorithm for identifying specific animals using a labeled database. It uses a *competitive scoring mechanism* derived from the *Local Naive Bayes Nearest Neighbor* algorithm.

Scheirer, Walter J., et al. (2011) discusses the important theory and practice methods followed while analyzing scores for image recognition [39]. The AI systems are assigned scores according to how well they fare at a task. For the task of Image Processing and Instance Recognition, this research article discussed how the scores are created, assigned, analyzed, and interpreted. Both theory and practice methods have been discussed conveniently.

Teichmann, Marvin, et al. (2019) discusses how Efficient Regional Aggregation can be used to identify images in an image classification dataset [40]. This is yet another technique used for image search. “*Teichmann, Marvin, et al.* ” demonstrate how a trained landmark detector, using their new dataset, could be

leveraged to index image regions and improve retrieval accuracy while being much more efficient than existing regional methods. *Zitová, Barbara, and Jan Flusser (1999)* discussed landmark recognition with features that remain invariant in data images [41]. This technique helped to further the research on Landmark Recognition by using similar features in different images.

The research article *Cao, Jiuwen, et al. (2015)* discusses more complex techniques namely sparse representation classification and extreme learning machines [42]. "*Cao, Jiuwen, et al.*" showed how these techniques, combined, produced good results for the field of *Landmark Recognition*. *Srinivasan, Sridhar, and Laveen Kanal. (1997)* discusses visual cues for image recognition [43]. The article explains how there are some visual cues in an image that may be used for Qualitative Landmark Recognition.

Cao, Jiuwen, Tao Chen, and Jiayuan Fan. (2014) the article proposed a fast online learning framework based on the Single hidden Layer Feedforward Neural Networks (SLFNs) [44]. This technique is also based on the bag-of-words (BoW) method which is used for feature extraction for *Landmark Recognition*. *Chen, Tao, and Kim-Hui Yap. (2013)*, discusses discriminative BoW framework for mobile *Landmark Recognition* [45].

Chen, Tao, Kim-Hui Yap, and Dajiang Zhang. (2012) was a research article published in 2012, which used the technique Discriminative bag-of-visual phrase learning for *Landmark Recognition* [46]. *Li, Xiaowei, et al. (2008)*, published in 2008, developed *Landmark Recognition and Modelling* using *iconic scene graphs* [47].

2.2.2. Survey for Instance Recognition

Table 5: Research articles with the area of Instance Recognition Addressed

Research Article	Area Addressed
Chen, Tao, et al. (2009) [48]	Mobile landmark recognition
Li, Hao, and Simon X. Yang. (2002) [49]	Visual Landmark Recognition systems
Chen, Tao, Kim-Hui Yap, and Dajiang Zhang. (2014) [50]	Mobile landmark recognition
Hui Yap, and Lap-Pui Chau. (2011) [51]	Mobile landmark recognition
Stefani, Pierre, and Cristina Oprean. (2019) [29]	Landmark recognition in photo albums
Guo, Jianya, Xi Mei, and Kun Tang. (2013) [53]	Identification for 3D things
Kragstov, Jens, et al. (1997) [54]	Landmark identification in craniofacial items
Déniz, Oscar, et al. (2011) [55]	Face recognition
Van Asselen, Marieke, Eva Fritschy, and Albert Postma. (2006) [56]	Better navigations
Epstein, Russell A., and Lindsay K. Vass. (2014) [57]	Landmarks Based way finding
Magliani, Federico, Tomaso Fontanini, and Andrea Prati. (2019) [58]	Small Scale to Large Scale Datasets
Zheng, Yan-Tao, et al. (2009) [59]	Web-scale Landmark Recognition
Wu, Jianixn, and James M. Rehg. (2008) [60]	Spatial Navigation

Chen, Tao, et al. (2009) was another survey article published in 2009 for information retrieval for mobile landmark recognition [48]. *Chen, Wei-Chao, et al., (2009)* was a research article published in 2009 which developed benchmarks for summarising *Landmarks* from community photo collections [49]. *Li, Hao, and Simon X. Yang. (2002)*, had discussed the evolution of visual *Landmark Recognition* systems [50].

Chen, Tao, Kim-Hui Yap, and Dajiang Zhang. (2014), published in 2014, discussed the technique named Discriminative soft bag-of-visual phrase which has been used for mobile *Landmark Recognition* [51]. *Chen,*

Tao, Kim-Hui Yap, and Lap-Pui Chau. (2011) published in 2011, uses integrated content and content analysis for mobile *Landmark Recognition* [52].

Stefani, Pierre, and Cristina Oprean. (2019) discussed landmark recognition in photo albums which was both fast and comprehensive [29]. In this article, a DELF network is used with a VLAD layer in a CNN network. Guo, Jianya, Xi Mei, and Kun Tang. (2013) had discussed 3D techniques for identification for 3D things [53]. This article used landmark annotations and a technique named dense correspondence registration for building 3D human facial images. The article describes how landmarks can be used in other fields.

Kragstov, Jens, et al. (1997) discussed important landmark identification in craniofacial items using cephalometric radiographs and CT scans [54]. This article, just like [53], is a case of landmarks other than inanimate objects. Déniz, Oscar, et al. (2011) discusses face recognition using oriented gradient technology in histograms [55]. This article, again like [55,56], is another instance of recognition of landmarks for animate objects.

Van Asselen, Marieke, Eva Fritschy, and Albert Postma. (2006) discusses psychological behaviors of humans in how they acquire spatial knowledge for navigation and how machines can be developed for better navigations using this learning [56]. Similarly, Epstein, Russell A., and Lindsay K. Vass. (2014) discusses neural systems for landmark-based wayfinding in humans [57]. This is a technique that is used not only used by animate organisms but using this technique, we may even be able to train an Artificial Intelligent computer System.

Magliani, Federico, Tomaso Fontanini, and Andrea Prati. (2019) discusses the datasets ranging from scale to large scale that have been used to date for Landmark Recognition and Retrieval [58]. This was also a comprehensive study that discussed all the simple and difficult traits of Landmark Recognition. Also, it discussed the achievements and research gaps of various datasets ranging from very small ones to ones having a million images. Zheng, Yan-Tao, et al. (2009) published in 2009, also aimed to develop a web-scale *Landmark Recognition* engine that could be used for later uses as well [59]. Wu, Jianxin, and James M. Rehg. (2008) used PACT technology to determine the whereabouts of a person or robot [60]. The technique uses *histograms* to know where a person is in space.

2.3. Survey for Robotic Technologies

Li, Hao, and Simon X. Yang (2003). was a research article published in 2003 which built a fully autonomous mobile robot [61]. It developed various modules that could implement different levels of competence and behaviors in the robots. Similarly, Nasr, Hatem, and Bir Bhanu. (1988) also worked on developing autonomous mobile robots [62]. This research article, however, was published in 1988, when not too much hardware existed. This article was more dealt with more theoretical aspects in building an autonomous mobile robot. The authors didn't practically implement anything, but the article paved way for a better understanding of future aspects. Trahanias, Panos E., Savvas Velissaris, and Thodoris Garavelos. (1997) published in 1997 had more success in *Landmark extraction and recognition* [63]. This article was one of the early articles published for *autonomous robot navigation* development. It was quite successful in bettering autonomous robot navigation.

Lee, Jung-Sub, et al. (2009) published in 2009, developed *In-pipe robot navigation* based on *Landmark Recognition using shadow images* [64]. This article, also for navigations of robots, dealt with developing robot navigation using a technique named *In-pipe robot navigation*. The technique used images of shadows of landmarks in an image to navigate through the surroundings. Elmoggy, Mohammed, and Jianwei Zhang. (2008) was a research article published in 2008, which developed robust and real-time techniques for *Landmark Recognition* which were used for *humanoid robot navigation* [65]. The authors discuss *humanoid robot navigation* which intends to make robots navigate their environment like humans. Mata,

published in 2001 discussed visual *Landmark Recognition* systems for topological navigation of mobile robots [66]. It discussed how robots can be trained to deal with all the 3-dimensional obstacles that enter their pathways. *Liu, Tingting, et al. (2019)* discuss how *Infrared Spectral Imaging* can be used to sense objects by robots [67]. This would lead to better *robotic vision*.

Table 6: Research articles with the area of Instance Recognition Addressed

ML Approach	Training Accuracy (in%)	Test Accuracy (in %)
SVM	51	11
KNN	45	11
Decision Tree	42	10
Random Forest	58	12

Yang, Simon X., and Hao Li. (2002) published in 2002, succeeded in developing mobile robots that not only were autonomous but could avoid obstacles in their path with a fuzzy logic behavior. This article was also used as a visual *Landmark Recognition* System [68].

Do, Quoc V., and Lakhmi C. Jain. (2006), published in 2006, was yet another attempt to develop landmark recognition systems for autonomous robot navigations [69]. *Luo, Ren C., Harsh Potlapalli, and David W. Hislop (1992)* had discussed how neural networks could be used for robot navigation [70]. This was developed using *Landmark Recognition* for robotic technologies developed using neural networks.

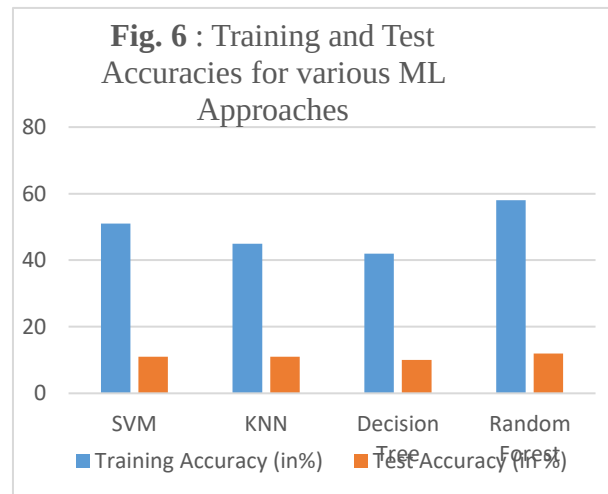
Table 7: Performance analysis of classical ML approaches on 1000 images out of Google Landmarks Dataset V2.

Research Article	Work Accomplished
Li, Hao, and Simon X. Yang (2003) [61]	Fully autonomous mobile robot
Nasr, Hatem, and Bir Bhanu. (1988) [62]	Autonomous mobile robots
Trahanias, Panos E., Savvas Velissaris, and Thodoris Garavelos. (1997) [63]	Autonomous robot navigation
Lee, Jung-Sub, et al. (2009) [64]	In-pipe robot navigation
Elmoggy, Mohammed, and Jianwei Zhang. (2008) [65]	Humanoid robot navigation
Mata, Mario, et al. (2001) [66]	Topological navigation of mobile robots
Liu, Tingting, et al. (2019) [67]	Infrared (IR) spectral imaging sensing
Yang, Simon X., and Hao Li. (2002) [68]	Mobile robots
Do, Quoc V., and Lakhmi C. Jain. (2006) [69]	Autonomous robot navigations
Luo, Ren C., Harsh Potlapalli, and David W. Hislop (1992) [70]	Robot navigation

3. Implementation of classical Machine Learning approaches

In this section, we analyze the performance of *SVM*, *KNN*, *Decision Tree*, and *Random Forest*, the four most popular classical *Machine Learning* Approaches. We analyzed the performance of these approaches on 1000 images out of *Google Landmarks Dataset V2*. The performance results are as shown in Table VII and Fig. 6.

From figure 6, we observe that *Random Forest* achieved 58% training accuracy on the training dataset and 18% on test data with 10 trees. Further, *SVM* performed the second best with 51% training accuracy. *KNN* and *decision tree* fared worst with 45% and 42% accuracies respectively. Though the dataset was very small, yet the accuracies are extremely less as compared to *Neural Network*-based models.



4 Future Scope

We have seen that a lot of work has been done in the field of Landmark Recognition. We see that with time, the sizes of datasets are growing as they might be expected to. However, as the dataset sizes increase, accuracies seem to go down. The techniques that seem to do well on smaller datasets fail miserably on larger ones. Hence, there is a need for developing techniques that do well not only on datasets of smaller sizes but on datasets of relatively all sizes.

Table 8: Various types of optimizations discussed by various research articles

Year	Authors	Research Article	Optimization
2005	Peterson, Michael R., Travis E. Doom, and Michael L. Raymer. [71]	Ga-facilitated knn classifier optimization with varying similarity measures.	Optimized KNN with Genetic Algorithm
2008	Lin, Shih-Wei, et al. [72]	Particle swarm optimization for parameter determination and feature selection of support vector machines.	Optimized SVM with Particle Swarm Optimization
2012	Samadzadegan, Farhad, Hadiseh Hasani, and Toni Schenk. [73]	Simultaneous feature selection and SVM parameter determination in classification of hyperspectral imagery using ant colony optimization.	Optimized SVM with Ant Colony Optimization
2017	El Houbby, Enas MF, Nisreen IR Yassin, and Shaimaa Omran. [74]	A hybrid approach from ant Colony optimization and K-nearest neighbor for classifying datasets using selected features.	Optimized KNN with Ant Colony Optimization
2017	Wei, Yan, et al. [75]	An improved grey wolf optimization strategy enhanced SVM and its application in predicting the second major.	Optimized SVM with Grey Wolf Optimization
2019	Sasirekha, K., and K. Thangavel. [76]	Optimization of K-nearest neighbor using particle swarm optimization for face recognition.	Optimized KNN with Particle Swarm Optimization
2019	Tao, Zhou, et al. [77]	GA-SVM based feature selection and parameter optimization in hospitalization expense modeling.	Optimized SVM with Genetic Algorithm
2020	Al-wajih, Ranya, et al. [78]	Binary Grey Wolf Optimizer with K-Nearest Neighbor classifier for Feature Selection.	Optimized KNN with Grey Wolf Optimization

As we are trying to develop AI models using and for natural elements in the surroundings, it might not be a thing of surprise that using a touch of nature should better the AI models currently being used for Image Processing for Landmark Recognition. No wonder, most of the algorithms and techniques developed today, for instance, recognition, is based on the natural processes and phenomena occurring in our environment.

Today, there exist many Nature Inspired Computing (NIC) Algorithms [71-78]. These algorithms are based on the working of natural phenomena in the environment and can be used for different areas like routing in computer networks, test case generations, image processing, etc [79-86]. The NIC algorithms can also be used to optimize the existing AI techniques like KNN, SVMs, etc. [72]

The processes and phenomena on which they are based include how organisms reproduce [84], how the bats see [85] and how our universe started [86]. These techniques provide a strong basis for the development of better AI models since AI is expected to work like natural processes. No wonder, Nature-Inspired Computing is widely used to optimize AI techniques today [71-112]. NIC algorithms work more like phenomena inspired by nature and hence are closer to functioning in a way that delivers nature-like results. Viewing is a natural phenomenon that we aim to pass on to machines. So, Nature Inspired Techniques can go a long way in developing Artificially Intelligent models.

We predict that Nature Inspired Computing techniques which include the Neural Networks will supposedly be used mostly in all future developments. They would be used to manufacture machines that would have abilities like those of living beings. An AI model that might be able to work close to the abilities that nature has provided should be inspired by nature.

5 Conclusion

We have surveyed the research articles ranging from 1988 till the present date related to Landmark Recognition and Instance Recognition. From the survey, we observed that since the sizes of datasets in the field of Landmark recognition have been growing (as they are expected to), the accuracies have been decreasing and there is a constant need for better technologies. Though better AI techniques have

continuously been developing, however, to expect robotic vision, we either need even better AI techniques or better hardware, or both. From the survey, we have concluded that the field of Landmark Recognition is given high importance in AI but it still has a long way to go due to various issues including the large size of datasets. Datasets in this field are extremely large and AI computer systems fail to train themselves in a proper manner and limited time. To make an AI system visualize things close to human-like precision, we need much better AI models and/or better hardware configurations. We also presented an implementation of classical Machine Learning Techniques on a part of Google Landmarks Dataset V2 and we could infer that these techniques will not work on datasets of such large sizes. A lot of research has already been done in this field and a lot more has yet to be done. If researchers reach appropriate results, this field can change the future of AI and robotic technologies.

References

- [1] Russell, Stuart, and Peter Norvig. "Artificial intelligence: a modern approach." (2002).
- [2] Zeng, Kaiman, et al. "Introducing AI to Undergraduate Students via Computer Vision Projects." AAAI. 2018.
- [3] Ge, Yuying, et al. "Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images." Proceedings of the IEEE conference on computer vision and pattern recognition. 2019.
- [4] Mikołajczyk, Agnieszka, and Michał Grochowski. "Data augmentation for improving deep learning in image classification problem." 2018 international interdisciplinary PhD workshop (IIPhDW). IEEE, 2018.
- [5] Zhou, Bolei, et al. "Places: A 10 million image database for scene recognition." IEEE transactions on pattern analysis and machine intelligence 40.6 (2017): 1452-1464.
- [6] Chen, Kaibing, et al. "2nd Place and 2nd Place Solution to Kaggle Landmark Recognition and Retrieval Competition 2019." arXiv preprint arXiv:1906.03990 (2019).
- [7] Kim, Jinmook. "A Study on Designing Metadata Standard for Building AI Training Dataset of Landmark Images." Journal of the Korean Society for Library and Information Science 54.2 (2020): 419-434.
- [8] Ozaki, Kohei, and Shuhei Yokoo. "Large-scale landmark retrieval/recognition under a noisy and diverse dataset." arXiv preprint arXiv:1906.04087 (2019).
- [9] Weyand, Tobias, et al. "Google Landmarks Dataset v2-A Large-Scale Benchmark for Instance-Level Recognition and Retrieval." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
- [10] Noh, Hyeonwoo, et al. "Large-scale image retrieval with attentive deep local features." Proceedings of the IEEE international conference on computer vision. 2017.
- [11] Weyand, Tobias, and Bastian Leibe. "Visual landmark recognition from internet photo collections: A large-scale evaluation." Computer Vision and Image Understanding 135 (2015): 1-15.
- [12] Radenović, Filip, Giorgos Tolias, and Ondřej Chum. "Fine-tuning CNN image retrieval with no human annotation." IEEE transactions on pattern analysis and machine intelligence 41.7 (2018): 1655-1668.
- [13] Gordo, Albert, et al. "Deep image retrieval: Learning global representations for image search." European conference on computer vision. Springer, Cham, 2016.
- [14] Radenović, Filip, et al. "Revisiting oxford and paris: Large-scale image retrieval benchmarking." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
- [15] Torii, Akihiko, et al. "24/7 place recognition by view synthesis." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015.
- [16] Li, Yunpeng, et al. "Worldwide pose estimation using 3d point clouds." European conference on computer vision. Springer, Berlin, Heidelberg, 2012.
- [17] Agarwal, Sameer, et al. "Building rome in a day." Communications of the ACM 54.10 (2011): 105-112.
- [18] Chen, David M., et al. "City-scale landmark identification on mobile devices." CVPR 2011. IEEE, 2011.
- [19] Avrithis, Yannis, Giorgos Tolias, and Yannis Kalantidis. "Feature map hashing: Sub-linear indexing of appearance and global geometry." Proceedings of the 18th ACM international conference on Multimedia. 2010.
- [20] Knopp, Jan, Josef Sivic, and Tomas Pajdla. "Avoiding confusing features in place recognition." European Conference on Computer Vision. Springer, Berlin, Heidelberg, 2010.
- [21] Philbin, James, et al. "Lost in quantization: Improving particular object retrieval in large scale image databases." 2008 IEEE conference on computer vision and pattern recognition. IEEE, 2008.
- [22] Jegou, Herve, Matthijs Douze, and Cordelia Schmid. "Hamming embedding and weak geometric consistency for large scale image search." European conference on computer vision. Springer, Berlin, Heidelberg, 2008.
- [23] Cao, Jiuwen, et al. "Landmark recognition with sparse representation classification and extreme learning machine." Journal of the Franklin Institute 352.10 (2015): 4528-4545.

- [24] Boiarov, Andrei, and Eduard Tyantov. "Large scale landmark recognition via deep metric learning." Proceedings of the 28th ACM International Conference on Information and Knowledge Management. 2019.
- [25] Magliani, Federico, and Andrea Prati. "An accurate retrieval through R-MAC+ descriptors for landmark recognition." Proceedings of the 12th International Conference on Distributed Smart Cameras. 2018.
- [26] Magliani, Federico, Navid Mahmoudian Bidgoli, and Andrea Prati. "A location-aware embedding technique for accurate landmark recognition." Proceedings of the 11th International Conference on Distributed Smart Cameras. 2017.
- [27] Magliani, Federico, Tomaso Fontanini, and Andrea Prati. "Efficient nearest neighbors search for large-scale landmark recognition." International Symposium on Visual Computing. Springer, Cham, 2018.
- [28] Vandaele, Rémy, et al. "Landmark detection in 2D bioimages for geometric morphometrics: a multi-resolution tree-based approach." Scientific reports 8.1 (2018): 1-13.
- [29] Stefani, Pierre, and Cristina Oprean. "Fast Landmark Recognition in Photo Albums." International Conference on Computer Analysis of Images and Patterns. Springer, Cham, 2019.
- [30] Cumbo, Kodi CA, et al. "Bee-inspired landmark recognition in robotic navigation." Proceedings of the 2016 on Genetic and Evolutionary Computation Conference Companion. 2016.
- [31] Wen, Xian-Bin, Hua Zhang, and Ze-Tao Jiang. "Multiscale unsupervised segmentation of SAR imagery using the genetic algorithm." Sensors 8.3 (2008): 1704-1711.
- [32] Maeda, J., et al. "Perceptual image segmentation using fuzzy-based hierarchical algorithm and its application to dermoscopy images." 2008 IEEE Conference on Soft Computing in Industrial Applications. IEEE, 2008.
- [33] Talbi, Hichem, Mohamed Batouche, and Amer Draa. "A quantum-inspired evolutionary algorithm for multiobjective image segmentation." International Journal of Mathematical, Physical and Engineering Sciences 1.2 (2007): 109-114.
- [34] Moreira, Jander, and Luciano Da Fontoura Costa. "Neural-based color image segmentation and classification using self-organizing maps." Anais do IX Sibgrapi 12.6 (1996): 47-54.
- [35] Babenko, Artem, et al. "Neural codes for image retrieval." European conference on computer vision. Springer, Cham, 2014.
- [36] Suau, Pablo. "Robust artificial landmark recognition using polar histograms." Portuguese Conference on Artificial Intelligence. Springer, Berlin, Heidelberg, 2005.
- [37] Chen, Yuntao, et al. "SimpleDet: A Simple and Versatile Distributed Framework for Object Detection and Instance Recognition." Journal of Machine Learning Research 20.156 (2019): 1-8.
- [38] Crall, Jonathan P., et al. "Hotspotter—patterned species instance recognition." 2013 IEEE workshop on applications of computer vision (WACV). IEEE, 2013.
- [39] Scheirer, Walter J., et al. "Meta-recognition: The theory and practice of recognition score analysis." IEEE transactions on pattern analysis and machine intelligence 33.8 (2011): 1689-1695.
- [40] Teichmann, Marvin, et al. "Detect-to-retrieve: Efficient regional aggregation for image search." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.
- [41] Zitová, Barbara, and Jan Flusser. "Landmark recognition using invariant features." Pattern Recognition Letters 20.5 (1999): 541-547.
- [42] Cao, Jiuwen, et al. "Landmark recognition with sparse representation classification and extreme learning machine." Journal of the Franklin Institute 352.10 (2015): 4528-4545.
- [43] Srinivasan, Sridhar, and Laveen Kanal. "Qualitative landmark recognition using visual cues." Pattern Recognition Letters 18.11-13 (1997): 1405-1414.
- [44] Cao, Jiuwen, Tao Chen, and Jiayuan Fan. "Fast online learning algorithm for landmark recognition based on BoW framework." 2014 9th IEEE Conference on Industrial Electronics and Applications. IEEE, 2014.
- [45] Chen, Tao, and Kim-Hui Yap. "Discriminative BoW framework for mobile landmark recognition." IEEE transactions on cybernetics 44.5 (2013): 695-706.
- [46] Chen, Tao, Kim-Hui Yap, and Dajiang Zhang. "Discriminative bag-of-visual phrase learning for landmark recognition." 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2012.
- [47] Li, Xiaowei, et al. "Modeling and recognition of landmark image collections using iconic scene graphs." European conference on computer vision. Springer, Berlin, Heidelberg, 2008.
- [48] Chen, Tao, et al. "A survey on mobile landmark recognition for information retrieval." 2009 Tenth International Conference on Mobile Data Management: Systems, Services and Middleware. IEEE, 2009.
- [49] Chen, Wei-Chao, et al. "Visual summaries of popular landmarks from community photo collections." 2009 Conference Record of the Forty-Third Asilomar Conference on Signals, Systems and Computers. IEEE, 2009.
- [50] Li, Hao, and Simon X. Yang. "An evolutionary visual landmark recognition system." IEEE International Conference on Systems, Man and Cybernetics. Vol. 7. IEEE, 2002.

- [51] Chen, Tao, Kim-Hui Yap, and Dajiang Zhang. "Discriminative soft bag-of-visual phrase for mobile landmark recognition." *IEEE Transactions on Multimedia* 16.3 (2014): 612-622.
- [52] Chen, Tao, Kim-Hui Yap, and Lap-Pui Chau. "Integrated content and context analysis for mobile landmark recognition." *IEEE Transactions on Circuits and Systems for Video Technology* 21.10 (2011): 1476-1486.
- [53] Guo, Jianya, Xi Mei, and Kun Tang. "Automatic landmark annotation and dense correspondence registration for 3D human facial images." *BMC bioinformatics* 14.1 (2013): 232.
- [54] Kragstov, Jens, et al. "Comparison of the reliability of craniofacial anatomic landmarks based on cephalometric radiographs and three-dimensional CT scans." *The Cleft palate-craniofacial journal* 34.2 (1997): 111-116.
- [55] Déniz, Oscar, et al. "Face recognition using histograms of oriented gradients." *Pattern recognition letters* 32.12 (2011): 1598-1603.
- [56] van Asselen, Marieke, Eva Fritschy, and Albert Postma. "The influence of intentional and incidental learning on acquiring spatial knowledge during navigation." *Psychological Research* 70.2 (2006): 151-156.
- [57] Epstein, Russell A., and Lindsay K. Vass. "Neural systems for landmark-based wayfinding in humans." *Philosophical Transactions of the Royal Society B: Biological Sciences* 369.1635 (2014): 20120533.
- [58] Magliani, Federico, Tomaso Fontanini, and Andrea Prati. "Landmark recognition: from small-scale to large-scale retrieval." *Recent Advances in Computer Vision*. Springer, Cham, 2019. 237-259.
- [59] Zheng, Yan-Tao, et al. "Tour the world: building a web-scale landmark recognition engine." 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009.
- [60] Wu, Jianixn, and James M. Rehg. "Where am I: Place instance and category recognition using spatial PACT." 2008 Ieee Conference on Computer Vision and Pattern Recognition. IEEE, 2008.
- [61] Li, Hao, and Simon X. Yang. "A behavior-based mobile robot with a visual landmark-recognition system." *IEEE/ASME transactions on mechatronics* 8.3 (2003): 390-400.
- [62] Nasr, Hatem, and Bir Bhanu. "Landmark recognition for autonomous mobile robots." *Proceedings. 1988 IEEE International Conference on Robotics and Automation*. IEEE, 1988.
- [63] Trahanias, Panos E., Savvas Velissaris, and Thodoris Gavelos. "Visual landmark extraction and recognition for autonomous robot navigation." *Proceedings of the 1997 IEEE/RSJ International Conference on Intelligent Robot and Systems. Innovative Robotics for Real-World Applications. IROS'97. Vol. 2. IEEE, 1997.*
- [64] Lee, Jung-Sub, et al. "In-pipe robot navigation based on the landmark recognition system using shadow images." 2009 IEEE international conference on robotics and automation. IEEE, 2009.
- [65] Elmogy, Mohammed, and Jianwei Zhang. "Robust real-time landmark recognition for humanoid robot navigation." 2008 IEEE International Conference on Robotics and Biomimetics. IEEE, 2009.
- [66] Mata, Mario, et al. "A visual landmark recognition system for topological navigation of mobile robots." *Proceedings 2001 ICRA. IEEE International Conference on Robotics and Automation (Cat. No. 01CH37164). Vol. 2. IEEE, 2001.*
- [67] Liu, Tingting, et al. "Flexible FTIR spectral imaging enhancement for industrial robot infrared vision sensing." *IEEE Transactions on Industrial Informatics* 16.1 (2019): 544-554.
- [68] Yang, Simon X., and Hao Li. "An autonomous mobile robot with fuzzy obstacle avoidance behaviors and a visual landmark recognition system." 7th International Conference on Control, Automation, Robotics and Vision, 2002. ICARCV 2002. Vol. 3. IEEE, 2002.
- [69] Do, Quoc V., and Lakhmi C. Jain. "A visual landmark recognition system for autonomous robot navigation." 2006 International Conference on Computational Intelligence for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce (CIMCA'06). IEEE, 2006.
- [70] Luo, Ren C., Harsh Potlapalli, and David W. Hislop. "Neural network based landmark recognition for robot navigation." *Proceedings of the 1992 International Conference on Industrial Electronics, Control, Instrumentation, and Automation*. IEEE, 1992.
- [71] Peterson, Michael R., Travis E. Doom, and Michael L. Raymer. "Ga-facilitated knn classifier optimization with varying similarity measures." 2005 IEEE congress on evolutionary computation. Vol. 3. IEEE, 2005.
- [72] Lin, Shih-Wei, et al. "Particle swarm optimization for parameter determination and feature selection of support vector machines." *Expert systems with applications* 35.4 (2008): 1817-1824.
- [73] Samadzadegan, Farhad, Hadiseh Hasani, and Toni Schenk. "Simultaneous feature selection and SVM parameter determination in classification of hyperspectral imagery using ant colony optimization." *Canadian Journal of Remote Sensing* 38.2 (2012): 139-156.
- [74] El Houby, Enas MF, Nisreen IR Yassin, and Shaimaa Omran. "A hybrid approach from ant Colony optimization and K-nearest neighbor for classifying datasets using selected features." *Informatica* 41.4 (2017).
- [75] Wei, Yan, et al. "An improved grey wolf optimization strategy enhanced SVM and its application in predicting the second major." *Mathematical Problems in Engineering* 2017 (2017).
- [76] Sasirekha, K., and K. Thangavel. "Optimization of K-nearest neighbor using particle swarm optimization for face recognition." *Neural Computing and Applications* 31.11 (2019): 7935-7944.

- [77] Tao, Zhou, et al. "GA-SVM based feature selection and parameter optimization in hospitalization expense modeling." *Applied Soft Computing* 75 (2019): 323-332.
- [78] Al-wajih, Ranya, et al. "Binary Grey Wolf Optimizer with K-Nearest Neighbor classifier for Feature Selection." 2020 International Conference on Computational Intelligence (ICCI). IEEE, 2020.
- [79] S. He, Q. H. Wu, and J. Saunders, "Group search optimizer: an optimization algorithm inspired by animal searching behavior," *IEEE transactions on evolutionary computation*, vol. 13, no. 5, pp. 973–990, 2009.
- [80] S. Binita, S. S. Sathya et al., "A survey of bio inspired optimization algorithms," *International Journal of Soft Computing and Engineering*, vol. 2, no. 2, pp. 137– 151, 2012.
- [81] Amar Singh, Shakti Kumar, S.S. Walia, Sandeep Chakravorty, "Face Recognition : A Combined Parallel BB-BC & PCA Approach to Feature Selection", *International Journal of Computer Science & Information Technology*, ISSN 1694 – 2380, Volume 2, spl issue 2 (2015), pp. 1-5.
- [82] Amar Singh, Shakti Kumar, Ajay Singh , Sukhbir Singh Walia, "Three-parent GA: A Global Optimization Algorithm", *Journal of Multiple-Valued Logic and Soft Computing*, Volume 32, 2019, pp. 407 – 423.
- [83] Shakti Kumar, Amar Singh, Sukhbir Singh Walia, "Parallel Big Bang - Big Crunch Global Optimization Algorithm: Performance and its Applications to routing in WMNs", *Wireless Personal Communications*, Springer, 100(4), 2018, pp. 1601- 1618.
- [84] Amar Singh, Shakti Kumar, Sukhbir Singh Walia, "Parallel 3-Parent Genetic Algorithm with Application to Routing in Wireless Mesh Networks", *Implementations and Applications of Machine Learning*, Springer, 2020, pp. 1 – 28.
- [85] X.-S. Yang, "A new metaheuristic bat-inspired algorithm," *Nature inspired cooperative strategies for optimization (NICSO 2010)*, pp. 65–74, 2010.
- [86] Erol O. K., Eksin I., "A new optimization method: Big Bang-Big Crunch", *Advances in Engineering Software*, Vol. 37, 2006, pp. 106-111
- [87] Kaveh, A. and Farhoudi, N., (2013), "A new optimization method: dolphin echolocation," *Advances in Engineering Software*, vol. 59, pp. 53–70.
- [88] Sur, C., Sharma, S., and Shukla, A., (2013), "Egyptian vulture optimization algorithm—a new nature inspired meta-heuristics for knapsack problem," *Proceedings of 9th International Conference on Computing and Information Technology, (IC2IT2013)*, King Mongkut's University of Technology North Bangkok, Springer, pp. 227–237.
- [89] Hernández, H. and Blum, C., (2012), "Distributed graph coloring: an approach based on the calling behavior of japanese tree frogs," *Swarm Intelligence*, 6(2), pp. 117–150.
- [90] Parpinelli, R. S. and Lopes, H. S., (2011), "An eco-inspired evolutionary algorithm applied to numerical optimization," *Proceedings of Third World Congress on Nature and Biologically Inspired Computing (NaBIC)*, , Salamanca, IEEE, pp. 466–471.
- [91] Bastos Filho, C. J., de Lima Neto F. B., Lins, A. J., Nascimento, A. I., and Lima, M. P., (2008), "A novel search algorithm based on fish school behavior," *Proceedings of IEEE International Conference on Systems, Man and Cybernetics, SMC 2008*, Singapore, pp. 2646–2651.
- [92] Carmelo, J. A. Bastos Filho, Fernando, B. de Lima Neto, Anthony, J. C. C. Lins, Antonio, I. S. Nascimento, Marilia, P. Lima, (2008), "A novel search algorithm based on fish school behavior," *IEEE International Conference on Systems, Man and Cybernetics*, Singapore, pp. 2646–2651.
- [93] Yang X. S., (2012), "Flower pollination algorithm for global optimization," *Unconventional Computation and Natural Computation*, Series Title : Lecture Notes in Computer Science, Springer, pp. 240–249.
- [94] Yang, X. S., Karamanoglu M., and He, X., (2013), "Multi-objective flower algorithm for optimization," *Procedia Computer Science*, 18(Supplement C), pp. 861–868.
- [95] Ferreira, C., (2002), "Gene expression programming in problem solving," *Proceedings of Soft Computing and Industry: Recent Applications*, Springer, London, pp. 635–653.
- [96] Mozaffari, A., Fathi, A., and Behzadipour, S., (2012), "The great salmon run: a novel bioinspired algorithm for artificial system design and optimisation," *International Journal of Bio-Inspired Computation*, 4(5), pp. 286–301.
- [97] He, S., Wu, Q. H., and Saunders, J., (2009), "Group search optimizer: an optimization algorithm inspired by animal searching behavior," *IEEE transactions on evolutionary computation*, 13(5), pp. 973–990.
- [98] Zhang, L. M., Dahlmann, C., and Zhang, Y., (2009), "Human-inspired algorithms for continuous function optimization," *Proceedings of IEEE International Conference on Intelligent Computing and Intelligent Systems*, vol. 1, Shanghai, China, pp. 318–321.
- [99] Mehrabian, A. R. and Lucas, C., (2006), "A novel numerical optimization algorithm inspired from weed colonization," *Ecological informatics*, 1(4), pp. 355– 366.
- [100] Maia, R. D., de Castro, L. N., and Caminhas, W. M., (2012), "Bee colonies as model for multimodal continuous optimization: The optbees algorithm," *Proceedings of IEEE Congress on Evolutionary Computation (CEC)*, Brisbane, QLD, Australia, pp. 1–8.

- a. Premaratne, U., Samarabandu, J., and Sidhu, T., (2009), "A new biologically inspired optimization algorithm," Proceedings of International Conference on Industrial and Information Systems (ICIIS), IEEE, Sri Lanka, pp. 279–284.
- [101] Havens, T. C., Spain, C. J., Salmon, N. G., and Keller, J. M., (2008), "Roach infestation optimization," Proceedings of Swarm Intelligence Symposium, SIS 2008, IEEE, St. Louis, MO, USA, pp. 1–7.
- [102] Jung, S. H., (2003), "Queen-bee evolution for genetic algorithms," Electronics letters, 39(6), pp. 575–576.
- [103] Eusuff, M. M. and Lansey, K. E., (2003), "Optimization of water distribution network design using the shuffled frog leaping algorithm," Journal of Water Resources planning and management, 129(3), pp. 210–225.
- [104] Hedayatzadeh, R., Salmassi, F. A., Keshtgari, M., Akbari, R., and Ziarati, K., (2010), "Termite colony optimization: A novel approach for optimizing continuous problems," Proceedings of 18th Iranian Conference on Electrical Engineering (ICEE), IEEE, Isfahan, Iran, pp. 553–558.
- [105] Ahmadi-Javid, A., 2011, June. Anarchic Society Optimization: A human-inspired method. In Evolutionary Computation (CEC), 2011 IEEE Congress on (pp. 2586-2592). IEEE.
- [106] Civicioglu, P., (2013) "Artificial cooperative search algorithm for numerical optimization problems". Information Sciences, 229, pp.58-76.
- [107] Civicioglu, P., (2012), "Transforming geocentric cartesian coordinates to geodetic coordinates by using differential search algorithm". Computers & Geosciences, 46, pp.229-247.
- [108] O'Neill, M. and Ryan, C., (2001), "Grammatical evolution", IEEE Transactions on Evolutionary Computation, 5(4), pp.349-358.
- [109] Kashan, A.H., (2014), "League Championship Algorithm (LCA): An algorithm for global optimization inspired by sport championships", Applied Soft Computing, 16, pp.171-200.
- [110] Xu, Y., Cui, Z. and Zeng, J., (2010), "Social emotional optimization algorithm for nonlinear constrained optimization problems" International Conference on Swarm, Evolutionary and Memetic Computing, Springer, Berlin, Heidelberg, pp. 583-590.
- [111] Civicioglu, P., (2013), Backtracking search optimization algorithm for numerical optimization problems. Applied Mathematics and Computation, 219(15), pp.8121-8144.
- [112] Atashpaz-Gargari, E. and Lucas, C., (2007), "Imperialist competitive algorithm: an algorithm for optimization inspired by imperialistic competition", Evolutionary computation, CEC 2007. IEEE Congress on IEEE, pp. 4661-4667.

Maintenance Directives of Analysis and Design for Complex Embedded System through Reverse Engineering

Surinder Pal Singh¹, Ankur Mangla², Vishal Gupta³, Charanjit Singh⁴

¹Computer Science and Engineering Department, Maharishi Markandeshwar University

²Computer Science and Engineering Department, Maharishi Markandeshwar University

³Computer Science and Engineering Department, Maharishi Markandeshwar University

⁴Computer Science and Engineering Department, Maharishi Markandeshwar University

s.p.singh@mmumullana.org¹, ankurmangla2006@mmumullana.org², vishal.gupta@mmumullana.org³,
charan2910@mmumullana.org⁴

Abstract : A Semi-Automatic Complex embedded systems has been superseded and difficult to replace because of its wide use. A Thousands of engineers are working for several years so that a complex embedded system will get launched. A special-purpose system such as semi-automatic complex embedded system are developed with the help of software and/or hardware so that they can control all the logical/physical processes of the system. It become more problematic for older Semi-Automatic Complex embedded systems because they are difficult to understand (and hence, to maintain) due to lack of good quality documentation and not availability of engineers or company. This paper describe an approach to understand the system data structure, passing of information, their outputs and halt state.

Keywords : Embedded system, Facts Extraction , State model through tuning machine.

1. Introduction

A Semi-Automatic Complex embedded system are developed and executed in such a way that set of problems will transfer messages mainly through queue or shared memory. There are many off-line techniques such as switch off buttons can change the status of the working system. If certain conditions will met then also behavior of the system can be changed. In practice, these systems are implemented on top of microprocessor that uses hardware and software inputs which are preemptive at design time, but priorities may also change dynamically. The developer or maintainers have to deal with many tasks with many constraints which include priority, deadline of the task etc. The system correctness also depends on worst-case execution time without any interruption. It's a matter of data structure predictability, average performance therefore poses a problem or responsibility that causes difficulty or hard work on verification via code analyses and testing[2].

Software Maintenance of complex embedded software systems is faced with challenges, which includes runtime behavior, analysis of the system , quality , reliability , integrity .In complex Semi-Automatic embedded systems visualization and verification techniques are used for data gathering so that good quality documentation can be done[3]. With the use of visualization and verification of the system, a new set of data structure and new enhancement for the system can be done with new improved design. Our main focus is on improving human understanding about the system and how this information is processed. It tackles the question of what information is stored and how this information is used by the microprocessor to get output. There are several tools already available in the markets that are used for understanding complex programs. These tools mainly focuses on transferring information about a system's design into the mind of the developer. The envisioned design are done by the data gathering and the code back into an abstract design. A programmer can easily use interface to handle interrupts, controls, access peripherals through system design. Software quality and robustness can be achieved by heuristic analysis of the system.

Simulation will be done to inspect for undoubted stipulating patterns that appear in directives and tells about the quality of the code. All Test input can be given by scripts well as by application resource which may be on line arguments or called by value/reference. Simulation gives an auto-generated test report which also helps in data gathering and working of the system. Now, this is the best stage at which quality code for hardware should be done [4].

2. Existing Work

Reverse engineering, also called back engineering, is the process by which man-made software is analyzed by deconstruction of constituent parts to get unknown or secret information. Reverse Engineering can be defined as "Gathering of information from software system through analyzing the source code"[1]. The process of reverse engineering starts when the complex embedded system's code with basic level of commands is collected which also includes model domain. Architecture of complex embedded systems needs all the belongings and effects of programs that connect hardware and/or software and/or resources[6]. In this section, we summarize static and dynamic analysis of complex embedded system. The problem of minimizing the advance seeking of analyzing the complex end-user devices is to get responses (information) from the complex devices so as to improve the Maintenance of complex Systems[14].

Static analysis. Static information are gathered from source code by finding its formal and actual arguments, also help in building scripts and this approach is also helpful in finding specification of complex embedded systems.

Dynamic analysis. Dynamic information are gathered in log files, traces and error messages which also helps in finding system output and its data structure.

Static information may be obtained without compiling and running the system. When program runs repeatedly then we have chance to collect information which was real-time and it helps in maintaining the system. These both Static and Dynamic analysis gives a clear picture about the system and its output.

2.1 Workflow of Process

There are many ways on which workflow of reverse engineering can be divided but mainly it can be break-up into three processes: Extraction, Analysis, and Visualization. Extraction means removal or pulling out of the information from the process. Analysis means getting information through statistics or past experience[8]. Visualization means how the system works in real time. The reverse engineer steps are repeated and refined several times to get the appropriate outcome. These innovative steps are necessary for the reverse engineering process.

2.2 Tool Support

These workflow output gives facts extraction of the software, analysis of the control flow of the software and visualize of the software which help to assist the reverse engineers. These output are stored in a central repository which also taken in consideration of functionality as well as about the working environment. The central repository stores requirements as a blueprint for the system working. The information in the repository is structured enough to determine what system does but will not determine what thing went wrong with the software model. It has information in a structured way so that a data model, schema, meta-model can be made and new software can be made with appropriate environment.

3. Directives for Analysis and Design of Semi-Automatic Complex Embedded System

Our main focus is on Complex Embedded system, which consists of many varieties of equipment ranges from automobile industries to household automated units and from biological imaging technologies to electricity generated plants. Complex embedded software systems consists of mainly hardware, firmware and software which controls the different working processes with the help of devices that stimulate heat, light, sound, pressure, magnetism, or a particular motion etc. In our approach, we maintain the complex embedded system when the code of the system is available and (i) There is inadequate documentation of the original design (ii) The original design documentation has been lost or never existed. (iii) The original CAD model is not sufficient to support modifications or current manufacturing methods. Our framework is used to get system analysis and design of complex embedded system. It is divided into mainly two parts (i) Facts Extraction (ii) State model through tuning machine. Mainly CAD intense on using portray and Artificial Intelligence method to captured data using geometrical methods and try to do reverse engineering from it[15].

3.1 Facts Extraction

It is done with help of five main directives:

(i) Interface Analysis. In this directive, we have to imagine or have a picture in our mind. How this complex embedded system works and checks the flow of different tasks done by the system. We also check the functionality of the controls/Button used in the system by seeing the inputs given by the single control.

Interface requirements is by seeing the existing system, then building a context diagram which can be easily broken down into data flow diagrams for more detailed analysis.

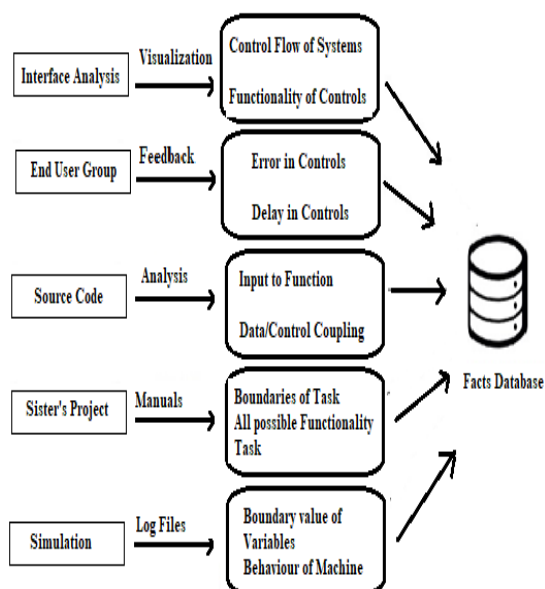


Fig 1. Facts Extraction

(ii) End-user Group. In this directives, On the basis of End-User group experience which works on the system tells about the failure in the system, Error or Delay shown by the controls when particular input was given to the system. End-user group gives a lot of documentation which is available for users to help them understand and properly.

(iii) Source Code. In this directive, we study the source code of the embedded system and gets the passing arguments of the every job/task, in further, we gets the range of every task and control flows from one tasks to another tasks. Requirement gathering from source code helps the software engineers to trace a requirement from its start to its final code[7].

(iv) Sister's Project. In these directives, we basically have a manual of another embedded system which is similar to our existing system. Through this ,we basically finds all possible functionality of the existing embedded system and also gets the boundary values of the individual task/job. This also helps in the verification of requirement as well as hazards and performance analysis of the embedded system.

(v) Simulation. In this directive, source code is checked through simulator and regression testing is done on the system. From this procedure, we gets the log files which helps in finding boundary values of variables taken by the programmer and also gets the behavior of the machine when wrong input given by the simulator[9].

All these facts are stored in repository and given to the focus group (Team leaders). They further do analysis or refine and give their individual feedback about the system. It can be collected to refine and validate the already elicited requirements. Before we go to design directives of complex embedded system .There are some application where we can use tuning machine and state diagram for embedded system.

Traffic control system. Integrated traffic lights Embedded system can check which directions have the large flow of vehicles on the road and On this information embedded system can change the traffic lights and their timings. The delay can depend on traffic road conditions, the physical layout of the intersection, and legal requirements. This will be a great beneficial to cities which are overcrowded and timings when there are maximum traffic on the road and this is not easily controlled by fixed timer traffic lights.

Smart homes. IOT (Internet Of Things) are having interlocking loops with embedded systems. We put our home network at our commands through command by voice, remote control, tablet or smartphone. All the appliances of a smart home which can be controlled via the internet are integrated with embedded systems.

Automotive industry. Automobile industry is also equipped with embedded systems. Ranging from wiper controls to complex anti-lock brake controls and air bags, embedded systems have gained overall control of automobiles.

Healthcare. A lot of embedded systems are used in health care and doctors are using imaging tools to diagnose health problems and to track treatment progress of the patient. With the help of smart embedded system, patients can be managed easily at home. These Smart embedded devices permit users to repeatedly checking their heart rate, blood pressure, oxygen rate and various other parameters.

4. Design Directives of complex embedded system:

Design directive for embedded system: State model through Turing Machine.

4.1 State model through Turing Machine. It is composed of state and transitions where states are different outputs which have a combination of different outputs. Transitions from one state to another state when there are some input is labeled on the arc. If this is exact input that system requires then the associated output state will be achieved otherwise system goes to interrupt with associated output with particular condition. State models are used to give a conceptual explanation of the working of an embedded system.

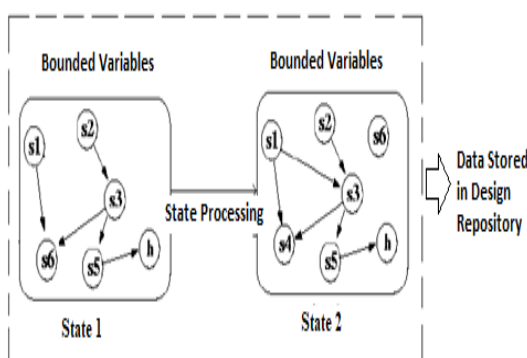


Fig. 2 State Model of Turing Machine

We have to find input for particular state of embedded system which can be variable inputs and gives different output at different condition so we have check all the conditions of the embedded machine with the help of repository that we have get from fact extraction. This working or processes are examined and represented with a series of events which can occur in one or more different states[11]. The information entered into module reads all the possible inputs, which includes synchronous inputs and asynchronous events. The input module updates the corresponding variables and feeds them to the state transfer module. Most important, we have to find all the variables which are dependent on timings, voltage, heat and pressure and also define the bounded integer variable which gives different output with small variation. By this strategy we can easily finished the detection strategy obtained by filling the state model of tuning machine system with the identified output and can be inserted in the results repository.

5. Conclusion & Future Scope

Maintenance for complex embedded system requires high degree of complexity of computing function at low cost and low consumption. Therefore need exact analysis and design of complex embedded system. We have presented directives for gathering of requirement using fact extraction and state model for tuning machine which are stored in repository. We have presented two modes of directives for requirement gathering. Facts Extraction that analysis the system for removing the bottlenecks and provide theoretical suggestions to the programmer. A programmer takes advantages of this information and improves its performance. The second approach to complex embedded system for design directives includes state model and tuning machine of different processes which helps in designing the well-defined behavior of the system. Using these directives, the programmer can design its complex embedded system in a simple way with all tasks completed. These

directives classify and analyze the current embedded system mining of data. Finally, suggestions for the future are to extracting data of complex embedded system like (IOT devices)through these directives.

References

- [1] E. J. Byrne, Editor, "Software Reverse engineering: A case study", Software-Practice and Experience, 1991.
- [2] A.C. Telea, "Reverse Engineering - Recent Advances and Applications", Ed. Intech2012.
- [3] Kraft, J., Kienle, H. M., Nolte, T., Crnkovic, I. & Hansson, H.. Software maintenance research in the PROGRESS project for predictable embedded software systems, 15th IEEE European Conference on Software Maintenance and Reengineering (CSMR 2011) ,pp. 335–338.
- [4] Andersson, J., Huselius, J., Norström, C. & Wall, A. Extracting simulation models from complex embedded real-time systems, 1st International Conference on Software Engineering Advances (ICSEA 2006).
- [5] D. D. Gajski and F. Vahid, "Specification and design of embedded software-hardware systems," IEEE Design & Test of Computers, Spring(1995).
- [6] M. D. Ernst. Static and dynamic analysis: synergy and duality. In ICSE Workshop on Dynamic Analysis (WODA), Portland, Oregon, USA(2003), May.
- [7] G. Ramalingam, J. Field, and F. Tip. Aggregate structure identification and its application to program analysis. In Proceedings of the 26th ACM SIGPLAN-SIGACT symposium on Principles of Programming Languages,(1999).
- [8] G. Canfora, M. Di Penta, New frontiers of reverse engineering, in: Future of Soft. Eng., FOSE, IEEE Computer Society,(2007), pp. 326–341.
- [9] Várady T, Martin RR, Cox J. Reverse engineering of geometric models – an introduction. Computer Aided Design,(1997)
- [10] A. V. Someren and C. Attack, The ARM RISC Chip, A Programmer's Guide. Addison-Wesley, (1993).
- [11] D. Ratiu, S. Ducasse, T. Girba, and R. Marinescu. Using History Information to Improve Design Flaws Detection. In Proceedings of CSMR04. IEEE Computer Society, (2004) ,pp. 223–232.
- [12] Bohlin, M., Lu, Y., Kraft, J., Kreuger, P. & Nolte, T. Simulation-Based Timing Analysis of Complex Real-Time Systems, Proceedings of the 15th IEEE International Conference on Embedded and Real-Time Computing Systems and Applications (RTCSA'09), Boogerd, C. & Moonen, L. (2009) , pp. 321–328.
- [13] Kienle, H. M., Kraft, J. & Nolte, T. System-specific static code analyses: A case study in the complex embedded systems domain, Software Quality Journal,(2011).
- [14] Thomas, S.L., Van den Herrewegen, J., Vasilakis, G., Chen, Z., Ordean, M. and Garcia, F.D., 2021. Cutting Through the Complexity of Reverse Engineering Embedded Devices. IACR Transactions on Cryptographic Hardware and Embedded Systems, pp.360-389.
- [15] Li, X., Qiao, L., Sun, Y. and Guan, Q., 2020, July. Research on Automated Vulnerability Mining of Embedded System Firmware. In International Conference on

Blockchain for IOT Security & Privacy

Chaitali Choudhary, Dr. Inder Singh

Ph.D. Scholar, Assistant Professor (S.G.)
School of Computer Science, UPES Dehradun
chaitali.choudhary@gmail.com, inder@ddn.upes.ac.in

Abstract: IOT has gained lots of popularity now a days and there is tremendous increase in IOT sensory devices. IOT devices are initiating too much data such that in year 2020 IOT devices will be far more than total world's population in 2008 and it's expected to cross 50 billion by the year 2020. Block chain is the key technology which can help in solving various issues of Internet of Things. There are various issues in Internet of Things such as scaling up, privacy issues, and initiating a reliable network thus Blockchain is considered as one of the best technological innovations to get attached to IOT industry. Blockchain technique can be considered as an effective tool to secure vast number of Internet connected devices thus processing of transactions and compatibility between devices make considerable monetary benefits to IOT manufacturing industry and this technique would result in elimination of various failure nodes and a more reliable system for IOT machines. Various cryptographic technologies and algorithms used in blockchains will result in secure and reliable user data.

Keywords: Blockchain, IOT, Security, privacy, IOT devices

1. Introduction

Internet-of-Things (IoT) is a connection of different objects consisting of various electronics, items various software's and different kind of sensors. IoT gadgets are gaining growing popularity in electronics market during last few years. Increasingly more IT companies have added their IoT products line. According to end user's device usages there will be tremendous increase of IoT gadgets by 20.8 billion by 2020. Needless to say, it's a huge market and a field of promising technology innovation.

However, IoT devices suffers a lot of malicious attacks, either the conventional Internet attacks or the specific attacks targeted for IoT devices. Technologies have made lots of changes in our everyday life, particularly in today's world where everyone is so socially active and data-oriented (Minoli & Occhiogrosso, 2018). All of this is because of the technical upgradation in sensors and mass communication techniques which resulted in such a vast increment in count of devices connected and interacting to each other over a communication network which enables us with so many techniques to interact and share information between devices and humans (e.g., device-to-device).

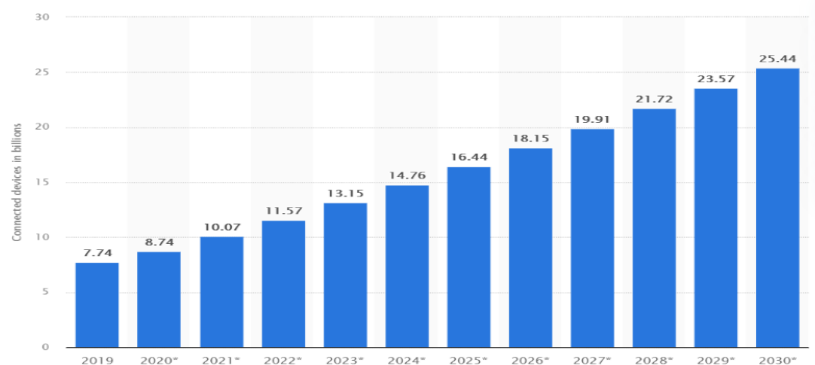


Fig 1. Number of Internet of Things (IoT) connected devices worldwide from 2019 to 2030(in billions)(Arne Holst, 2021)

Given the prevalence of such hand-held gadgets in our culture, data security and privacy are two of the most pressing issues. Data security and device security, as well as data privacy, are critical in data-sensitive situations like IoMT and IoBT. Malware creators and cyber attackers of today are creative and always looking

for new ways to get around current security measures (e.g., generating different versions of malware using mutation). Security and privacy remain significant problems for the Internet of Things (IoT), since IoT data is too large and dispersed across many nodes in the network (Zhu et al., 2019). Although blockchain-based methods offer decentralized security and anonymity, they come at a high cost in terms of energy, latency, and computing overhead, making them unsuitable for most resource-constrained IoT devices.

2. IOT attacks:

IoT is a very essential concept to end users and is very helpful in various dimensions of life but it also has lots of security challenges associated with it. One of the biggest concerns is to share data securely between trusted nodes. This trust factor means that data can be transferred between them on recognizing and affiliating their identity without any authentication issues and malware detection test (Alladi et al., 2019). The worst-case scenario is when your private data is leaked from your IoT devices is misused and can lead to various privacy issues and even to cases where life is endangered. Many different kinds of assaults have existed for a long time. What's new is the size and relative ease of assaults in the Internet of Things (IoT) - the millions of devices that might be hacked in the same way as conventional cyber-attacks, but on a much bigger scale and with little, if any, security. At its heart, the Internet of Things is all about connecting and networking things that haven't always been linked. This implies that all of those gadgets, whether it's your brand-new linked refrigerator or your connected car, are establishing a new network entry point, presenting a growing security and privacy risk (Rathee et al., 2021).

3. Block chain:

Blockchain technology has lately been a popular study subject among academic institutions and business organizations. Simply said, blockchain is a decentralized database with connected data blocks, each block containing a set of legitimate and digitally signed transactions. Nodes in a peer-to-peer network keep track of the blockchain itself. The use of a decentralized model in which apps may function effectively without the requirement for a central authority is what distinguishes blockchain. It allows for a trustless network in which members may transact despite their lack of trust for one another (Agrawal, R., Verma, P., Sonanis, R., Goel, U., De, A., Kondaveeti, S. A., & Shekhar, 2018). The Digital Cryptocurrency Bitcoin's backbone technology is Blockchain. The blockchain is a decentralized database that stores and distributes records of all transactions and digital events. Every transaction is double-checked by the system's majority of users. It has each transaction's complete record. The most well-known cryptocurrency and a blockchain example is Bitcoin (Narayanan & Clark, 2017). In 2008, a person or group of people known as 'Satoshi Nakamoto' released a white paper titled "Bitcoin: A peer to peer electronic currency system," which introduced blockchain technology to the public (Satoshi Nakamoto, 2008). Blockchain Technology creates an incorruptible digital ledger that records transactions and is disseminated throughout the network. Anything of value, such as land assets, automobiles, and other items, may be recorded as a transaction on Blockchain. Recently, blockchain technology has gotten a lot of press. It became well-known when Satoshi Nakamoto, a person or group of individuals who used the pseudonym Satoshi Nakamoto, introduced Bitcoin in 2009. Many people mix up blockchain with bitcoin and think it is the same thing. Bitcoin, on the other hand, is a blockchain-based application. Apart from payment systems, blockchain has a wide range of applications and use cases.

4. Blockchain for security:

An entry in the chain is generated when a piece of data is verified and then included in the chain. To add a piece of block, every blockchain end user must utilize their own key allotted to them as well as sophisticated computer equipment to solve highly complex mathematical calculations which is a part of algorithms. As soon as any user solve the complex mathematics problem they can add this new block to the existing chain, and the data contained inside it is permanently saved on the network which is immutable that means it cannot be changed or delete afterwards. When it comes to blockchain users, another characteristic is their capacity to upload entire all data of their network on the computer if needed (Atlam & Wills, 2020). A result of this is that two things happen: they may be able to earn money by renting out their "extra" storage space, and they may also be able to ensure that the chain does not go under completely. The whole system analyses each and every block of data in order to detect the one that is different from the others, for example, when someone

who doesn't own the data tries to manipulate a piece of data (such as a hacker) or tries to change or manipulate it (or from the majority). Any time the system comes across one of these blocks, it removes it from the existing blockchain and labels it as a fraud entry. As we know blockchain is constructed in this manner that there is no requirement of central administrator or a central storage center. Every individual user on the network makes a small but significant contribution to the blockchain's storage capacity. The responsibility for double-checking the data they store and/or disseminate rests with each individual to guarantee that no fake data is added and that no current data is removed is shared across the board. Blockchain technology is one of the most effective tools we have for protecting data from hackers, avoiding possible fraud, and reducing the risk of data loss or breach.

A hacker would have to delete all of the data contained on every computer linked to the global network to fully destroy or damage a blockchain. Millions of computers may be involved, with each computer keeping a duplicate of a portion or the whole of the information. Undamaged computers, referred to as "nodes," would continue to operate in order to check and keep track of all of the data on the network until the hacker was able to bring the whole network down at the same time as the rest of the network (which is almost difficult). The complexity of a task such as shutting down an entire chain increases as the number of users on a network gets larger. Because of the difficulties involved in joining a blockchain network, bigger blockchain networks with a greater number of users are at an exponentially lower risk of being targeted by hackers. We believe that blockchain technology has the potential to be the most secure way of storing and transferring data online that we have discovered so far, owing to its complex structure. The consequence is that entrepreneurs are beginning to use the technology in a number of sectors in order to fight fraud and enhance data security (Abhishek Dubey, Ayush Gupta, Nitish Raturi, 2018).

Guardtime has shown in the past that it is capable of protecting sensitive data via the use of blockchain technology. The company does away with the necessity for keys for verification purposes. As an alternative, they disseminate all of the data among all of the system's nodes. Any effort to tamper with the data will result in a thorough investigation, which will include a comparison of all chains with the metadata packet, and elimination of those that do not match. Therefore, the only way to completely wipe the blockchain clean is to destroy each and every node one at a time, as described above. No matter how many nodes are compromised, if just one of them continues to operate and provide correct data, the whole system may be restored to normal functioning. It is the goal of Guardtime's technology to detect and verify data changes at all times. There is no clear way of tampering with blocks in the chain as a result of this, and the data is thus safe. DDoS attacks are based on a simple yet powerful principle that has been proven time and time again. An attack by hackers may take a variety of forms, including sending a high number of spam requests to a website, which raises traffic to the point where the site can no longer cope with the volume of demands. After then, the attack will continue until the website gets overloaded and eventually collapses. DDoS (distributed denial of service) assaults have grown increasingly common in recent years, impacting large companies such as Twitter, Spotify, and SoundCloud, amongst others. There are issues with the current Domain Name System that make it difficult to avoid distributed denial of service (DDoS) assaults (DNS). In the Internet, the DNS (Domain Name System) serves as a phone book, providing a decentralized one-to-one mapping of IP addresses to domain names in a somewhat decentralized manner. Using this method, it is feasible to convert human-readable domain names (for example, steelkiwi.com) into machine-readable Internet Protocol addresses (IP addresses) (made up of numbers). However, despite the fact that DNS is only partially decentralized, hackers may still attempt to compromise the centralized component of DNS (which contains the majority of the data) and cause a chain reaction of website failures (Agrawal, R., Verma, P., Sonanis, R., Goel, U., De, A., Kondaveeti, S. A., & Shekhar, 2018).

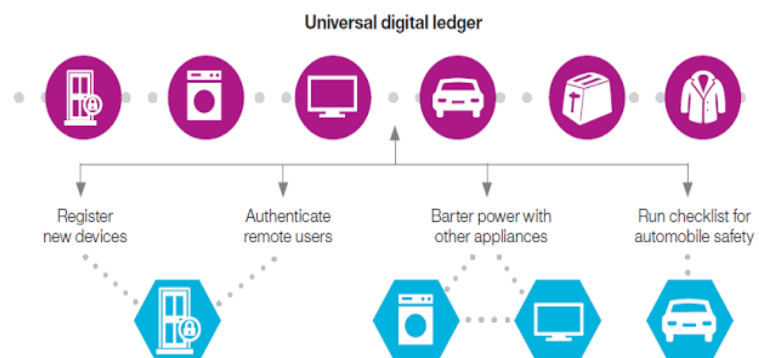


Fig 2. Blockchain for IoT transactions (Brody & Pureswaran, 2015)

5. Prevention of DDoS attack:

Decentralizing DNS via the use of blockchain technology would scatter information across many nodes, making it almost difficult for hackers to carry out attacks on the DNS server. The domain altering powers would be limited to the domain owners and would allow no other domain owners to get in on the action. Unauthorized third parties are much less likely to access or modify data with these safeguards in place. A blockchain system ensures the data remains impregnable to hackers until every node is wiped out, which is practically impossible. This area is using blockchain technology to avoid DDoS assaults since companies are using the blockchain in this sector. For example, the Blockstack offers a completely decentralized DNS service. The plan is to remove all third parties from web server, identity, and database management (Dorri et al., 2017). This would completely decentralize the internet (among other things). The blockchain may be used as a substitute for the existing DNS system and would allow everyone to register domain names, but only authorized people would be allowed to change the domains after they are registered. Because every other person possesses a copy of the blockchain's information, it would be difficult to hack or delete it entirely. The United Kingdom-based MaidSafe Business is another company in the same field. The project's goal is to "decentralise the internet" as they say, and do so in the service of creating a "new internet" that people can use. Users of this internet would store their data, run applications, and perform the other activities that they might usually do online but in a safer environment (Jayakumar et al., 2016). They may decide how much space on their own devices they want to provide to the network when they join up. The network's coin distribution is meant to compensate each user for the value they provide to the network. You encrypt and fragment all of your data, which is then shared across all of the MaidSafe network's members. Only the person who is in charge of managing data in their company may put it back in a form that can be read, preventing anybody else from seeing it.

6. IOT security using Blockchain:

Bitcoin (BTC) is a distributed database system that creates ledger entries that are very difficult to tamper with. Every transaction may be stored in immutable records, with each record being spread over a large number of participating nodes (Sharma et al., 2018). The use of strong public-key cryptography, a strong cryptographic hash, and full decentralization provide the necessary security. The technology is built on the idea of blocks. Small groups of transactions that have taken place inside the system are referred to as transactions groups. A hash of the prior transaction is included in each new block, which serves as a reference to the previous transaction. The term comes from the fact that it results in the creation of a 'chain' of blocks (Huh et al., 2017). A block's generation requires many specialized processors and a considerable length of time since it is computationally challenging to do so. The fact that it is difficult to generate a block and that, in order to tamper a single block, one must first tamper the previous block and then follow the chain in order to fully alter it, blockchain technology is regarded to be tamper resistant. Those who mine are those who use sophisticated computers to generate chunks of information.

7. Conclusion & Future Scope:

A truly autonomous smart gadget that can exchange data and even conduct financial transactions would be enabled by the use of blockchain technology, eliminating the need for a centralized intermediary. As a result of the distributed nature of the blockchain network, nodes may independently check the authenticity of a transaction without the need for a central authority to do the validation. An interesting capability provided by the blockchain is the capacity to keep a fully decentralized, trustworthy record of all transactions that take place inside a network of nodes. Without this capacity, it would be impossible to meet the many legal and regulatory requirements of industrial IoT applications without resorting to a centralized architecture.

References:

- [1] Abhishek Dubey, Ayush Gupta, Nitish Raturi, and P. S. (2018). *Springer Nature Singapore Pte Ltd. 2018, ICACCT* (Vol. 899). Springer Singapore. <https://doi.org/10.1007/978-981-13-2035-4>
- [2] Agrawal, R., Verma, P., Sonanis, R., Goel, U., De, A., Kondaveeti, S. A., & Shekhar, S. (2018). Continuous Security in IoT Using Blockchain. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6423–6427. <https://doi.org/10.1109/icassp.2018.8462513>
- [3] Alladi, T., Chamola, V., Parizi, R. M., & Choo, K. K. R. (2019). Blockchain Applications for Industry 4.0 and Industrial IoT: A Review. *IEEE Access*, 7, 176935–176951. <https://doi.org/10.1109/ACCESS.2019.2956748>
- [4] Arne Holst. (2021). *Number of Internet of Things (IoT) connected devices worldwide from 2019 to 2030(in billions)*. Aug 25. <https://www.statista.com/statistics/1183457/iot-connected-devices-worldwide/>
- [5] Atlam, H. F., & Wills, G. B. (2020). IoT Security, Privacy, Safety and Ethics. In *Internet of Things*. Springer International Publishing. https://doi.org/10.1007/978-3-030-18732-3_8
- [6] Brody, P., & Pureswaran, V. (2015). Device democracy: Saving the future of the Internet of Things - IBM Global Business Services Executive Report. *IBM Global Business Services Executive Report*, 3–25. <http://m2mworldnews.com/download/white-papers/IBM-Saving-the-future-of-IoT.pdf>
- [7] Dorri, A., Kanhere, S. S., Jurdak, R., & Gauravaram, P. (2017). Blockchain for IoT security and privacy: The case study of a smart home. *2017 IEEE International Conference on Pervasive Computing and Communications Workshops, PerCom Workshops 2017*, 618–623. <https://doi.org/10.1109/PERCOMW.2017.7917634>
- [8] Huh, S., Cho, S., & Kim, S. (2017). Managing IoT devices using blockchain platform. *International Conference on Advanced Communication Technology, ICACT*, 464–467. <https://doi.org/10.23919/ICACT.2017.7890132>
- [9] Jayakumar, H., Raha, A., Kim, Y., Sutar, S., Lee, W. S., & Raghunathan, V. (2016). Energy-efficient system design for IoT devices. *Proceedings of the Asia and South Pacific Design Automation Conference, ASP-DAC, 25-28-Janu*, 298–301. <https://doi.org/10.1109/ASPDAC.2016.7428027>
- [10] Minoli, D., & Occhiogrosso, B. (2018). Blockchain mechanisms for IoT security. *Internet of Things*, 1–2, 1–13. <https://doi.org/10.1016/j.iot.2018.05.002>
- [11] Narayanan, A., & Clark, J. (2017). Narayanan_Clark_2017_Bitcoin's Academic Pedigree. *Acmqueue*, 15(July-august), 1–30.
- [12] Rathee, G., Balasaraswathi, M., Chandran, K. P., Gupta, S. D., & Boopathi, C. S. (2021). A secure IoT sensors communication in industry 4.0 using blockchain technology. *Journal of Ambient Intelligence and Humanized Computing*, 12(1), 533–545. <https://doi.org/10.1007/s12652-020-02017-8>
- [13] Satoshi Nakamoto. (2008). *Bitcoin: A Peer-to-Peer Electronic Cash System*. https://doi.org/10.1162/ARTL_a_00247
- [14] Sharma, P. K., Chen, M. Y., & Park, J. H. (2018). A Software Defined Fog Node Based Distributed Blockchain Cloud Architecture for IoT. *IEEE Access*, 6(c), 115–124. <https://doi.org/10.1109/ACCESS.2017.2757955>
- [15] Zhu, L., Gai, K., & Li, M. (2019). Blockchain Technology in Internet of Things. In *Blockchain Technology in Internet of Things*. <https://doi.org/10.1007/978-3-030-21766-2>

Gradient Integrated Regression Sustainable Approach with Machine Learning towards Software Quality Assurance

Alankrita Aggarwal¹, Kanwalvir Singh Dhindsa², P. K. Suri³, Pardeep Singh⁴

¹PhD Research Scholar, Department of Computer Science and Engineering, IKGPTU,
Kapurthala Jalandhar-144603, Punjab, India, Orchid id: 0000-0002-0931-1118,
alankrita.agg@gmail.com

²Department of Computer Science and Engineering, Baba Banda Singh Bahadur Engineering
College, Fatehgarh Sahib-140406, Punjab, India, Orcid id: 0000-0002-7911-9244,
kdhindsa@gmail.com

³Department of Computer Science & Applications, Kurukshetra University, Kurukshetra 136038
Haryana, India, pksurikuk@gmail.com

⁴Assistant Professor (S.S.), Cybernetics, School of Computer Science, University of Petroleum and
Energy Studies, Dehradun -248001, pardeep.maan@gmail.com

Abstract : With the increase of enormous software suites and modules, there is a need to evaluate the cumulative quality and overall performance of newly developed modules using advanced algorithms and approaches including machine learning, statistical analysis, deep learning, and many others. In this work, the software quality assurance with the datasets evaluation using statistical analysis and regression integrated approach is presented. The aim is to acquire a much-enhanced acceptance of the subject matter by providing an investigation of software quality analysis using regression which is done in the last fifteen or twenty years.

Keywords: Software Quality Assurance, with Regression, Machine Learning, Gradient Integrated Regression

1. Introduction

Software service management and code quality with web-based applications and code evaluations is an important task in developing distributed applications that can be accessed via the Internet. As a part of it, software risk management is a prominent area in engineering science that involves historical detection, processing and risk management, and software creation vulnerabilities. In network engineering, software defect estimation is used to forecast software module deformity i.e. faults present during or after the distribution of the module. In this regard, the prediction methods for fault detection can be followed to assess faults in advance for the hold of good software [1][2].

The multiple dimensions of Regression testing of retesting software which is extended by new features during software development and the time taken in retesting the complete program may not be feasible with the amount of time and cost required so to overcome, a set of all test cases is executed for regression testing. The work is presenting the effectual outcome and performance-aware results and these can be used in the advanced predictions associated with the bugs and software quality management [3]

2. Literature review

Various researchers, scholars, and practitioners are working in enormous efforts towards operating with suggestive remarks on the study of the same domain, yet there is tremendous scope for progress as the methods, techniques, and paradigms need to be systematically analyzed.

Takenen et al.[4] suggested a component-based model which is doing Asset-Centric-based assessment and evaluation of all areas of risks and their relation with metric-based analysis of risk in software components.

Odzaly et al. [5] emphasize the creation of software programs and effort stresses incorporate data collection from a working project context to efficiently interpret data collection. In this, the combination of data

assimilated from live projects to highlight effectually by software agents for agile-based risk assessment analytics on data.

Dalla Palma et al. [6] have deployed an Infrastructure as Code (IaC) by offering automatic and configuration file infrastructure and tools to interact.

Zavvar M et al. [7] emphasize comparing and projected results on the basis of the soft computing approach for optimization of results classifying risk related to the software development phase threats. The method suggested is a soft calculation approach to optimization of outcomes using Support Vector Machine (SVM) and examines and region parameters of the Curve called Area Under Curve (AUC) and the Assignment Preciosity Rate (CAR). K-Means and findings can be correlated with Self Organizing Maps (SOM) and Classification Accuracy Rate (CAR)

Ghobadi S.& Mathiassen lars [8] presented a representation for the removal of risks in agile development using four agile projects and two software companies researched risk analysis and knowledge sharing in software development for certain issues in software development, usefulness and created a resolution strategy to minimize the risks. Outcomes raised the knowledge-sharing risk management for different project performances.

Mittal, M. et al.[22] presented a clustering approach which is a changed and updated single-pass clustering based on variable threshold method.

Juan Xuan. Et al. [23] illustrates the device fault forecasts and analytics using the data reduction-based software methodology. We derive characteristics from historic faulty data sets and create a statistical replica for the current data set to assess how instance choices and function selections can be implemented.

Mittal M. et al. [27] described the rising of big data technologies and cloud computing environments and comparison of spark and Hadoop and resulting of limitations of Hadoop over Spark can risk can be minimized by applying the same environments used by any kind of professionals

Abdelrafe M. S. Elzamly [28] depicted the fuzzy regression-based modeling methods with 49 risk factors related to software project development for risk analysis

Elzamly, B. et al.[29] uses an approach for the risk prediction called Linear Stepwise Discriminate Analysis for the identification of threats that affects the execution of software project to combine it for early prediction of software risks classified as low, high, and medium-class approaches for the risk control

Yucalar F. et al. [30] suggested a manual risk forecasting method by using automated predictors to aim at error-prone components so that engineers can focus on the faulty component.

3. Proposed Algorithm

A brief introduction of all the Statistical Methods is covered below which draws attention to the variety of statistical and machine learning methods used for quality and defect reduction. Statistical methods for software quality prediction can be done using regression analysis methods in which the values of any unidentified variable are predicted which are dependent on one or more known variables values.

Figure 1 is depicting the flow of work and the steps mentioned are an iterative process with an assortment of phases in which existing paradigms and technologies for Quality of Service (QoS) factors for the software services are studied and considered.

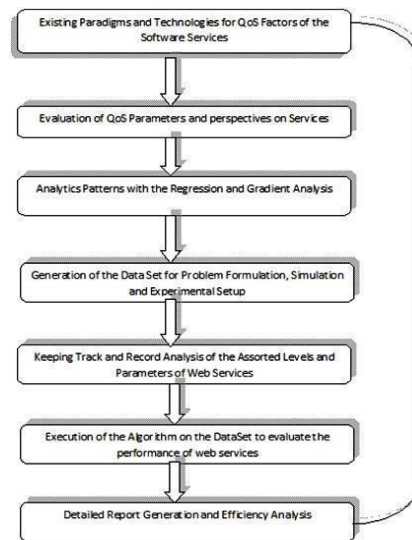


Figure 1: Algorithmic Flow of Work

4. Results and Analysis:

The work proposes a qualified revision of diverse quality prediction methods for error minimization based on any project and its related data sets. Then there are factors that are Evaluated on the basis of QoS Parameters and perspective on Services. After that, the patterns are analyzed the regression and Gradient Analysis methods. Following is the analytics of the simulation done on the datasets and thereby the evaluation patterns are presented as shown in table 1

Table 1: Performance Analytics on Stability Points

Iterations	Execution Time (In MicroSeconds)	Stability Point (Error Evaluation)
10	1.21	8
30	1.29	23
60	1.20	24
70	1.42	25

The training error is evaluated to analyze the overall patterns of the datasets under evaluations and presented in Figures 2.

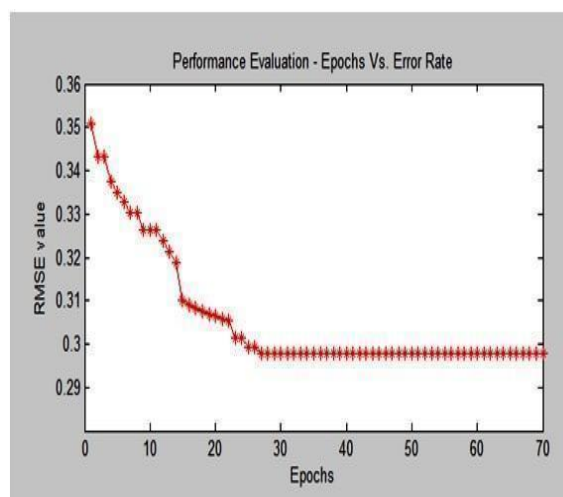


Figure 2: Epochs and Performance Evaluation

5. Conclusion & Future scope

The domain of software quality is quite prominent whereby the software suites are required to be rigorously tested. Earlier statistical methods were deployed for doing the process but present work requires a more dependable approach so that the modeling for the prediction of software quality can be done. The application of statistical and machine learning usage in error forecasting requires a colossal quantity of statistics and analyzing this massive data is essential with the help of a better forecast model.

References

- [1] Ali, M. M., Huda, S., Abawajy, J., Alyahya, S., Al-Dossari, H., & Yearwood, J. (2017). A parallel framework for software defect detection and metric selection on cloud computing. *Cluster Computing*, 20(3), 2267-2281.
- [2] Lewis, W. E. (2017). *Software testing and continuous quality improvement*. CRC press.
- [3] Goyal, L. M., Mittal, M., & Sethi, J. K. (2016). Fuzzy model generation using Subtractive and Fuzzy C-Means clustering. *CSI transactions on ICT*, 4(2-4), 129-133.
- [4] Takanen, A., Demott, J. D., Miller, C., & Kettunen, A. (2018). *Fuzzing for software security testing and quality assurance*. Artech House.
- [5] Odzaly, E. E., Greer, D., & Stewart, D. (2018). Agile risk management using software agents. *Journal of Ambient Intelligence and Humanized Computing*, 9(3), 823-841.
- [6] Palma, S. D., Di Nucci, D., Palomba, F., & Tamburri, D. A. (2020). Towards a Catalogue of Software Quality Metrics for Infrastructure Code. *arXiv preprint arXiv:2005.13474*.
- [7] Zavvar, M., Yavari, A., Mirhassannia, S. M., Nehi, M. R., Yanpi, A., & Zavvar, M. H. (2017). Classification of risk in software development projects using support vector machine. *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, 9(1), 1-5.
- [8] Ghobadi, S., & Mathiassen, L. (2017). Risks to effective knowledge sharing in agile software teams: a model for assessing and mitigating risks. *Information systems journal*, 27(6), 699-731.
- [9] Lu, F., Bi, H., Huang, M., & Duan, S. (2017). Simulated annealing genetic algorithm based schedule risk management of IT outsourcing project. *Mathematical Problems in Engineering*, 2017.
- [10] Méndez Fernández, D., Tießler, M., Kalinowski, M., Felderer, M., & Kuhrmann, M. (2017). On Evidence-based Risk Management in Requirements Engineering. *arXiv*, arXiv-1707.
- [11] Rauter, T., Höller, A., Kajtazovic, N., & Kreiner, C. (2016, January). Asset-Centric Security Risk Assessment of Software Components. In *MILS@ HiPEAC*.
- [12] Elzamly, A., & Hussin, B. (2014). Managing software project risks (analysis phase) with proposed fuzzy regression analysis modelling techniques with fuzzy concepts. *Journal of computing and information technology*, 22(2), 131-144.
- [13] B. C. Krishna and K. Subrahmanyam (2016), "A Decision Support System for Assessing risk using Halstead approach and Principal Component Analysis," *Journal of chemical and pharmaceutical sciences*, 9(4): 3383-3387.
- [14] Purandare, P. (2016). An entropy based approach for risk factor analysis in a software development project. *International Journal of Applied Engineering Research*, 11(4), 2258-2262.
- [15] Choetkiertikul, M., Dam, H. K., Tran, T., & Ghose, A. (2015, November). Predicting delays in software projects using networked classification (t). In *2015 30th IEEE/ACM International Conference on Automated Software Engineering (ASE)* (pp. 353-364). IEEE.
- [16] Elzamly, A., Hussin, B., Abu-Naser, S. S., & Doheir, M. (2015). Predicting Software Analysis Process Risks Using Linear Stepwise Discriminant Analysis: Statistical Methods. *Int. J. Adv. Inf. Sci. Technol.*, Vol. 38, No. 38, page(s): 108-115, June 2015.
- [17] Mohanty, R., & Ravi, V. (2017). Machine Learning Techniques to Predict Software Defect. In *Artificial Intelligence: Concepts, Methodologies, Tools, and Applications* (pp. 1473-1487). IGI Global.
- [18] Singh, R., Raja, R., & Chopra, J. (2017). Software Defect Prediction Using Averaging Likelihood Ensemble Technique. *International journal*, 4, 213-223.
- [19] Ali, M. M., Huda, S., Abawajy, J., Alyahya, S., Al-Dossari, H., & Yearwood, J. (2017). A parallel framework for software defect detection and metric selection on cloud computing. *Cluster Computing*, 20(3), 2267-2281.
- [20] Osman, H., Ghafari, M., & Nierstrasz, O. (2017, February). Automatic feature selection by regularization to improve bug prediction accuracy. In *2017 IEEE Workshop on Machine Learning Techniques for Software Quality Evaluation (MaLTesQuE)* (pp. 27-32). IEEE.
- [21] Maddipati, S. S., Pradeepini, G., & Yesubabu, A. (2018). Software Defect Prediction using Adaptive Neuro Fuzzy Inference System. *International Journal of Applied Engineering Research*, 13(1), 394-397.

- [22] Mittal, M., Sharma, R. K., & Singh, V. P. (2015). Modified single pass clustering with variable threshold approach. *International Journal of Innovative Computing Information and Control*, 11(1), 375-386.
- [23] Xuan, J., Jiang, H., Hu, Y., Ren, Z., Zou, W., Luo, Z., & Wu, X. (2014). Towards effective bug triage with software data reduction techniques. *IEEE transactions on knowledge and data engineering*, 27(1), 264-280.
- [24] Neufelder, A. M. (2018). *Ensuring software reliability*. CRC Press.
- [25] Van Deursen, A., Aniche, M., Boone, C., Cunha, M. L., & Nadeem, A. (2019). Software Quality and Testing. *Delft University of Technology*.
- [26] Leicht, N., Blohm, I., & Leimeister, J. M. (2017). Leveraging the power of the crowd for software testing. *IEEE Software*, 34(2), 62-69.
- [27] Mittal, M., Balas, V. E., Goyal, L. M., & Kumar, R. (Eds.). (2019). *Big data processing using spark in cloud*. Springer.
- [28] Abdelrafe MS. Managing Software Project Risks Using Stepwise And Fuzzy Regression Analysis Modeling Techniques, January 2016.
- [29] Elzamly, B. Hussin, S. S. A. Naser, and M. Doheir, "Predicting Software Analysis Process Risks Using Linear Stepwise Discriminant Analysis : Statistical Methods," *Int. J. Adv. Inf. Sci. Technol.*, Vol. 38, No. 38, page(s): 108–115, June 2015.
- [30] Yucalar, F., Ozcift, A., Borandag, E., & Kilinc, D. (2020). Multiple-classifiers in software quality engineering: Combining predictors to improve software fault prediction ability. *Engineering Science and Technology, an International Journal*, 23(4), 938-950.
- [31] Rothmel, G. (2018, November). Improving regression testing in continuous integration development environments (keynote). In *Proceedings of the 9th ACM SIGSOFT International Workshop on Automating TEST Case Design, Selection, and Evaluation* (pp. 1-1). ACM.
- [32] Aggarwal, A., Dhindsa, K. S., & Suri, P. K. (2021). Performance-Aware Approach for Software Risk Management Using Random Forest Algorithm. *International Journal of Software Innovation (IJSI)*, 9(1), 12-19.
- [33] Aggarwal, A., Dhindsa, K. S., & Suri, P. K. (2020). Design for Software Risk Management Using Soft Computing and Simulated Biological Approach. *International Journal of Security and Privacy in Pervasive Computing (IJSPPC)*, 12(2), 44-54.
- [34] Escalated Methods for Software Defects Audit in Repercussion and Effects Construe to Nature Inspired and Behavior Driven Mechanisms. (2019). Regular Issue, 8(6), 1779–1783. doi:10.35940/ijeat.f8442.088619
- [35] Usage Patterns and Implementation of Random Forest Methods for Software Risk and Bugs Predictions. (2019). Special Issue, 8(9S), 927–932. doi:10.35940/ijitee.i1150.0789s19
- [36] Aggarwal, A., Dhindsa, K. S., & Suri, P. K. (2020). Design for Software Risk Management Using Soft Computing and Simulated Biological Approach. *International Journal of Security and Privacy in Pervasive Computing (IJSPPC)*, 12(2), 44-54
- [37] Aggarwal, A., Dhindsa, K. S., & Suri, P. K. (2020). An Empirical Evaluation of Assorted Risk Management Models and Frameworks in Software Development. *International Journal of Applied Evolutionary Computation (IAEC)*, 11(1), 52-62.
- [38] Aggarwal, A., Dhindsa, K. S., & Suri, P. K. (2020). A Pragmatic Assessment of Approaches and Paradigms in Software Risk Management Frameworks. *International Journal of Natural Computing Research (IJNCR)*, 9(1), 13-26.
- [39] Podgorny, I. A., Cessna, J., Gielow, C. C., & Cannon, M. (2018). U.S. Patent No. 10,162,734. Washington, DC: U.S. Patent and Trademark Office.
- [40] Mittal, M., Sharma, R. K., & Singh, V. P. (2014). Validation of k-means and threshold based clustering method. *International Journal of Advancements in Technology*, 5(2), 153-160
- [41] Felderer, M., & Fourmeret, E. (2015). A systematic classification of security regression testing approaches. *International Journal on Software Tools for Technology Transfer*, 17(3), 305-319.

COVID-19 Diagnosis: An Eye of Machine Learning Techniques

Anshul^{#1,2}, Raju Kumar^{*3}

^{#1}Research Scholar,

Chandigarh University, Gharuan, Mohali, Punjab, India

^{#2} Assistant Professor,

Chandigarh Group of Colleges, Mohali, Punjab, India

Email: mca.anshulgarg@gmail.com

^{*3}Associate Professor,

Chandigarh University, Gharuan, Mohali, Punjab, India

Email: rajuk12@gmail.com

Abstract- Coronavirus infection has a place with the Severe Acute Respiratory Disorder (SARS) family elevated a circumstance of well-being crisis in practically every one of the nations of the world. The COVID battle to control the promulgate of the coronavirus includes most states, organizations, and research and development Institutions. Various AI and machine learning-based methods are utilized to analyze COVID patients utilizing diverse image modalities alike CT SCAN, X-RAY, and so on. This paper gives the works made in COVID diagnosis, the job of machine learning and deep learning -based strategies to take care of this issue, and presents restrictions and future directions as for diagnosis of coronavirus.

Keywords: Machine Learning, Deep Learning, COVID-19, CT SCAN, X-RAY. Artificial Intelligence.

1. Introduction

COVID-19 has announced a epidemic in March, 2020 [1]. It was first perceived in China in late 2019 This infection is reaching out across the globe, creating monstrous cultural commotion and extending our capacity to convey proficient healthcare. On the planet, till 5 August 2021 complete affirmed cases were 201,876,177 in which deaths were 4,284,408 and the tally is expanding step by step [2]. The rushed spread of this infection all through the world is brought about by the shortage of information about this infection alongside an absence of antiviral treatments. Early contamination manifestations incorporate hack, fever, and other respiratory issues. Cold and cough are assuming a significant part in sending the Covid from one human to another

As the disease heightens, the effected person needs to confront astringent respiratory conditions, kidney failure, and lung issues, etc. A few patients begin losing their platelets and had uncontrollable body temperature. The circumstance is deteriorating as till time there is no viable medication is there to control this issue. The healthcare business is energetically searching for innovations also, methods to track the expansion of the coronavirus in this worldwide wellbeing emergency. One of the most uses worldwide innovations right currently is Artificial Intelligence (AI) that can follow the speed and distinguish the development pace of the Covid, recognize the danger and seriousness of Coronavirus patients. Computer-based intelligence can likewise expect death chances by satisfactorily breaking down past information. AI can help us with doing combating the infection by testing people, clinical help, information and data, and proposals about infectious prevention [3]. The following flowchart shows the different areas in which machine learning and deep learning are showing robust results.

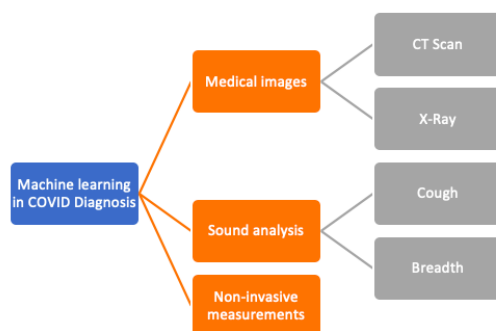


Fig 1: Machine learning in COVID diagnosis

2. Related Work

This section shows the current work in the field of Corona virus detection. The literature work shows that deep learning models are exposing encouraging results in the field of Corona virus diagnosis.

Table 1. Analysis on Corona Diagnosis

Sr. No	Paper References	Methodology	Dataset	Results
1.	Islam et al., 2020[1]	CNN+LS TM	4575 images	AUC: 99.9% Sensitivity : 99.3% F1-score: 98.9%. Accuracy : 99.4% Specificity : 99.2%,
2.	Saha et al., 2021 [3]	EMCNet	2460 X-Ray images approx	Accuracy: 98.91% Precision: 100% Recall: 97.82% F1-Score: 98.89%
3.	Barstugan et al., 2020 [4]	SVM	150 images (Patch size: 16x16, 32x32, 48x48, 64x64) which belong to the 53 infected cases	Accuracy- 99.08%
4.	Farooq & Hafeez, 2020 [5]	ResNet	5941 radiography images of 2839 patients with 4 classes.	Accuracy: 96.23%
5.	Pham, 2020 [6]	CNN	349 CT Images(216 Covid patients+ 397(non - covid))	Accuracy is best in the case of DenseNet.
6.	Rahimzadeh & Attar, 2020 [7]	DC NN	Total images 11302	Accuracy: 91.4%
7.	Y. Wang et al., 2020 [8]	BI-AT-GRU Simulation model RSM	Not mentioned	Accuracy- 94.5%, Precision- 94.4% Recall-95.1% and F1- 94.8%

8.	Han et al., 2020[9]	AD3D- MIL	460 CT Scan: 230(79 COVI D-Patient +100 (pneumonia) +130(Normal)	overall accuracy: 97.9%, AUC: 99.0%, Cohen kappa:95.7%.
9.	Caruso et al., 2020[10]	RT-PCR	158 (83(M)+75(F)	Sensitivity:97% Specificity:56% Accuracy: 72%
10.	Hou et al., 2020 [11].	COVNet	4356 chest CT Images	Sensitivity: 90 Specificity : 96 AUC: AUC of 0.96
11.	Xu, Jiang, Ma, Du, Li, Lv, Yu, et al,2020[12]	CNN	618 CT images	Accuracy: 86.7%.
12.	S. Wang et al., 2021[13]	CNN	453 CT (COVID-19 Positive)	Sensitivity: 84% Specificity :80.5% Accuracy: 82.9%
13.	Minaee et al., 2020 [14]	CNN	5,000 X-rays	Sensitivity: 98%, specificity: 90%
14.	Alqudah et al., 2020 [15]	CNN +SVM	310 (from Cohen) + 310 from Kaggle	Accuracy:98.71 Sensitivity: 98.71 Specificity:99.35 Precision:98.73 F1-Score:98.71 MCC: 0.9808
15.	Medhi et al., 2020 [16]	CNN	150 COVI D-19 confirmed cases using Kaggle	Accuracy: 93%
16.	Asif & Wenhui, 2020 [17]	CNN	864 (COVID infected images+ 1345 (pneumonia) + 1341(Normal)	96% accuracy
17.	El-Din Hemdan et al., n.d. [18]	COVIDX -Net	50 X-Ray(25 Covid)	F1 Score By: VGG19 :0.89 DenseNet: 0.91
18.	Jaiswal et al., 2020 [19]	DenseNet 201	Kaggle	Accuracy: 97%
19.	Sun et al., 2020[20]	SVM	336 (COVID positive) and 220 clinical laboratory observations /records	recall : 93.33(training) and 100 % (testing datasets).

20.	Li et al., 2020 [21]	COVIDM obileXper t	179 CXR images(Nor mal)+179(P neumonia)+ 179 (Covid)	classification accuracy: 0.843, AUROC : 0.97
21.	Gunraj et al., 2020 [22]	COVID-Net	13,975 CXR (13,870 patient cases)	Accuracy:93.3% Sensitivity : 91.0 PPT: 98.9%
22.	Zheng et al., 2020 [23]	DeCoVN et	499 CT Images for training + 131 CT for testing	Accuracy :0.901, a PPV positive predictive :0.840 Negative predictive value: 0.982
23.	Chen et al., 2020 [24]	DNN	46,096 anonymous images	Per Patient:- sensitivity :100%, specificity :93.55%, accuracy: 95.24%, PPV :84.62%, NPV :100%; Per-image :- Sensitivity: 94.34 Specificity: 99.16% Acc. :98.85% NPV :99.61% PPV :88.37%
24.	Karim et al., 2020 [25]	DeepCOVID Explainer	16,995 CXR images	PPV -89.61% Recall -83%,
25.	Sethy et al., 2020 [26]	CNN+SV M	GitHub and Kaggle	Accuracy : 95.33, Sensitivity:95.33%, FPR:2.33% F1 score :95.34
26.	Ucar & Korkmaz, 2020 [27]	COVIDia gnosis- Net	Covidx	Accuracy:98.26% COR:98.26 COM:98.26 SPE:99.13 MCC:97.39 F1:98.25
27.	Narayan Das et al., 2020 [28]	Xception	X-ray images	Accuracy: 99.25% F-Measure:98.6% Sensitivity:99.1% Specificity:99.4% Kappa statics:98.04%

3. Artificial Intelligence Applications in Diagnosis of Corona

AI and profound learning models hold incredible guarantee in COVID diagnostics. Many AI applications are utilized to screen, forestall and treat viral contaminations. The fundamental issue with this infection is that it voyages so rapidly from one individual to another that early recognition of the infection can save many lives. AI indicative apparatuses are especially exact in recognizing the presence of infections utilizing X-beam pictures and can function as an appraisal device. In clinics and other wellbeing frameworks during this pandemic, AI devices are a superior other option, as radiologists are costly and not accessible all over the place. These instruments utilize different imaging modes like CT SCAN, X RAY, and CBX. Notwithstanding, most analysts like to utilize CT and X-beams [27]. AI devices can likewise assist clinicians with rapidly distinguishing patients experiencing other medical conditions.

Fake hereditary insight plays a critical obligation in clinical picture investigation. The ML models can comfort the local area in an assortment of ways, including early admonition and cautioning, illness guess and conclusion, checking and expectation, treatment and care, and social control [8]. Canadian man-made reasoning model BlueDot tried its mental fortitude and advised clients about the flare-up before WHO. In like manner, another AI model, Healthmap, predicts the beginning of COVID 19. The ML model additionally adds to medicate disclosure Covid19. Scientists are utilizing computerized reasoning models that can accelerate research for COVID19 immunizations and medicines. Google's DeepMind, a profound learning apparatus that predicts the design of the COVID19 protein, uncovers the COVID19 antibody. These models additionally work in the field of contact following through portable applications. Table 2 shows a portion of the versatile applications utilized in various nations.

Table 2: Application of Contact Tracing [29]

Sl. No	Country	Contact tracing App	Location tracking	Launch on
1	Australia	COVIDSafe	BlueTrace protocol: Bluetooth	April 14, 2020
2	Austria	Stopp Corona	Bluetooth, Google/Apple	March, 2020
3	Bahrain	BeAware Bahrain	Bluetooth & GSM	March 31, 2020
4	Bulgaria	VirusSafe	GSM	May, 2020
5	China	conjunction with Alipay	GPS, GSM, credit-card-transaction-history	Very little Info
6	Cyprus	CovTracer	GPS, GSM	May, 2020
7	Colombia	CoronApp	GPS	April 12, 2020
8	Czech Republic	eRouška (eFacemask)	BlueTrace protocol: Bluetooth	April 15, 2020
9	Estonia	Estonia's App	Google/Apple, DP-3T, Bluetooth	April, 2020
10	Finland	Ketju	DP-3T, Bluetooth	May, 2020
11	France	StopCovid	Bluetooth	May, 2020
12	Germany	CoronaApp	Bluetooth, Google/Apple	May, 2020
13	Ghana	GH Covid-19 Tracker App	GPS	April 12, 2020
14	Hungary	VirusRadar	Bluetooth	May 13, 2020
15	Iceland	Rakning C-19	GPS	Apr-20
16	India	Aarogya Setu	Bluetooth & location-generated social grap	April 2, 2020
17	Iran	Mask.ir	GSM	May, 2020
18	Ireland	HSE Covid-19 App	Bluetooth, Google/Apple	May, 2020
19	Israel	HaMagen	Standard location APIs	March, 2020
20	Italy	Immuni	Bluetooth, Google/Apple	May, 2020
21	Jordan	AMAN App - Jordan	GPS	May, 2020
22	Latvia	Apturi Covid	Bluetooth	May, 29, 2020
23	Malaysia	MyTrace	Bluetooth, Google/Apple	May 3, 2020
24	Mexico	CovidRadar	Bluetooth	May, 2020
25	New Zealand	NZ COVID Tracer	Contact details and physical address	May 20, 2020

4. Discussion and Conclusion

COVID pandemic started in March 2020 and till time first wave and second wave already came. A lot of loss of human lives has already been done throughout the world. In many countries, the third wave also started. Table 3 shows the most affected countries due to corona, the data is based on the worldometer accessed on August 7, 2021.

Table 3. Top 10 Country Wise Analysis for CORONA -19[2]

Country	Total Confirmed Cases	Total Death
USA	36,176,471	631,299
India	31,812,114	426,321
Brazil	20,026,533	559,715
Russia	6,379,904	162,509
France	6,207,416	112,046
UK	5,952,796	130000
Turkey	5,822,487	51,767
Argentina	4,975,616	106,747
Colombia	4,815,063	121,695

Below graphical illustration of the topmost 10 countries which are affected by corona. till August 7, 2021.

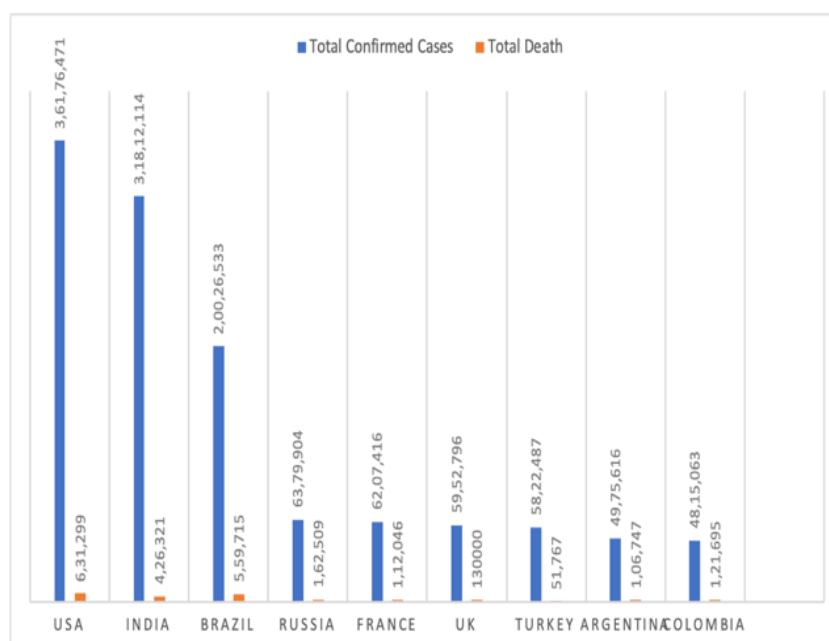


Fig 2 Corona Death Analysis[2]

The effects of the pandemic have led researchers to make a large number of discoveries to understand, test, and prevent the disease.

Islam et al [1] used CNN with LSTM showing all results above or equal to 99% in terms of AUC, F1-Score, accuracy and specificity but to improve results larger dataset and different angles of X- Ray images can be used. Other diseases except can also be classified using this method.

Saha et al.[3] used another deep learning model named EMCNet using 2460 X- ray images showing prominent results in terms of different parameters like accuracy, precision, recall and F1-Score. Barstugan et al. [4] used machine learning method SVM to accurately diagnose the coronavirus . In both papers [3] [4] multiclassification can be done using larger dataset. Farooq et al. [5] , Pham et al [6] Rahimzadeh et al [7] ,Hou et al [11], Xu et al. [12] , Wang et al [13] and many other researchers used CNN and its various versions to detect the coronavirus. and help the doctors to save the human lives but still there is lot of room for improvement as different variants of corona are here to threaten the lives of humans . So more dataset is required such that more lives can be saved.

In any case, it remains ambiguous whether the studies provided so far adequately fill the existing information gaps. The disease can be confirmed using the RT-PCR test, but detection by RTPCR is a time-consuming task. The ML and DL methods that allow automatic identification of COVID 19 are very useful for examining the quantification and classification of subjective non-contact interactions [2]. For fast and accurate results, DL technology for the ancestors of COVID 19 graphics capabilities by CT SCAN and X-ray imaging methods. CT scans are slightly more advanced than X-rays and can use machine learning to deliver more accurate results for tissues and organs of the body, but are more expensive than x-rays [3]. The X-Ray is economical than CT but can lead to false results. Although AI methods have always been shown to be promising in the healthcare industry, there are still many open challenges. ML and DL are very good at detecting COVID19, but DL shows better results than ML [5]. Convolutional neural networks perform best in this area so far, but the problem is that deep learning-based models require very large amounts of data and are tagged as training recommendations. ResNet shows the best results ever with 98% accuracy [5]. Troubleshooting the dataset allows you to perform data expansion, but the results can be affected by the gain. Machine learning also plays an effective role in monitoring drug discovery, severity, and improvement of COVID treatment, despite the early diagnosis. The COVID-19 AI algorithm was developed using a variety of applications, data collection processes, and performance metrics [5]. Therefore, there is probably no 100% effective drug that can be clinically distributed at this time. Possible reasons are :

- (i) Small distortion of the dataset and large variability of the dataset of international origin.
- (ii) Insufficient integration of multithreaded data, especially image data.

ML and DL have shown promise in solving the COVID problem, but this problem is difficult to combat due to unknown biological origins and their mysterious behavior. The main symptoms of coronavirus infection are thought to be fever, cold, cough, and difficulty breathing, but you can get infected without these symptoms [10], [24] Certain precautions such as wearing a face cover, hygiene, isolation and social distancing can condense the risk of a pandemic epidemic [20] The future face of AI-related to COVID19 diagnostics is to test and train various ML and DL architectures using diverse and widespread image datasets with unique nuances of the disease. For a clearer diagnosis of COVID19, a multi-classification method is needed along with statistics.

References

- [1] M. Z. Islam, M. M. Islam, and A. Asraf, "A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images," *Informatics in Medicine Unlocked*, vol. 20, no. August, p. 100412, 2020, doi: 10.1016/j.imu.2020.100412.
- [2] Worldometer, "COVID-19 CORONAVIRUS PANDEMIC," 2021. <https://www.worldometers.info/coronavirus> (accessed August 7 , 2021).
- [3] P. Saha, M. S. Sadi, and M. M. Islam, "EMCNet: Automated COVID-19 diagnosis from X-ray images using convolutional neural network and ensemble of machine learning classifiers," *Informatics in Medicine Unlocked*, vol. 22, no. December, p. 100505, 2021, doi: 10.1016/j.imu.2020.100505.
- [4] M. Barstugan, U. Ozkaya, and S. Ozturk, "Coronavirus (COVID-19) Classification using CT Images by Machine Learning Methods," *arXiv*, no. 5, pp. 1–10, 2020.
- [5] M. Farooq and A. Hafeez, "COVID-ResNet: A deep learning framework for screening of COVID19 from radiographs," *arXiv*, 2020.
- [6] T. D. Pham, "Classification of COVID-19 chest X-rays with deep learning: new models or fine tuning?," *Health Information Science and Systems*, vol. 9, no. 1, 2021, doi: 10.1007/s13755-020-00135-3.
- [7] M. Rahimzadeh and A. Attar, "a New Modified Deep Convolutional Neural Network for Detecting Covid-19 From X-Ray Images," *arXiv*, 2020, doi: 10.1016/j.imu.2020.100360.
- [8] Y. Wang, M. Hu, Q. Li, X.-P. Zhang, G. Zhai, and N. Yao, "Abnormal respiratory patterns classifier may contribute to large-scale screening of people infected with COVID-19 in an accurate and unobtrusive manner," 2020, [Online]. Available: <http://arxiv.org/abs/2002.05534>
- [9] Z. Han et al., "Accurate Screening of COVID-19 Using Attention-Based Deep 3D Multiple Instance Learning," *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2584–2594, 2020, doi: 10.1109/TMI.2020.2996256.
- [10] D. Caruso et al., "Chest CT Features of COVID-19 in Rome, Italy," *Radiology*, vol. 296, no. 2, pp. E79–E85, 2020, doi: 10.1148/radiol.202021237.

- [11] H. Hou et al., "Artificial Intelligence Distinguishes COVID-19 from Community Acquired Pneu-monia on Chest CT," *Applied Intelligence*, vol. 2019, pp. 1–5, 2020, [Online]. Available: <http://arxiv.org/abs/2003.13865>
- [12] X. Xu et al., "Deep learning system to screen coronavirus disease 2019 pneumonia," *arXiv*, pp. 1–29, 2020.
- [13] S. Wang et al., "A deep learning algorithm using CT images to screen for Corona virus disease (COVID-19)," *European Radiology*, pp. 1–19, 2021, doi: 10.1007/s00330-021-07715-1.
- [14] S. Minaee, R. Kafieh, M. Sonka, S. Yazdani, and G. J. Soufi, "Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning," *arXiv*, pp. 1–9, 2020.
- [15] A. M. Alqudah, S. Qazan, and A. Alqudah, "Automated Systems for Detection of COVID-19 Using Chest X-ray Images and Lightweight Convolutional Neural Networks," 2020, doi: 10.21203/rs.3.rs-24305/v1.
- [16] K. Medhi, M. Jamil, and M. I. Hussain, "Automatic Detection of COVID-19 Infection from Chest X-ray using Deep Learning," *medRxiv*, pp. 1–6, 2020, doi: 10.1101/2020.05.10.20097063.
- [17] S. Asif and Y. Wenhui, "Automatic Detection of COVID-19 Using X-ray Images with Deep Convolutional Neural Networks and Machine Learning," *medRxiv*, p. 2020.05.01.20088211, 2020.
- [18] E. El-Din Hemdan, M. A. Shouman, and M. E. Karar, "COVIDX-Net: A Framework of Deep Learning Classifiers to Diagnose COVID-19 in X-Ray Images."
- [19] A. Jaiswal, N. Gianchandani, D. Singh, V. Kumar, and M. Kaur, "Classification of the COVID-19 infected patients using DenseNet201 based deep transfer learning," *Journal of Biomolecular Structure and Dynamics*, vol. 1102, 2020, doi: 10.1080/07391102.2020.1788642.
- [20] L. Sun et al., "Combination of four clinical indicators predicts the severe/critical symptom of patients infected COVID-19," *Journal of Clinical Virology*, vol. 128, no. March, p. 104431, 2020, doi: 10.1016/j.jcv.2020.104431.
- [21] X. Li, C. Li, and D. Zhu, "COVID-MobileXpert: On-Device COVID-19 Patient Triage and Follow-up using Chest X-rays," Apr. 2020, [Online]. Available: <http://arxiv.org/abs/2004.03042>
- [22] H. Gunraj, L. Wang, and A. Wong, "COVIDNet-CT: A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases From Chest CT Images," *Frontiers in Medicine*, vol. 7, no. 0123456789, pp. 1–12, 2020, doi: 10.3389/fmed.2020.608525.
- [23] C. Zheng et al., "Deep Learning-based Detection for COVID-19 from Chest CT using Weak Label," *medRxiv*, pp. 1–13, 2020, doi: 10.1101/2020.03.12.20027185.
- [24] J. Chen et al., "Deep learning-based model for detecting 2019 novel coronavirus pneumonia on high-resolution computed tomography: a prospective study," *medRxiv*, 2020, doi: 10.1101/2020.02.25.20021568.
- [25] M. R. Karim, T. Döhmen, D. Rebholz-Schuhmann, S. Decker, M. Cochez, and O. Beyan, "DeepCOVIDExplainer: Explainable COVID-19 Predictions Based on Chest X-ray Images," *arXiv*, 2020.
- [26] P. Kumar Sethy and S. Kumari Behera, "Detection of coronavirus Disease (COVID-19) based on Deep Features," 2020, doi: 10.20944/preprints202003.0300.v1.
- [27] F. Ucar and D. Korkmaz, "COVIDagnosis-Net: Deep Bayes-SqueezeNet based diagnosis of the coronavirus disease 2019 (COVID-19) from X-ray images," *Medical Hypotheses*, vol. 140, no. April, p. 109761, 2020, doi: 10.1016/j.mehy.2020.109761.
- [28] N. Narayan Das, N. Kumar, M. Kaur, V. Kumar, and D. Singh, "Automated Deep Transfer Learning-Based Approach for Detection of COVID-19 Infection in Chest X-rays," *Irbm*, 2020, doi: 10.1016/j.irbm.2020.07.001.
- [29] S. Lalmuanawma, J. Hussain, and L. Chhakchhuak, "Applications of machine learning and artificial intelligence for Covid-19 (SARS-CoV-2) pandemic: A review," *CHAOS SOLITONS & FRACTALS*, vol. 139, Oct. 2020, doi: 10.1016/j.chaos.2020.110059.

Software reliability model with FSM: A Study

Rahul Kumar Singh¹, Pardeep Singh², Gourav Bathla³ and Alankrita Aggrawal⁴

^{1,2,3} School of Computer Science, University of Petroleum and Energy Studies, Dehradun, Uttarakhand
248001, India

⁴ Department of Computer Science, PIET, Panipat, Haryana, India
pardeep.maan@gmail.com

Abstract. Software reliability refers to the capacity of software to operate without fail in a variety of environments and conditions. In this paper two approaches are considered for measuring the software reliability which shows some effective results for each approach. Prediction model can be considered under the traditional approach. The research has considered the data set for measuring the reliability using the Goel-Okumoto model the same data set is applied through the FSM to obtain an optimized results. Analyzing those results, the performance of FSM is better than the traditional approach.

Keywords: Software reliability, MTTF, MTTR, MTBF, finite state automata.

1 Introduction

Software reliability means failure free operation under different environments or conditions. Reliability of software shows its quality. There are some phases of software development life cycle feasibility study, design, coding, testing, implementation, maintenance by following that cycle the new product/software is developed. In this process there are some errors of different types. The probability of reliability depends upon the error. There is no software which has 100% reliability or the probability of error is zero. Human always make some mistakes which further affects the quality of software. For the quality purpose, the reliability of software should be better. It will be better if there will be less number of errors in the product/software. Quality of a system can be measured in the terms of degree to which it is reliable in performing its various function without any error or failure. This is an important measurement to check the reliability of a software as its required to confirm accuracy of the output generated by a software for an input data [1]. Fault in software means eliminating the user requirements. Faults can come due to poor documentation or poor communication between user and programmer. The process of software testing is used to find out mistakes. Testing neither guarantees the correctness of software program. If there are errors then the reliability is low.

To estimate the reliability of the software there are some models like Jelinski and moranda, goel and okumoto model, schick and wolverton, littlewood-verrall Bayesian model. Each model has advantages and disadvantages. To find accurate value of reliability need some mathematical or statistical model. Researchers want to find out accurate value of reliability [23,24].

In this paper there is technique used to determine the software reliability. The probability of failure or reliability is estimate by automata. An automaton is the best way to find out the discrete value of the software reliability [25].

This paper consists of four sections. Section 1 explains about the software reliability and its models. The basic model of reliability is explained in Section 2 along with examples. FSM is defined in Section 3. Comparison of different model is described in section 4. Section 5 conclude the article.

2 Software Reliability Model

There are some reliability models which are used to determine the reliability of software. Reliability can measure through metrics.

Rate of Occurrence of Failure (ROCOF): it is a measurement of the frequency of failures occurring in the system.

Mean time Between Failure (MTBF): it can be defined as an average of time between two failures in a repairable system.

Mean Time to Failure (MTTF): its a variation of MTBF which is applied to non-repairable system where a failure in a part can lead to failure in whole system.

Mean Time to Repair (MTTR): its average time taken to fix the failure i.e. a duration between repairing starts till the system resumes to normal functioning.

Probability of Failure on Demand (POFOD): it defines the probability of a system to fail when a safety action is failed to act against a failure.

Reliability depends upon the failure occur in the program. If MTTF is known and MTTR is known then the reliability of product can be measure. Data set is given below. This is obtained by using testing tool. From testing tool, the number fault is calculated and time taken to fix the failure is obtained by tester. Let's consider the data set and put the value in the equation of goel-okumoto model.

Total number of faults =8

Total number of test cases =105

Total execution time = 00h: 02m: 55S

Total time to repair = 20s

Equation of goel-okumoto model is

$$\mu(t) = a(1 - e^{-bt}), a > 0, b > 0$$

Where,

$\mu(t)$ = total actual fault, using okumoto model.

a= initially a is taken total number of fault detection

b=rate at which the defect rate decrease or increase.

From this the value of $\mu(t)$ is 8 means the reliability of the given data set is 0.08 this is depend upon the intensity function $\lambda(t)$.

Table 3. Time to Repair the test cases.

Number of test cases	Number of faults	Execution time	Time to Repair
5	0	00h:00m:05s	0s
10	5	00h:00m:15s	9s
15	2	00h:00m:21s	3s
20	0	00h:00m:34s	0s
25	0	00h:00m:45s	0s
30	1	00h:00m:55s	8s
5	0	00h:00m:05s	0s
Total=105	8	00h:02m:55S	20s

Procedure of using software reliability:

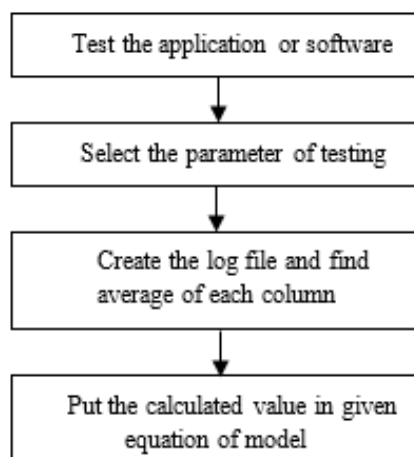


Fig. 1. Methodology for Traditional approach

From above chart in Fig. 1, the first step, test the application means collect the data in the form of code after that check the error by performing some testing process. Automated testing can be done through tools otherwise check the expected result manually. In the second step, select the parameter on the behalf of testing performed. In the third step, create the log file of each test case. Last step, put the calculated values in the given equation.

3 Finite State Machine Reliability Model

Software reliability estimation using automata based model has the ability to control various reliable software operations by transducing software states to transition states of an automata model [2]. To implement this technique, we utilize FSM reliability model. In FSM reliability model, there are four phases as shown in fig. 2. The flow starts from FSM Representation and ends at Software maintenance.

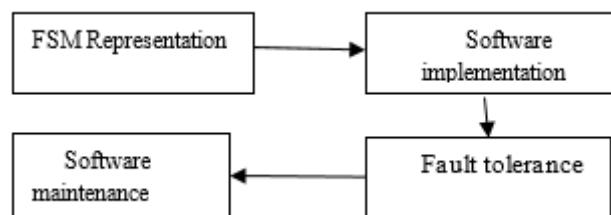


Fig. 2. Methodology for FSM

Different types of activities of reliability models are explained in table 2.

Table 4. Description of different FSM reliability model.

Sr. No.	Activity	Description
1.	FSM Representation	Give input as code, output should be shows FSM representation and next state transition
2.	Software implementation	To monitor software execution next state transition should be use.
3.	Fault tolerance	Find the path using shortest path algorithm and repeat the steps.
4.	Software maintenance	Record the path

From these activities, the FSM model can be represented. As the example, let's consider a code of any language, now divide the code in parts by states $q_0, \dots, q_n$. As the finite automata has five tuples $(Q, \Sigma, \delta, q_0, q_n)$

- Q is a finite set of states
- Σ is a finite non empty set of inputs
- δ is transition function
- q_0 is a initial state
- q_n is a final state

Similarly FSM model will be consider these tuples in order to complete their representation. FSM representation is shown in the Fig. 3.

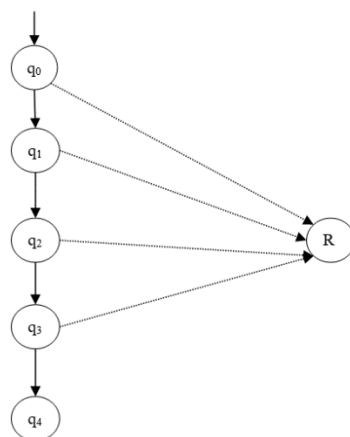


Fig. 3. Representation of software modules

Diagram conclude that each state is divide into its own reliability then total reliability R can be written as :

$$R = 1/4R_1 + 1/4R_2 + 1/4R_3 + 1/4R_4$$

The model is not only capable of predicting and estimating the reliability factor but also it can be trained to learn the evolving behavior of software dynamically which make it suitable for fault tolerance and during the software operations it become able to retract the software from any erroneous state to a safe state [4]. In [4], application of Finite State Machines has been also propmoted for the purpose of solving common design issues. FSM should follow two conditions:

- Each input, transition δ has to be defined.
- Defined transition has to handle cases where timeout occur.

FSM is mathematical model it gives the discrete value of the reliability. Through FSM, the error detection get easier as compared to other technique. The result of each state is different due to parsing because the FSM model is provide the parsing between two stages. Hence, the result will be different.

4 Comparison

The aim of reliability estimation for a software system is to predict about any possible error or failure that can occur during the operation process of software [21]. Due to defining discrete vaues for reliability and a better representation of software modules, FSM is considered as a better mathematical approach for measuring reliability in comparison to general predication model. Not only FSM, but many other automata models has been emerged as a better application in relibitiy prediction and overcome the issues of inaccuracy by traditional prediction models [22]. Comparison between these models is just to understand the difference between the automata approach and traditional approach. If the results are compared then it clearly shows that the automata approach is better methodology, because the concept of self - learning and self - healing can be proved by automata approach. The FSM based representation of software model to predict the failure is better than traditional because the later approach uses brute-force technique that applies exhaustive amount of failure data from history for prediction, i.e. it is based on assumptions. Whereas, the automata approach or FSM not based upon the assumptions. The results of both approaches are different but the better result can consider of FSM or automata approach.

5 Conclusion

Software reliability refers to the ability of software to operate without lose in a variety of environments. In this work, two methods for assessing software reliability are considered, with each method yielding some useful results. In this study, a data set is considered for measuring reliability using the Goel-Okumoto model, and then the same data set is applied using the FSM, with differing findings. Compare the performance of both the model and found the FSM model perform the better results. In the future, we can also use the Jumping Automata to improve the performance.

References

- [1] Amrit L. Goel, Software Reliability Models: Assumptions, Limitations and Applicability, IEEE Transactions on Software Engineering, Vol. SE11, No. 12, pp. 1411-1423, 1985..
- [2] P. Ahmed, Automata Based software Reliability Paradigm for Ubiquitous Systems, IJCSET, ISSN : 2229-3345, Vol.4 No. 07 .jul 2013.
- [3] W. Ritika, A Tool for runtime reliability estimation and control using automata based software reliability model, mathematical models in engineering and computer science, ISBN:978-1-61804-194-4, 2013.
- [4] K. Kandt, "Software Design Principles and Practices", Jet Propulsion Laboratory, NASA, [http:// trs-new.jpl.nasa.gov/dspace/bitstream/2014/10550/1/02-2593.pdf](http://trs-new.jpl.nasa.gov/dspace/bitstream/2014/10550/1/02-2593.pdf), (2003).
- [5] D. E. Caiazzo, J. L. Falcone, J. Hegewald, E. Lorenz, B. Stahl, D. Wang, J. Bernsdorf, B. Chopard, J. Gunn, R. Hose, M. Krafczyk, P. Lawford, R. Smallwood, D. Walker and A. G. Hoekstra, "A Complex Automata approach for Instant Restenosis: two-dimensional multiscale modeling and simulations", Journal of Computational Science, vol. 2, no. 1, (2011).
- [6] M. E. Cambroner, G. Diaz, V. Valero and E. Martinez, "Validation and Verification of Web services choreographies by using timed automata", The Journal of Logic and Algebraic Programming, (2010).
- [7] T. Carmely, "Using Finite State Machines to Design Software", In EE Times: Embedded Systems Design, vol. 23, no. 6, (2009).
- [8] C. V. Ramamoorthy and F. B. Bastani, "Software reliability: Status and perspectives," IEEE Trans. Software Eng., vol. SE-8, July 1982.
- [9] B.M. Gouthami and P.Kumar, Effective Path Selection to Estimate Software Reliability, Special Issue of International Journal of Computer Applications, ICCCMIT, 2012.
- [10] M. Satyanarayanan, Pervasive Computing: Vision and Challenges, IEEE Personal Communications, Vol. 8, No. 4, Aug. 2001.
- [11] J. L. Cook and J. E. R. Marquez, "Reliability Analysis of Cluster-based Adhoc Networks", Reliability Engineering and System Safety, vol. 93, no. 10, (2010).
- [12] M. Crochemore and D. M. Gabbay, "Reactive Automata, Information and Computation", vol. 209, no. 4, (2011).
- [13] I. K. El-Far and J. A. Whittaker, "Model-based Software Testing", Encyclopedia on Software Engineering, John Wiley & Sons, Inc, (2002).
- [14] D. Ghosh, R. Sharman, H. R. Rao and S. Upadhyaya, "Self-Healing Systems- Survey and Synthesis", Decision Support Systems, vol. 42, no. 4, (2007).
- [15] S. S. Gokhale, "Accurate Reliability Prediction Based on Software Structure", [http:// www.engr.uconn.edu/~ssg/cse300/397-232.pdf](http://www.engr.uconn.edu/~ssg/cse300/397-232.pdf), (2004).
- [16] N. Langberg and N. D. Singpurwalla, "Unification of some software reliability models via the Bayesian approach," SIAM J. Sci. Statist. Comput., vol. 6., pp. 781-790, July 1985.
- [17] P. Zhang, H. Muccini and B. Li, "A Classification and Comparison of Model Checking Software Architecture Techniques", The Journal of Systems and Software, vol. 83, no. 5, (2010).
- [18] D. J. Smith, "Reliability, Maintainability and Risk: Practical Methods for Engineers", Butterworth- Heinemann, (2001).
- [19] J. Strassner, "An Overview of the Challenges and Promises of Autonomic Computing", CSI Communications, (2011).
- [20] R. H. Reussner, H. W. Schimdt and I. H. Poernomo, "Reliability -Prediction for Component-based Software Architectures", The Journal of Systems and Software, vol. 66, no. 3, (2003).
- [21] Escalated Methods for Software Defects Audit in Repercussion and Effects Construe to Nature Inspired and Behavior Driven Mechanisms. (2019). Regular Issue, 8(6), 1779–1783. doi:10.35940/ijeat.f8442.088619
- [22] Usage Patterns and Implementation of Random Forest Methods for Software Risk and Bugs Predictions. (2019). Special Issue, 8(9S), 927–932. doi:10.35940/ijitee.i1150.0789s19

- [23] Aggarwal, A., Dhindsa, K. S., & Suri, P. K. (2020). Design for Software Risk Management Using Soft Computing and Simulated Biological Approach. *International Journal of Security and Privacy in Pervasive Computing (IJSPPC)*, 12(2), 44-54
- [24] Aggarwal, A., Dhindsa, K. S., & Suri, P. K. (2020). An Empirical Evaluation of Assorted Risk Management Models and Frameworks in Software Development. *International Journal of Applied Evolutionary Computation (IJAECE)*, 11(1), 52-62.
- [25] Aggarwal, A., Dhindsa, K. S., & Suri, P. K. (2020). A Pragmatic Assessment of Approaches and Paradigms in Software Risk Management Frameworks. *International Journal of Natural Computing Research (IJNCR)*, 9(1), 13-26.
- [26] Aggarwal, A., Dhindsa, K. S., & Suri, P. K. (2021). Performance-Aware Approach for Software Risk Management Using Random Forest Algorithm. *International Journal of Software Innovation (IJSI)*, 9(1), 12-19.

A Survey on Jumping finite automata

Rahul Kumar Singh¹, Pardeep Singh², Alankrita Aggrawal³, Isha Sharma⁴, Shivali Gaba⁵

^{1,2}SoCS, University of Petroleum and Energy Studies, Dehradun, Uttarakhand 248001, India.

^{3,5}Department of Computer Science, PIET, Panipat, Haryana, India

⁴Department of Computer Science, Chandigarh University, Mohali, Punjab, India.
pardeep.maan@gmail.com

Abstract— The class of regular languages are best described by finite automata. In this study, we have investigated new research area of automata theory which is considered as Jumping Finite Automata (JFA). The behavior of this automaton is equivalent to the classical finite automata and it is used to analyze the information in a random manner. This automation can jump in any direction of the input tape after reading a symbol and proceed to go forward anyway. An input symbol cannot be re-read again during the whole computation process, if it is already read by R/W head earlier. Otherwise, it will be considered as finite automaton in the traditional sense. With an appropriate example, this research illustrates a comparison between the traditional finite automata and Jumping Finite Automata. It also includes various additional forms of automata, such as, nested automata, on-input revolving automata, and bag automata that describe the reasons of information processing discontinuity. A systematic survey of jumping automata and comparison with finite automata is conducted to analyze these various forms.

Keywords— Classical finite automata, computational processing, formal languages, automata theory, tape reading discontinuously.

1. Introduction

AUTOMATION is significant in computing theory, designing of compiler, artificial intelligence, parsing, and other fields also. In previous centuries, many classical computational methods were designed in such a way that they perform their actions on a continuous basis. [1]. For instance, finite automaton is well recognized for encapsulating the regular language family and finite automata read the input alphabet from left to right symbol-by-symbol manner. On other hand, information processing in modern computer techniques are defined in the different way i.e. they work with data in a discontinuous way [2]. For example, some computational process are executed anywhere in the middle of a running process, while the following step may be conducted far away [3]. As a result, in order to execute the next computing step, the system must jump over a significant amount of data to the desired place, causing a discontinuity in the system [4].

Hence, traditional data processing approaches that perform their action continuously reflect discontinuous data processing of current computational methods [5]. Because of the discontinuity in data handling, the concept of adopting a classical computer model in a discontinuous manner has emerged [6]. Jumping finite automata are a novel type of classical computer model that operates in a discontinuous manner [1], [2]. It functions in the same way as traditional finite automata, except the input symbols are read in a random order. [7].

In this paper, we have investigated new version of classical finite automata named JFA and various other types of automata that work discontinuously such as bag automaton [5], nested automaton, on input revolving automaton [6].

The paper is further divided into four sections. In section I, we have explained finite automata's generalized version. Section II describes the formal definition of various kind of automata with suitable explanation. Next section III, describes the related work in the fields of automata theory that gives idea of processing information in a discontinuous. Finally, Section IV describes comparison between finite automata and jumping finite with suitable examples and compare them. At last, we conclude the article.

2. Preliminaries

Definition 2.1 Quintuple can be used to represent a finite automaton $F, (Q, \Sigma, \delta, s, f)$ where [8], [9], [10]

- Q define set of states (non-empty and finite)
- Σ represents the input alphabet (non-empty and finite).
- δ represents a mapping $\delta[Q \times \Sigma \rightarrow Q]$ of transition function which illustrates how states change throughout a transition.
- $s \in Q$ describe the starting state.
- $f \subseteq Q$ represent the final states' collection.

For example, finite automata (FA) $(Q, \Sigma, \delta, s, \{s\})$ of the language $L = (0(10)^*)$ is defined in the following way [11]:

Q is collection of states and which is defined as $\{S, T\}$.

δ is a transition function which is defined as $\delta(Q \times \Sigma) \rightarrow Q$ and for the language it is represented as $\delta(T \times 0) \rightarrow S$ and $S \times 1 \rightarrow T$.

T is considered as starting state [12]

S is considered as final state of the finite automata.

$\Sigma = (0, 1)$ is the input alphabets.

The process of the transition graph on the input string 0101010 is defined as follows [13]:

Consider the input string (0101010) for a given regular language L and find out the acceptance of the given string on finite automata.

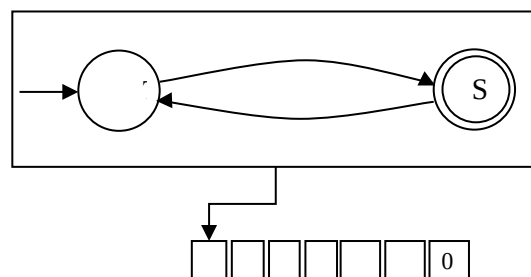


Fig.1.Reading the input 0

In the Fig.1., finite automata start with reading the symbol 0 from the input tape and moves to next state as per transition diagram of automata shown. Considering, T as the starting state, automata read the input symbol 0 from the string 0101010 and reaching to the next state S .

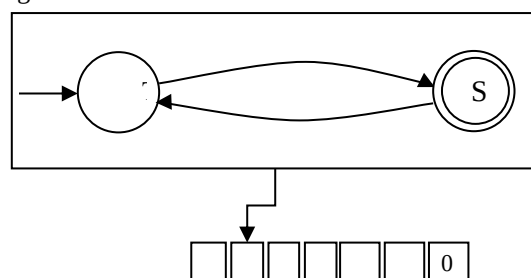


Fig.4. Reading the input 1 from input tape

Further, the next symbol will be considered by the finite automata from state S. So, when it reads the input alphabet 1, it reaches to the next state T as illustrated in the Fig.2.

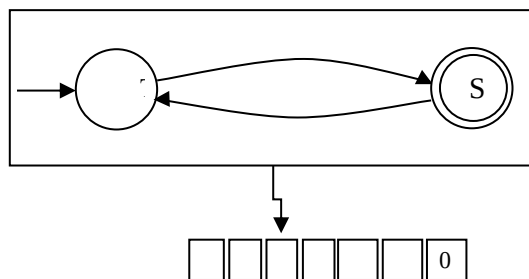


Fig.5.Reading the input 0 from input tape

Now the input pointer is on state T and it further reads the input string 0 and reach to new next state S as depicted in Fig.3.

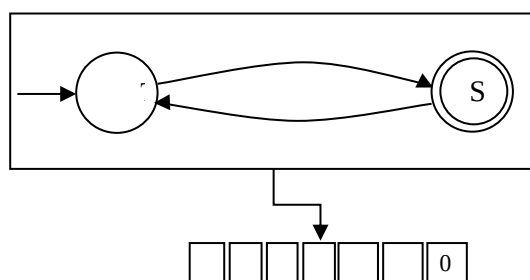


Fig. 6.Reading the input 1 from input tape

After that, finite automata read symbol 1 in the input string and reach to the next state T. Fig.4. represent the working of it.

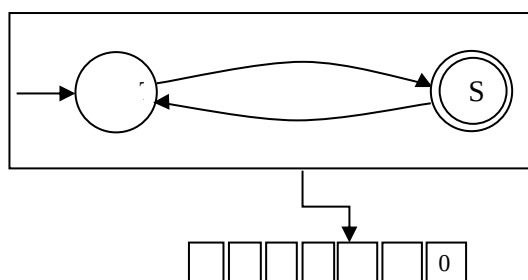


Fig.7. Reading the input 0 from input tape

Now, finite automata is in the state T and read the alphabet 0 and reach to the next state S as illustrated in Fig.5.

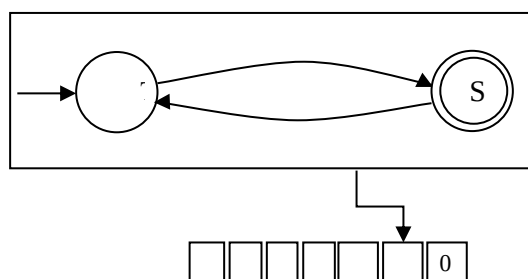


Fig. 8. Reading the input 1 from input tape

In Fig.6., finite automata reads next input alphabet 1 and reach to the state T.

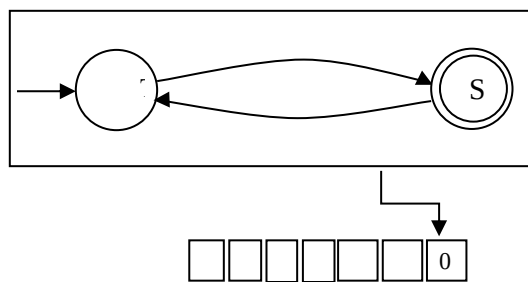


Fig. 9. Reaching at the final state by reading input string 0

Thus, at last, as depicts from the Fig.7., the finite automata read the input alphabet 0 and reach to the final state S, which shows that machine is going to accept the given string (0101010).

The Regular Language $L(F)$ accepted by this machine can be described using regular expression as: $0(10)^+$. So, this is clearly analysed from the working of FA that a FA reads the input strings one by one from left to right and at last string is going to accept by the machine, if it reaches at the final state after reading last symbol from the tape [14].

Definition 2.2: A quintuple $G = (Q, \Sigma, R, s, f)$ is defined the formal definition of General Jumping Finite Automata (GJFA) [15], [16] where

- Q represents collection of states (finite and non-empty)
- Σ defines input alphabet.
- Collection of rules is defined by R and it is in the form $pb \rightarrow q$ ($b \in \Sigma^*$, set of states Q are p and q).
- Starting state is represented by s .
- Set of final states are defined by $f \subseteq Q$ [1], [3].

The basic jumping connection is indicated by the symbol $\sim$. The jumping relation is defined as:

If $z, a', z', b \in \text{input alphabet } \Sigma^*$ and production rules (R) are $az = a'z'$ and $pb \rightarrow q \in R$ then G takes a jump from $apbz$ to $a'qz'$, which is represented as:

$$apbz \sim a'qz'$$

In a traditional way [17] [18],

- Jumping relation $\sim^n$ is a series of n ($n \geq 0$) jumps and technically, it is considered the n^{th} power of jumping relation $\sim$.
- The jumping relation $\sim^+$ is a series of non-empty jumps; precisely, It is considered the transitive closure of $\sim$.
- The jumping relation $\sim^*$ is a series of jumps (ϵ - moves allow); precisely, It is considered the reflexive-transitive closure of $\sim$.

Grammar (G) accepts the language which is indicated by $L(G)$ and it is described as:

$$L(G) = \{ab : a, b \in \Sigma^*, asb \sim^* f, f \in F\}.$$

Let $w \in \Sigma^*$, if $w \in L(G)$ then it means G accepts the string w . if $w \in \Sigma^* - L(G)$, it means G reject the string w .

If $L(G) = L(G')$, it implies that two GJFAs G' and G are equivalent to each other.

A General jumping finite automata $G = (Q, \Sigma, R, s, f)$ is of degree m , where $m \geq 0$, if $pb \rightarrow q \in R$ denotes $|b| \leq m$.

Example: The General jumping finite automata $G = (\{s, p, f\}, \{a, b, c\}, R, s, \{f\})$ and production rule $R = \{sabc \rightarrow p, pcc \rightarrow f, fa \rightarrow f\}$ is of degree 3.

Formal description 3: The formal explanation of General jumping finite automata is defined as: $G = (Q, \Sigma, R, s, f)$. If $pb \rightarrow q \in R$ and $|b| \geq 1$ that means G is an ϵ -free GJFA. General jumping finite automata is of degree m , where $m \geq 0$, if $pb \rightarrow q \in R$ implies that $|b| \leq m$ i.e. if all rules $pb \rightarrow q \in R$ satisfy $|b| \leq 1$ (in a nutshell, if the degree is 1), then G is a jumping finite automata (JFA) [1],[2],[3].

For example, the formal description of jumping finite automata is explain below with transition graph for the language $L = \{n_a(w) = n_b(w)\}$

Q is collection of states and which is defined as $\{S, T\}$.

δ is a transition function which is defined as $\delta(Q \times \Sigma) \rightarrow Q$ and for the language it is represented as $\delta(S \times a) \rightarrow T$ and $T \times b \rightarrow S$ [19].

S is considered as starting state.

T is considered as final state.

$\Sigma = (0, 1)$ is the input alphabets.

The process of the transition graph on the input string (babbaa) is defined as follows:

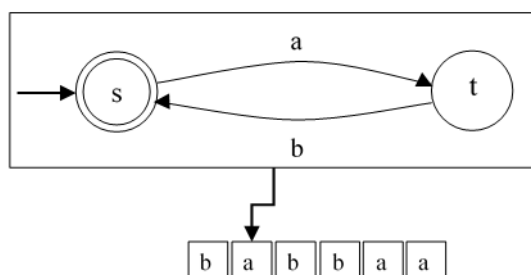


Fig.10. Reading the input a from input tape

In Fig. 8., the machine is in start state (s). From the given input string (babbaa) by making jump over input alphabet (b) it reads a first symbol (a). After that it reach to the next state (t).

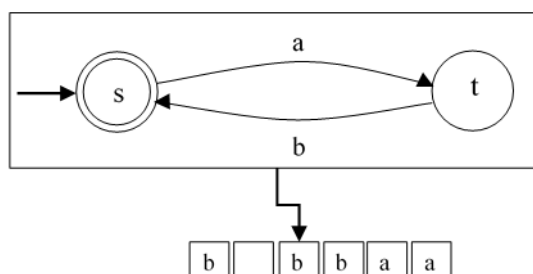


Fig.11. Reading the input b from input tape

In Fig.9., machine reads the input symbol b from the specified string. At the present time machine is in state t and after reading the current input it reaches to state (s).

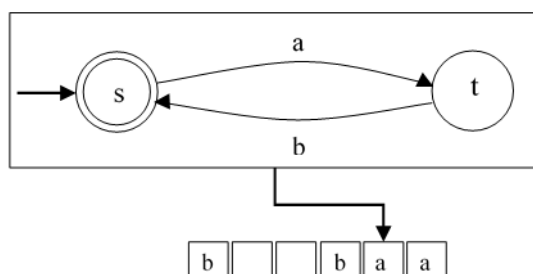


Fig.12. Reading the input a from input tape

At the present, read head is in state (s) and reads the input symbol (a) from the specified string. Machine is making the jump and reads the input (a). After these steps, machine reaches to the next state (t) as explained in Fig 10.

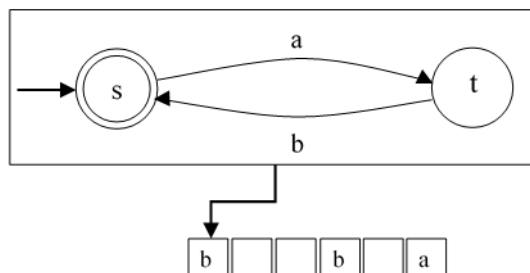


Fig. 13.Reading the input b from input tape

In Fig. 11., the read head in the state (t) and it reads the input symbol b from the specified string. Machine is making the jump and reads the input (b). After that, machine reaches to the next state (s).

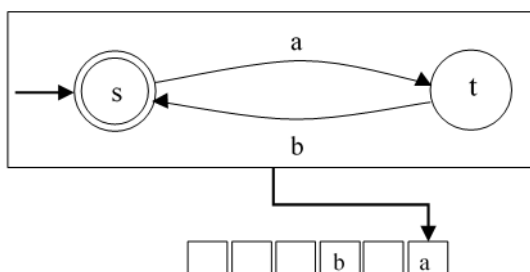


Fig.12. Reading the input, a from input tape

Now, the read head in the state (s) and it reads the input symbol (a) from the specified string. Machine is making the jump and reads the input (a). After that, machine reaches to the next state (t) as illustrated in Fig.12.

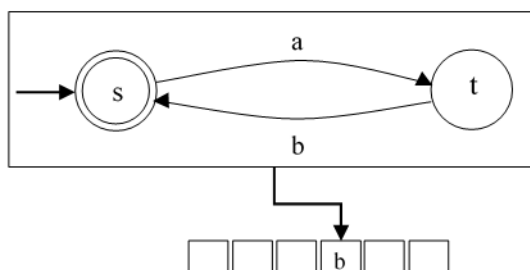


Fig. 14. Reading the input b from input tape

In Fig.13., the read head in the state (t) and it reads the input symbol (b) from the specified string. Machine is making the jump and reads the input (b). After that, machine reaches to the next state (s). Fig.14 illustrated that machine is accepting the string.

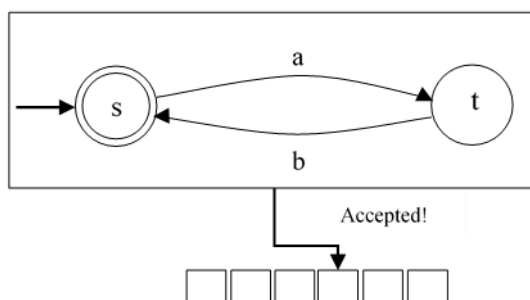


Fig. 15. Machine in final state and accept the string

In the discontinuous manner, machine is reading the input symbols and making the jumps if necessary. This is the way how reading process is going on [20].

The language is accepted by the machine (L): $\{n_a = n_b \mid w \in \{a, b\}^*\}$

3. Related Work

3.1 Finite automata with input-revolving

Bensch *et. al.* [4] examines input revolving finite automata, which is considered as a new technique in 2009. This technique perform three different actions on input symbols and works similar to finite automata —

- right revolving,
- circular interchanging,
- left revolving.

For instance, suppose, on the input tape, x is a present input symbol. We shift the rightmost input alphabet of string x at the starting of the string x in the left revolving, as depicted in Fig.15., whereas the present input alphabet of the string x is moved to the end of the string x in right revolving as presented in Fig. 16. Finally, the present input alphabet and the last input alphabet of string x are exchanged in circular interchanging as depicted in Fig.17. As a result, these three special operations alter x in a discontinuous manner, resulting in discontinuous processing data.

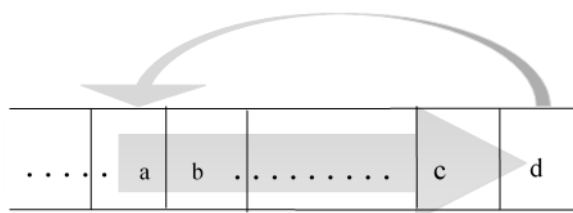


Fig .16.Left revolving [4]

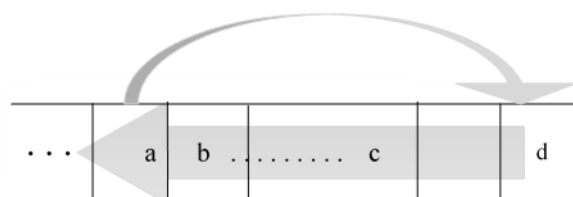


Fig. 17. Right revolving [4]

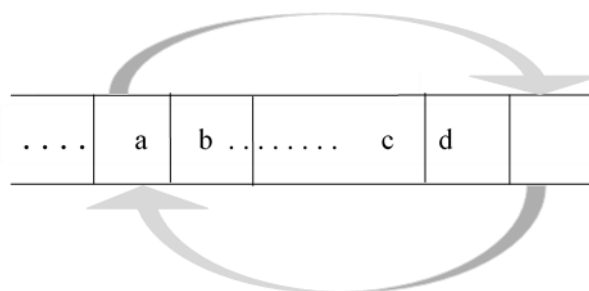


Fig. 18. Circular interchanging [4]

3.2 Bag automata

In [5], the author presented bag automata, which are similar to pushdown automata but utilizes multisets instead of pushdown lists. This automaton may change his multisets during a motion by inserting or removing

an input alphabet in between. This type of change can be made anywhere around the multiset, resulting in system discontinuity; that means the multisets are changed on a regular basis.

3.3 Nested word automata

In [6], authors introduced new automata known as nested word, that functions are similar to conventional finite automata in which they consider its input string from left to right. Nested word automata, on the other hand, work with nested words rather than regular words, as their name suggests. Additional information describing linear sequencing and nestedness measurements which is added to these nested words. Suppose that such automata are reading a input alphabet that has two preceding symbols, because of the linear sequencing, one of them appears there and because of the nestedness, the other is present. The next state is determined by the states of the automata as they operate on their two predecessors under this premise [7].

3.4 Watson-Crick Jumping Finite Automata (WKJFA)

In [22,23], authors have applied the approach of JFA over Watson-Crick Automata (WKA) for processing the double stranded sequences, which have their main application in processing DNA computing. They defined four variants for their approach and compare them with variants of Chomsky hierarchy and among themselves for their power. Other than this, researchers also analysed and investigated various operations over WKJFA such as closure, under duplication, reverse duplication and pseudo duplication.

4. Comparison

Classical finite automata (F) consist of three components: reading head, and an input tape, control unit [9], [10]. An input alphabet Σ is carried by input tape which is splitted into number of cells. In the input tape, both the ends are represented with the special symbols $\mathcal{C}$ and $\mathcal{S}$ which indicate the end of the tape. Infinite length of input tape does not contain these special symbols. Between the end markers the input string is present that need to be processed from left to right sequence. Only one square is examined at a time by the read head, which can travel in either route, right or left. The present input alphabet is the symbol under the head. A finite collection of states and a set of computational rules are used to describe finite state control. The input alphabet underneath the read head and present state of a machine will be the input to control unit.

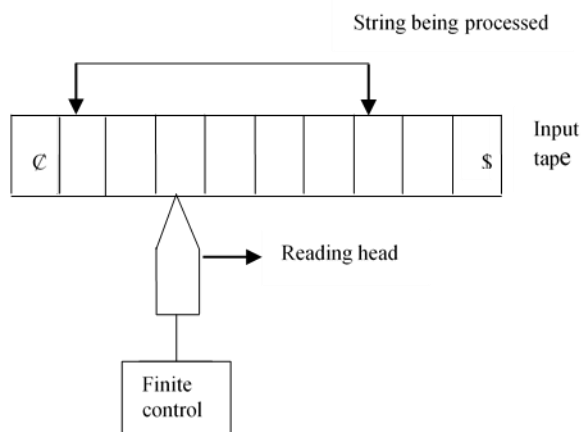


Fig.18. Finite automation architecture [10].

Finite automata is a method of computing that involves performing a series of motions. Each motion is dictated by a processing rule that specifies whether the current input sign is read or not. If the machine reads the symbol, the read head will move one cell to its original location. In finite automata, only one starting state is allowed whereas multiple final state can be assigned. If finite automata can read input string w by performing a series of moves from starting to final state, finite automata accepts w ; otherwise, finite automata will not accept w string [10].

Another side, a jumping finite automata read the input string in discontinue manner except that it works similar to classical finite automata [2]. Jumping can be performed by Jumping finite automata over a portion

of the tape after analysis a input alphabet. Once a symbol's occurrence on the tape has been read, it cannot be read again during the full process. Otherwise, it is a finite automaton in the classical sense.

The key difference between classical finite and jumping finite automata's, as discussed above, is that in classical automata's, the machine works with respect to string, while in jumping finite automata's, the input string is analysed with respect to the request of the system by making jumps where required. As a result, finite automation accommodates modern computer models' discontinuities [21].

5. Conclusion

We investigated the study of Jumping Finite automata that is considered as novel research field of automata theory. It also presents the comparison between classical formal models and modern information processing methods. This paper discussed a number of related automata, such as input revolving automata, bag automata, stacked word automata, Watson-Crick Jumping Finite Automata and so on, to give a concept of how to process information in a random fashion. This novel theory of automata proposes a discontinuous adaptation of classical formal models.

References

- [1] A. Meduna "Automata and Languages: Theory and Applications", Springer, London, 2000.
- [2] A. meduna and P. zemek "Jumping Finite Automata", International Journal of Foundations of Computer Science Vol. 23, No. 7 ,1555–1578, 2012.
- [3] A. Meduna, L. Vrabel, and P. Zemek "Jumping Finite Automata pdf", Brno University of Technology, Faculty of Information Technology, 2012.
- [4] S. Bensch, H. Bordihn, M. Holzer, and M. Kutrib "On input-revolving deterministic and nondeterministic finite automata", Information and Computation, 207(11):1140–1155, 2009.
- [5] M. Daley, M. Eramian, and I. McQuillan "Bag automata and stochastic retrieval of biomolecules in solution", In O. H. Ibarra and Z. Dang, editors, Implementation and Application of Automata, Eighth International Conference CIAA 2003, Santa Barbara, CA, July 16-18, 2003, Springer-Verlag, 2003.
- [6] R. Alur and P. Madhusudan "Adding nesting structure to words", Journal of the ACM, 56,2009.
- [7] R. Alur and P. Madhusudan "Visibly pushdown languages", In Annual ACM Symposium on Theory of Computing, Chicago, IL, USA, 2004 (3):16:1–16:43, 2009.
- [8] Daniel I. A. Cohen "Introduction to computer theory", Hunter College City University of New York, John Wiley & Sons, Inc., 1986.
- [9] K.L.P. Mishra and N. Chandrasekaran "Theory of computer science automata, languages and computation", 3rd edition, Prentice Hall of India Private limited, New Delhi - 110 001, 2008.
- [10] P. Linz "An Introduction to Formal Languages and Automata", 3rd Edition University of California at Davis, Jones and Bartlett Publisher, 2001.
- [11] Formal Languages and Algebraic Systems: Proceedings of AFLAS 2008, pages 113–123. Hong Kong University of Science and Technology, 2010.
- [12] C. D. Manning, P. Raghavan, and H. Schütze. Introduction to Information Retrieval. Cambridge University Press, New York, 2008.
- [13] T. Masopust and A. Meduna. Self-regulating finite automata. Acta Cybernetica, 18(1):135–153, 2007.
- [14] T. Masopust and A. Meduna. On pure multi-pushdown automata that perform complete pushdown pops. Acta Cybernetica, 19(2):537–552, 2009.
- [15] S. Beier, and M. Holzer, Nondeterministic right one-way jumping finite automata. Information and Computation, p.104687, 2021.
- [16] S. Beier, M. Holzer, and M. Kutrib, Operational state complexity and decidability of jumping finite automata. International Journal of Foundations of Computer Science, 30(01), pp.5-27, 2019.
- [17] S. J. Immanuel and D. G. Thomas. "Two-dimensional jumping finite automata." Mathematics for Applications 5, 105-122, 2016.
- [18] H. Fernau, M. Paramasivan, M. L. Schmid, and V. Vorel Characterization and complexity results on jumping finite automata. Theoretical Computer Science, 679, pp.31-52, 2017.
- [19] V. Vorel On basic properties of jumping finite automata. International Journal of Foundations of Computer Science, 29(01), pp.1-15, 2018.

- [20] S. Beier and M. Holzer Properties of right one-way jumping finite automata. Theoretical Computer Science, 798, pp.78-94, 2019.
- [21] S. Beier and M. Holzer Decidability of right one-way jumping finite automata. International Journal of Foundations of Computer Science, 31(06), pp.805-825, 2020.
- [22] Mahalingam, Kalpana, Ujjwal Kumar Mishra, and Rama Raghavan. "Watson–Crick jumping finite automata." International Journal of Foundations of Computer Science 31, no. 07 (2020): 891-913.
- [23] Mishra, U. K., K. Mahalingam, and R. Rama. "Watson–Crick Jumping Finite Automata: Combination, Comparison and Closure." The Computer Journal (2021).

Privacy Concerns in Big Data Security – Survey

^[1]Aashita Chhabra, ^[2]Manu Narula, ^[3]Aakanksha, ^[4]Vedant Sahrawat

^[1]Assistant Professor, Department of Information Technology, ADGITM, Delhi, India

^[2]Research Scholar, Department of Information Technology, DTU, Delhi, India

^{[3][4]}Systems Engineer, Infosys

^[1]Aashita.chhabra@gmail.com, ^[2]Manunarula1996@gmail.com,

^[3]Aakankshasingh874@gmail.com, ^[4]vedant.sahrawat1104@gmail.com

Abstract. In the past few years, big data technologies have become a center of attraction field as a research arena. The ever growing expanse of big data is strongly accompanied by an equivalent multifold in the opportunities to exploit the seclusion of concerned individuals. As we know big data need massive computing potential and tremendously vast storage, distributed systems become as ideal choice of architecture. The mere involvement of multiple parties in these environments poses an additional layer of potential breaches. Though there have been numerous developments made to incorporate security in big data layouts at different levels of working, especially focused on generating data, its storage, and data transforming stages of the life cycle of a standard big data application. The goal of this study is to come up with the theory of how privacy is employed up till recent times and discuss the challenges that concern these efforts. In the shell, we try to travel through the composition of the big data infrastructure and measures it incorporates in various stage of its life cycle to preserve a user's privacy.

Keywords. Big data, cloud computing, distributed storage, homomorphic encryption, public auditing.

1. Introduction

Accrued to the technological breakthroughs, the sheer bulk of data gathered and maintained by different entities such as social sites, Adhoc networks, Internet, etc., is exponentially increasing in real time. All this data generated from non-singular and vast sources in varied formats having high scale throughput is considered as Big Data. Big data has attracted quite a lot of researchers in recent years emerging as a hotspot among researchers from various fields. The pivotal factor being that data generation expanding at such a curve that it is shaping out to be extremely complex and tiresome to process using conventionally accepted methods [1]. Meanwhile, big data itself can imply different mechanisms based on end process for structuring the data, mainly categorized as structured data, semi-structured data, or unstructured data formats. Same approach adds a bit more challenges when handling large amount of storage and handling routines. These kind of jobs inevitably force user to opt for new methods to store and analyze datasets instantaneously. Big data, if maintained with a good enough efficiency during gathering and analysis phases, can be turned into fruitful opportunities, aiding business setups and corporations to improve the internal direction and decision making procedures and also provide a firm foundation ground for new untapped areas through data processing. It can also bear fruition in development of the scientific development areas and overall economic growth by merging standard business models with scientific paradigms [2].

Big data, over time has been subjected to many definitions but, for the study done here, we will use the definition standardized by “International Data Corporation” (IDC) in [3] The principle of big data is explained as technology and architectural techniques, designed to draw value in an economical way and effort out of volumes of a staggering variety of data, through convention of constantly arriving data capture and analysis. Keeping this definition at our base of philosophy, the significant expansion of big data is summarized by 3 V's, representing volume of data, velocity of its expanse and variety of components, as depicted in Fig. 1. Volume corresponds to the size of data that is generated or gathered or is to be processed by the system. With the digitization of world, a manifold upgrade in the size of the data has been recorded. The rate of acquiring new data is defined as the velocity of data. A standard characteristic of big data is the fact that the in hand data is never of same type but is pretty much diverse, i.e., they may house alphanumeric texts, audio recorded from different devices, images from camera or phone, or video etc. This multi-faceted nature of data gives meaning to 3rd V, the variety. Even though big data poses as extremely effective route for us to better evaluate and predict the working of the world and innovate based on this insight for the various aspects of human

endeavors, the sheer volume of data involved in the process exponentially increase the prospect of potential privacy breach. For example, Search engines like Bing, Google or avast can store user's shopping tendencies and browsing habits. Social networking sites such as Instagram archives everything about our personal life as well as social lifestyle. Major video library services such as YouTube and Vimeo provide recommendations to us based on the search history on our account. With the seat of power bestowed by big data, one can gather, store and use our personal information for commercial profits, needless to say this has substantially given rise to many threats to our privacy and social security. The recent data leaks from Facebook and Dominos and other such companies add to the evidence of same. Users' privacy may come under threat in any of the following situations [4]:

- Personal information of users may sometimes be combined with different datasets provided externally. Such a collaboration may lead to an unintentional inference of sensitive facts about the users.
- Personal information is many a times collected to keep track of history of user to suggest relevant things in businesses, such a record may reveal some sensitive information at times.
- The sensitive data involved in big data is stored securely but the processing and storage location itself may not be secured properly leading to potential leaks.

In efforts to implement a better big data privacy scheme, many attempts have been made and researched upon in recent years. These attempts can be categorized for explanation based on the particular stage of big data life cycle that they directly affect into like storage, processing or gathering. Data generation phase makes use of access restriction techniques and data falsifying methods to protect privacy. While access restriction limits the accessibility of the data, falsifying data workflow alters the original data before presenting them to a third party. The approaches in storage phase make use of encryption standards. These techniques can be broadly presented by 3 encryption families, 1) attribute based encryption (ABE), 2) Identity based encryption (IBE), and 3) storage path encryption. On top of this, as a secondary step to protect the sensitive data content, hybrid cloud architecture is implemented on storage location using private clouds for such sensitive data.

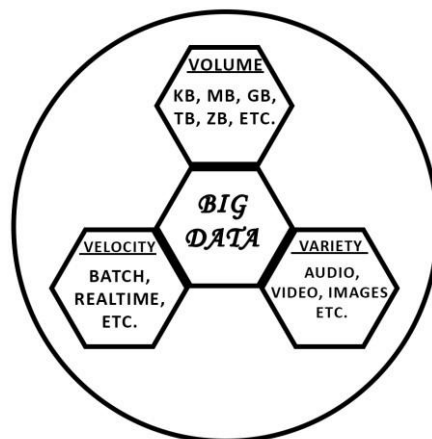


Fig. 1. The three V's of Big Data

The data processing stage consists of two parts; first one being privacy preserving data publishing (PPDP) and the other being extraction of knowledge from the data. In PPDP, techniques based on generalization and suppression concepts are applied to protect the privacy of data, these are commonly referred to as anonymization techniques. Though making sure that the utility of the data stays intact while safeguarding the privacy is still a great challenge to PPDP architects. Extracting knowledge is a very complex set of logical paths designed to extract useful information from a huge dynamic and complex dataset in a user friendly and detailed representation. These workflows can be divided into techniques based on clustering, classification and association rule mining. On one hand, classification and clustering splits the input values into different categories based on similarities in features, on the other hand, the methods based on association rule mining are focused on identifying the useful connections and trends over time in the input values.

Safeguarding a user's privacy in big data is without a doubt an exponentially expanding research area. Moreover, while the current papers effectively introduced the basic concepts and urgency of amping the level of privacy protection in big data, they somehow failed to highlight several pivotal aspects that are buried deep in this area. For example, neither of our references [2][5] provides a detailed introspect of big data measures in place for privacy safekeeping with respect to the fast emerging field of cloud computing.

In this paper, we will attempt to compile a comprehensive and complete overview of the technologies at play behind the currently implemented or in research measures at different stages of a big data lifecycle that preserve the sensitive information dealt with, on day to day basis in such applications. Moreover, we will look into the new issues or the overlooked issues that arose in a big data setup when the data is kept in storage and processed on an external cloud infrastructure, as cloud computing plays a very guiding role in the future of applications and expansions of big data eco system. Towards conclusion, we will highlight the various potential research directions.

2. Big Data : Infrastructure

For productive management of all major aspects of the three V's, efficient and effective systems are required to be designed in order to process the large inputs of data working at substantial speeds from heterogeneous sources. The life cycle of big data incorporates 3 phases, as illustrated in Fig. 2. Nowadays data often adopt a distributed nature as well as technologies keep on upgrading and evolving at a higher pace for storage and processing of the volumes of data involved. For consideration, cloud solutions, such as Hadoop or MapReduce, are exploited actively for these concerned tasks in big data.

This sub section will put some spotlight on the real life implementations of big data life cycle. On top of this, we will discuss the interaction that is in action between big data and cloud computing technologies as well as the demerits often seen cloud computing when used for storage and processing of big data.



Fig. 2. Big Data Life Cycle

2.1 Big Data Life Cycle

- **Data generation:** Data may be gathered or derived from multiple dissimilar sources. The typical generated data per day has increased over the years following a steep curve. For example, everyday 3 quintillion bytes of data come into existence on the web while 90 percent of this bulk is generated in the past few years only. Social networking sites alone hold a share of 25TB of new data every day individually. In routine, the data generated is vast, diverse and complicated. Therefore, it is beyond the limit of conventional systems to handle this work load. The data gathered for the process is generally in concern with a specific domain including but not limited to business, research, etc. depending on the aim of use and the organization in control.
- **Data storage:** This stage corresponds to the storage, safekeeping and maintenance of large-scale datasets. A data storage system comprises of two primary phases i.e., hardware infrastructure and data management [6]. Hardware infrastructure utilizes the communications (ICT) devices and networks for multitude of tasks including the distributed storage and processing solutions. Data management covers the deployed software above hardware to interact with these large data sets. It is ideal for this stage to provide the necessary interfaces for interacting with and analyzing the stored datasets.
- **Data processing:** As the name suggests, this phase infers to the workflow for collection of data, its transmission and other pre-processing tasks including cleaning as well as extraction of information. The task of collection is charged with responsibility to form the foundation of all operations and may incur

data flow from diverse sources such as texts, stills and motion media. In this stage, data is acquired from specific data production environments making use of dedicated collection technologies. Moving on to transmission stage, we require a high-speed transmission setup to shift the datasets to a specially designed storage for the multitude of analytical operations that are to be performed. Conclusively, the pre-processing tasks aim to filter out redundant sections of the data to optimize the storage space utilization. The filtered data in conjunction with a mix of analytical methods is used by several program sets to derive trends and other management needed information. Though the requirement definitions for analytics may differ for different fields, some of these domains may impose similar underlying mechanics to process and model data. Popular data research areas of present can be simplified into six broad fields: structured data analytics, text analytics, multimedia analytics, web analytics, network analytics, and mobile analytics [6].

2.2 Big Data with Cloud

Big data demand humongous computational power and storage solutions, which demands the services offered by the cloud computing architecture. Big data has its boons like cost effectiveness and scalability which make it a huge success in today's world. Other than cost and scalability, massive processing power and storage capability are the most obvious and catchy advantages the cloud presents. Modernized tasks such as real time virtualization, distributed network storage and multi-nodal processing are all born out of cloud computing which were once considered as non-achievable on conventional systems. However, on the other side, such high complexity and demand for volumes of data generates a different category of privacy issues altogether. Hiccups are often encountered in form of hesitation and uncertainty when it comes to transferring private to the cloud servers till a surety can be given about the security of the data on the said servers. There are very many challenges faced during development of such a trustworthy and secure system on cloud, some of which are discussed below [7].

- **Outsourcing:** In order to cut down costs, companies tend to outsource data to cloud systems. Evidently, third party involvement results in loss of physical control over their personal, vulnerable data. Data owners can face serious issues with the privacy of the data. The CIA principle, which stands for confidentiality, integrity, and authenticity in terms of which the outsourced data can be verifiable to the customer, becomes a necessity.

Multi-tenancy: Virtualization helps in accessing common cloud platform to multiple data owners at the same time by the help of virtual machines on a single host. The data belonging to different users is available at a common physical storage following some predefined allocation algorithms. Such a vulnerable environment gives way to malicious users to gain access to data that does not belong to them. A wide array of issues or attacks may be encountered such an environment, including but not limited to sorts of data breaches. Due to this, it is pivotal to install and implement mechanisms that handle such scenarios if need be.

Massive computation: Traditional means often fall short when it comes to handling data privacy for cloud environment in terms of computational power and speed. Cloud computing demands distinctive mechanism even if it is to achieve the same level of security that a user experiences in non-cloud activities.

3. Concerns During Data Generation

Data generation is categorized into bi-branched workflows; namely active data generation and passive data generation. Active data generation when a client consensually agrees to share data with the third party, while passive data generation where the generation of data occurs through different activities that a client performs over the internet like streaming and the owner of this data does not know of the sharing to a third party. The tendency of the data owner to restrict his data and minimize the access of his data to third parties comes forth as a big challenge during this stage. The data owners want high level of access to their own personal and sensitive data and have concerns over how much of this data is actually under his control and not available to outside users. We can minimize this risk of violation presented during data generation by restricting the access or by false encrypting data [5].

3.1 Access Restriction

Access restriction aids in empowering a data owner to shut down access to such data, which might reveal some sensitive information and is not meant to be shared. Data owners, in order to protect it from being stolen by some third party can adopt different access control methods. If the data is being passed on passively, some measures can be placed in flow ensuring privacy, by means of anti-tracking extensions, ad blockers and various other encryption algorithms [5].

Malicious users often make use of virus and malware to enter and access digitally stored data of another person. To avoid it, anti-virus and anti-malware soft wares can be used. These tools serve the same purpose on the hard stored data. Making sure our online activity and traces don't stay on the internet, we can adopt different security tools to get rid of them and hence protect our hard stored data from leaking.

3.2 Falsifying Data

In some unique cases, the data is so public or accessible that it cannot be prevented from being stolen. In such cases, data is often distorted or de-structured using certain dedicated tools before the data can be transferred to some third party. With distortion in place, the actual information becomes appreciably hard to identify and reconstruct. Some common techniques used for the task include tools like Socketpuppet, MaskMeetc.[5].

4. Concerns During Data Storage

Storing vast volumes data is not a challenge in recent times, thanks to the advances in storage solutions. However, securing this data is another thing altogether. In case the storage mechanism is anyhow compromised, the harm caused by the leak of personal information addressing countless users can prove to be virtually limitless. Therefore, it is essential to give attention to the safety procedures against damages like these. In today's information based networks, big data storing houses commonly known as data centres play key role in providing the base for these complex interactions as well as retrieval concerning huge amount of data. The distributed environment that exists now, requires an application to have several datasets that may belong to different data centres in different locations and therefore often present us with their own challenges.

The traditional security measures to protect data can be explained in regard with the area they interact with directly like file systems, databases, multi-media etc. [8]. The traditionally implemented workflows to provide the security [9], [10], [11] for conventional architectures [12] have attracted many eyes as a research area but possibly do not have an application towards big data analytics platform. Dynamic accommodation of a diverse application base requires that kind of configurability at its base. One of the technology which deal with this issues is storage virtualization, supported by the paradigm of cloud based applications [13]. Storage virtualization, in layman's terms is projecting multiple networks working together as a single unique network in itself. However, using a cloud service that is often provided as third party arrangements can incur the risks and attacks discussed above regarding the trust and effects on privacy. Therefore, in this document we will limit ourselves to discussions involving the concerns of privacy of data when stored on cloud. Our approach for discussion is twofold, first we will look onto how we can enable preservation of privacy in cloud infrastructure, then we will discuss the means of verification of information while retrieval.

4.1 Measures for preservation of privacy in storage on cloud

While stored on cloud systems, data security gets split into three fronts, confidentiality of information, integrity of data and ease of access whenever the data is required [7]. The former two fronts tackle the issues of privacy directly, the latter deals with retrieval instead.

One of the pre-requisites for a big data storage system is to safeguard an individual's privacy. There exist some traditional supports already in place to take care of that requirement. For example, the public key crypto-system. The additional approaches of encryption fortifying the user data when stored on the cloud are discussed as follows.

Attribute Based Encryption. Attribute Based Encryption [14], [15] is an encryption algorithm providing peer to peer escorting in cloud storage systems. In A.B.E. the data owner gets the privilege for setting the

policies of encryption and data is scrambled strictly abiding by the defined policies. The contents can only be de-cyphered by the users whose access level and attributes meet the predefined standards in the user managed policies. The immediate drawback for this algorithm is presented in form of the inability to update policies. One may require once in a while to modify the access policies in big data systems when dealing with changing times or clients. The currently put up attribute based encryption schemes [16], [17] do not support policy updating. The result of that is, as there is no local copy of the data once it is outsourced, the user needs to transfer back the entire data set and re-encrypt it using the new policy, then re-upload it into the big data system. This process incorporates a high communication overhead as well as computational cost.

To tackle this issue of policy modification, recently Yang et al. [18] came up with a slightly more secure and verifiable outsourcing method for such policies. In [18], the proposed system integrates query processing through which the data can be re-encrypted under new policy on the cloud itself thereby removing the overheads of transmission all over again.

Identity Based Encryption. This technique came out as a prospective replacement to the Public Key systems. IBE manages the keys by making use of a PKI based on digital certificates much like human identities such as emails or IP addresses taking place of the public keys. One such system is proposed in [19].

Encryption systems like discussed above natively do not incorporate the support for updating of cypher - receivers. However, some way arounds can be implied to perform this update. For example, if data owner adopts a snapping decryption re-encryption mode. Drawback being the massive costs of repeating calculations over huge amount of data. Moreover, this mode will require the owner to stay online the whole time. Mambo and Okamoto [20] developed a proxy-re encryption (PRE) scheme which is explained in detailed depths in [21]. This technique is aimed to juggle the issue of data accessibility between heterogenous receipts efficiently. In [20], the cyphertext generated for a receiver is morphed and sent to the receiver, while a copy of the same message is morphed in a slightly different way and sent to another recipient eliminating the risk of leak of information of the message or the keys involved. This was achieved by a routine jumping scheme applied from proxy to proxy eliminating the need for staying online the whole time. In [22], [23] and [24] this PRE is implemented over IBE and ABE techniques in some modified settings and discussed at length.

Homo-morphic Encryption. Public cloud has always been more prone to privacy breaches solely by the presence of multi-tenancy and virtualization. The users of cloud based technology often come under scenarios where they might share similar physical space according to the algorithm being followed by the cloud and in this setup, the probability of mutual data leaks become quite high. One solution to minimize this risk is to encrypt the data and store it on cloud and perform the computations over encrypted data itself. Such kind of encryption technique allows functions to directly perform on encrypted data set [25] is termed as Complete Homo-morphic Encryption. It implies that by using only the cyphertext of a message, one can perform operations directly on that and it will change the message accordingly. Thus, this technique provides high privacy on the cost of computational complexity and may not be compatible with other technologies implemented. A comparative study of discussed encryption methods are provided in Table 1.

Encryption Methods	Working Features	Drawbacks
Attribute Based Encryption	- Security protocols based on attributes set by the user. - User gets a secure and somewhat flexible control in hand.	- Computational overhead in multi categorial data. - Updates in policy and attributes not possible. - Data needs to be downloaded and decyphered before computations.
Identity Based Encryption	- Security protocols based on identity of the user. - User gets a complete access over everything.	- Time overhead for access increases with size of environment. - Changing cyphertext receiver is not possible. - Data needs to be downloaded and decyphered before computations.
Homo-Phobic Encryption	- Highly secure. - Computations and processing done on encrypted data itself.	Computational overhead is substantially high.
Proxy Re-Encryption	- Can be implemented in either IBE or ABE methods. - Cyphertext receiver can be updated.	- Computational overhead. - Data needs to be downloaded and decyphered before computations.

Table 1. Comparison of various encryption techniques

4.2 Verification of Integrity of Data in Big Data storage

Whenever cloud integration comes into play, the actual owner of the so important data immediately loses the control over data. Though there are many elements covered on top of basic security initiatives, there still exists a gap in trust from user's point of view towards the service. This situation demands for instatement of a verification based support which can prove to the owner, the intact nature of the stored data in accordance with the proposed contract therein agreed upon. Table 2 provides a summarized view of the various techniques employed for this verification purpose that are discussed at some length in this document. Many of the earlier attempts that have made significant contribution to this area over the past few years have been taken into account while providing a suitable overview of the issues [28]-[37]. In general non-cloud environment this verification is done in many ways including the transfer capacity and the extent of usage of the concerned device, some examples for same can be found in checksums or trapdoor hash methods among many others, while one can also find digital or electronic signatures equally reliable. But when we consider the sheer size of data involved in cloud environments, the overheads and computations become non-feasible, at-least without retrieving the whole chunk of data from cloud. This is why the main focus of majority of researchers in the field has been towards alienating this need to receive data from cloud [28],[29]. Ideally, the verification procedures implemented in cloud can only provide the substantial proof if the 100% of data is intact. It is often recommended to have a periodic verification in place to attain the desired level of satisfaction and security. We will be covering up the highlights of such mechanisms in place for dynamic data such as the cloud spot lighting on some of the popular integrity verification methods. Fig. 3 exhibits the general underlying framework of integrity verification methods.

Data owners are presented with option for performing integrity verification on their own or via a third party they trust. The basic groundwork for any verification workflow for integrity comprises of a tri-metry of parties namely the client, the cloud storage server (CSS) and the third party auditor (TPA). The client transfers the data to the cloud and the directive of TPA is to test and authenticate the integrity levels of stored datasets. The general phases involved in the life cycle of such a method can be described as follows [29].

- Metadata and data upload: To be able to verify the integrity of said data without transfer overheads, the user needs to compute the verification metadata. Metadata is computed by taking the original data as base and summarizing its defining features. It is then stored alongside the original data. For practical implementations, the metadata is preferred to be much smaller in size when compared to the original data.

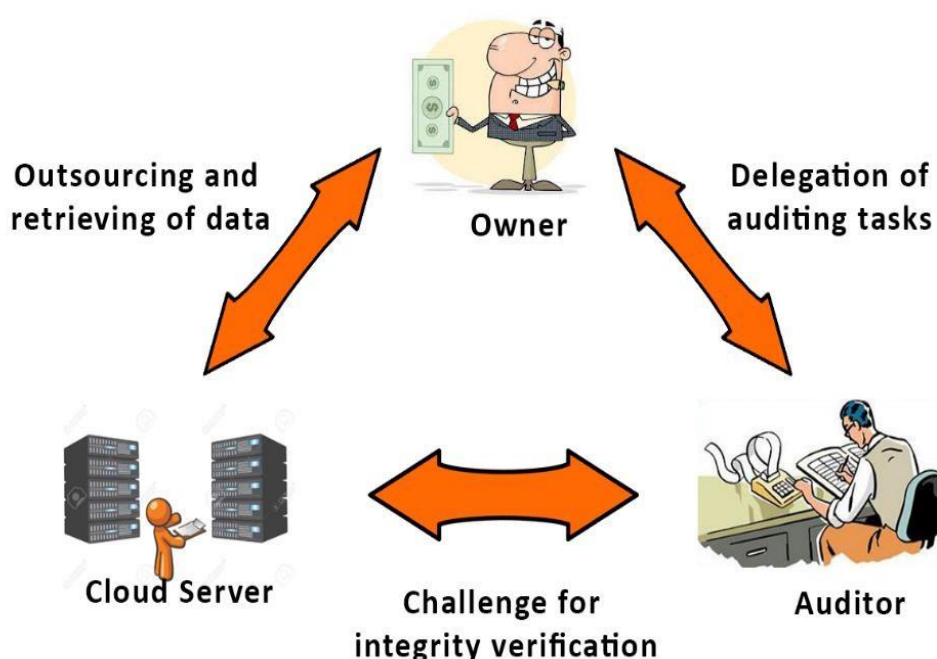


Fig. 3. Integrity verification methods.

Homo-morphic linear authenticator (HLA) or Homo-morphic verifiable tag (HVT) is used as virtues to compute metadata. The blocks of data transferred to the cloud are packaged with HVT or HLA tag. Schemes like Merkel Hash Tree (MHT) are used to add an extra layer of security which come under the authenticated data structures involved in verification [30].

- Appointing and authorizing of TPA: The TPA or Third-Party Auditor is the person who becomes responsible for verification of the data outsourced to the cloud and needs to be authorized by the data owner to manage and compute on the data. It is important to note that this TPA will only be involved if the user decides to not verify the data on their own and want a third party to handle the process, else this step is not required.

Challenge phrase and verification process: The server is sent a challenge message by the client or by the TPA on behalf of the client, it then throws a response with respect to the challenge message and the data block to the challenger. The response is then matched with the original expected response to verify the integrity. The method is said to have a public verifiability if this verification is possible to complete without the involvement of client's secret key. Most of the methods, such as provable data processing (PDP) and proofs of retrievability (POR), support this paradigm. Both of these methods will be dis-cussed in detail later. The major issue presented with public verification is that it may pave the way for a DDOSor distributed denial of service attack since the challenge phrase is simple and can be sent by anyone.

Integrity verification method	Working features	Drawbacks
PDP	• Best Security given for remote data verification. • Built on top of homo-morphic verifiable tags. • Best suited for Static Datasets.	• Susceptible to attacks in dynamic data environment. • Not as secured at the TPA end.
POR	• Guarantees accurate data possession. • Recovery of corrupted bits possible as error correcting codes are used in the process.	• Auditing is complex as ECC is involved. • Support limited to a few complex queries only.
Public Auditing	• Third party is involved in the auditing process. • Makes use of BLS signatures to prepare authentication values. • The method is proved to be completely secure.	• Data leak possible through third party auditor.

Table 2. Comparison of various integrity verification methods.

- Data updating: Data is said to be updated when standard data update like insert, delete and modify queries are used to modify the data.
- Metadata updating: After data update has been performed, it is important to initiate a metadata update for the changes according to the keys in order to keep the data set verifiable without having to retrieve all the data over again.
- Verification of updated data: Efficient verification of the data after updating is necessary as the cloud system cannot be completely trusted, it is an important step to check if the data update is done successfully or not.

PDP. First put forward in 2007 [31], [32],[33], PDP method was developed to achieve a simple random accessed verification style where any portion of any file could be verified for integrity by the means of some tags pre-calculated in certain manner like HVT or HVL tags. It is safe to say that these complex tags form the very base of the PDP architecture. Being based over RSA, it is believed that these tags have a near impossible chance of being forged as multiple tags are often bundled up to provide a random allotted verification sequence to file sections on the cloud. Having said that, it is important to note the defining properties of these tags.

- Block-independent verification. Unlike mainstream methods, HVT do not need access to the file blocks in order to provide the proof of it. This makes the process somewhat faster in comparison to block dependent methods.

- Additive verification capability: If two or more HVT or HVL tags are added in terms of their values, the resulting summation will correspond to the combined content of the concerned blocks.

POR. PORs use crypto systems instead of tags to provide the desired verification check. These are cypher proofs that are based on a challenge phrase – response phrase type of system that basically just say if a file can be accessed in its entirety without any corruptions or missing portions. The very first example of such an implementation was put forth by Jules and Kaliski [32].

However, the proposed model had a very acute limitation of application restricted to static data environments which made it impossible to implement in cloud structure at any level. This started a very hefty race in the field for feasible solutions, an improved version making use of private key crypto system was proposed by Shacham and Waters [35] that limited the verification authority to the owner with the help of a secret unique key. The mechanism provided relatively faster response and computations and thus is deemed efficient. A similar method based on POR involving a third party auditor was later studied and put forth by Armknecht et al. [34].

Public Auditing. Verification done with help of unbiased observer is deemed as public auditing [36], [37]. Wang et al. [36] proposed a method built upon BLS signature capable of supporting public auditing with full data dynamics, that is an upcoming big emerging works on public data auditing. However, being relatively too new on the field, this method lacks the compatibility for a fine-grained update or. Liu et al. [28] provides a yet another attempt at public auditing methods, this time with support for fine-grained updates over variable-sized file blocks. As an auxiliary measure, a mechanism is placed between client and the TPA to check the number of challenge phrases sent to the server to exterminate the possibility of distributed denial of service attack.

The one general issue with public auditing methods is that the linear combinations of blocks computed for definitive auditing may reveal some sensitive information, profoundly in a case where the combinations computed are related to the same block fields. Wang et al. worked to counteract or minimize this risk and came up with a privacy preserving public auditing method [37]. In this method, while computing the integrity proof, a random mask algorithm is applied simultaneously to restrict the parts of original files from being extracted in case same block is picked up in multiple requests.

5. Concerns During Data Processing

Privacy protection in transforming data may be understood by distributed in two places. In the former, the motive is to abstract information from accidental or unwanted disclosure as the data may contain sensitive information regarding the data owner. In the later area, the motive shifts to extraction of sensible information or knowledge of trends from the data without compromising the privacy in the process.

5.1 PPDP (Privacy Protection During Processing)

During PPDP, the gathered data may house sensitive information concerning the data owner. Hence modification is required before releasing the datasets for processing as they may violate privacy being released directly. On the flip side, the modifications and filtering done on the data should not conflict with the original purpose of the collection in the first place. Based on this, we can infer that the privacy and usability of data are inversely proportional with respect to each other, a relation that we will be discussing in this forthcoming section. Many researches have went by, conducted to observe the effects of modifying the data before storing or publishing for use [38], [39] for rectifying it at later stages. To conserve the privacy of a client, PPDP primarily makes use of various anonymization techniques. The original data is assumed to be completely sensitive to the user and often consists of multiple records. Each record may have all or some of the following four basic attributes [38].

- Identifier (ID): Some of the attributes act as the primary key e.g., driving license number, and Pincode etc.
- Quasi-identifier (QID): Those attributes which cannot act as primary keys but can act as a foreign key to the dataset when merged or cross referenced with other datasets. An example of same can be seen in Figure 4.

- Sensitive attribute (SA:) The main focus characteristics for the whole façade, the entries or data that a person may require to keep restricted e.g., addresses and disease.
- Non-sensitive attribute (NSA): Non-sensible attributes are fields that do not concern the privacy of an individual and can be freely published and used, all fields other than above mentioned attributes comes under this category.

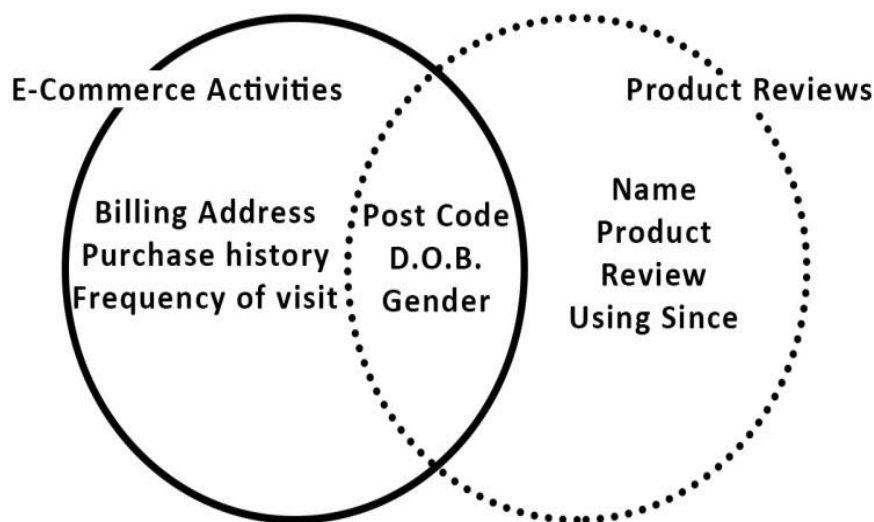


Fig. 4. Example of QID Linking

The data is abstracted by getting rid of the identifiers and manipulating the quasi-identifiers before making the data public or storing it for processing. Thanks to abstraction, the identity of the client as well as other deemed sensitive data is sneaked away from the potential stealers. The extent of abstraction is proportional to the desired privacy required in the dataset. The privacy models can be broadly divided under two groups established on the techniques unwanted intruder might employ to recognize an individual [38]. The first group falls under assumption that intruder might be able to identify clients if he forms links of the evidence with an external data source. While the second group assumes that the attack will be of a probabilistic nature that implies that the attacker can make an educated accurate guess about presence of records related to a certain client in the database.

Several models have been proposed to tackle the above mentioned problems. Catering to every aspect, some attempts for limiting record linking and others to prohibit probabilistic tracks have been employed over the years [38].

Anonymization Techniques. Before making the data public, the actual dataset is updated according to the selected level of seclusion requirements. To preserve the secrecy, we have got some techniques of anonymization operations is executed on to the data [38].

- Generalization: Generalization is the replacement of a highly specific attribute with a more generalized or broader, less specifying attribute value, for example, instead of writing cricketer or athlete, one can write sportsman. Generalization can happen at various levels such as siblings, cells, categories, sub branches etc.
- Suppression: Suppression replaces the values of a field with a predefined special character like '*' or '-' highlighting to the processing party or the general public that this field is of limits set by the owner of the data.

- **Anatomization:** Unlike generalization and suppression, anatomization do not change the values of sensitive attributes or qids, instead it splits them in two separate tables unlinking the relation between the two, the tables have one common attribute generally named as group id to re link the fields when data is transferred back.
- **Permutation:** Following on anatomization, in permutation, after making separate tables, the sensitive attributes are shuffled in their respective groups to de-link their records and making it impossible to re align them without the key to the algorithm used.
- **Perturbation:** In perturbation, the original sensitive values are replaced by synthetic auto-generated values that generally have no meaning in the real world scenarios, they are computed in such a way that the statistical analytics remain unchanged or merely changed by them. Though they play no part in actual calculations or meaningful to the clients, these just fulfil the role for data hiding for privacy execution by the user.

As pointed out earlier, the privacy level and the utility of the dataset to the actual cause it was all gathered has an inverse proportion with each other. Hence, it is important balance this trade off in a big data application. One parameter used for the purpose is data loss, that indicates the level of anonymization in the set. Many algorithms have been proposed over time to calculate the data loss with very many different approaches including minimal distortion [40], discernibility metric [41], the normalized average equivalence class size metric [42], and others [43]-[45]. To better handle the problems of this trade-off PPDP algorithms generally prefer the greedy address to attain near perfect trade-off. Most of these algorithms perform only by constructing multiple tables based on metrics given by the user regarding information privacy and metrics given by cloud for the processing pre-requisites, thus eventually coming up with a final set with minimal data loss and highest degree of privacy achievable satisfying the user.

6. Conclusion and Future Challenges

As the world is growing, so is the dependence and utility of technology around us. It is certain based on the data generated every second that the future lies in the domain of data driven mechanisms and algorithms. We have tried our best to compile a study on the current issues in the big data implementations regarding the privacy of the client's data at various stages of processing. Here in this section we aim to summarize our findings for research enthusiasts for possible and focus needed research areas.

To ensure that only authorized personnel has access to the data as well as to enforce an end to end secure transmission of data, access control methods and different encryption methods such as IBE, ABE, and PRE, are put in place. Encrypting large datasets paves way for problems as the dataset first needs to be retrieved or decyphered completely before execution of the present techniques that are available. These techniques often make it difficult for the users to perform fine grained tasks such as record sharing for data analytics. PRE, to some extent, has better performance regarding this issue. In some cases, operations require data to be shared among different parties, each with its own unique key. This results in a continuous chain of alternating encryption – decryption tasks which incurs a hefty cost in both time and computational complexities along with increasing the risk of data leak in the process. The most viable solution to this is to work on developing an environment or technique which can allow multiple entities to share data without employing this encryption loop.

Data anonymization defined as the process of shield private or sensitive data by erasing or encrypting identifiers that coordinate an individual to database. Identification of an individual becomes almost impossible from the anonymized data. Due to the ever increasing volumes of data and progress in data analytic technologies, it is becoming difficult to use these anonymization techniques to good effect. Hiding information or generalizing certain parameters is not enough. Anonymized data can be prone to many attacks and we need to study these various attack models and information loss metric in order to come up with a more effective and pragmatic approach to big data anonymization. The data around is highly dynamic and the techniques need to be built appropriately for them since many anonymization methods are only for static data. New protocols and measuring processes need to be introduced especially in the sphere of seclusion and utility. The anonymization of data is a process that can hefty and thus needs special automation tools to deal with the developing 3 V's.

Centralized storage of individual's data using cloud services is growing gradually and thus we need to understand the associated risk that come with it regarding privacy. For such personal data storage purposes centralized collection cannot be a solution and needs to be issued with. Hackers are using various sophisticated methods of attacks to breach and expose flaws in one's privacy which inevitably lead to catastrophic consequences. Personalized storage computation onto our own devices using intelligent agents should be promoted. Using such methods, encrypted cloud computing on personal devices can still be done and we can preserve our privacy by simultaneously profiting different business models available. OwnCloud and IndieWeb are some of the projects who are implementing the idea of decentralized storage. Algorithms that are developed need to be advance enough to work over extremely complex distributed data and build models that acquire the context more efficiently

Techniques like Machine Learning and Data mining are very effective on collected data and should be used. High amounts of data can be used to generate predictive models which are highly powerful. For e.g., in the fields of medical science and astronomy. However, computations are carried out by some skilled professionals belonging to third party resources and these results in threat to privacy of users. Classification, clustering and association rule mining are some of the techniques that fall under Machine learning and they need to be implemented in a manner which preserves the privacy of the clients. Sometimes the third party can face cost issues and legal constraints to procure data which might otherwise not be available to detect useful knowledge in that domain (eg. hospitals). Preservation of privacy should be of the utmost important while handling different datasets provided by an organization. Performing computations on the encrypted data available without the need to decrypt it should be deployed to solve such issues. These computation techniques might increase the complexity overhead and the main challenge is the keep it as low as possible.

References

- [1] J. Manyika et al., Big data: The Next Frontier for Innovation, Competition, and Productivity. Zürich, Switzerland: McKinsey Global Inst., Jun. 2011, pp. 1-137.
- [2] B. Maturdi, X. Zhou, S. Li, and F. Lin, "Big data security and privacy: A review," China Commun., vol. 11, no. 14, pp. 135-145, Apr. 2014.
- [3] J. Gantz and D. Reinsel, "Extracting value from chaos," in Proc. IDC IView, Jun. 2011, pp. 1-12.
- [4] A. Katal, M. Wazid, and R. H. Goudar, "Big data: Issues, challenges, tools and good practices," in Proc. IEEE Int. Conf. Contemp. Comput., Aug. 2013, pp. 404409.
- [5] L. Xu, C. Jiang, J. Wang, J. Yuan, and Y. Ren, "Information security in big data: Privacy and data mining," in IEEE Access, vol. 2, pp. 1149-1176, Oct. 2014.
- [6] H. Hu, Y. Wen, T.-S. Chua, and X. Li, "Toward scalable systems for big data analytics: A technology tutorial," IEEE Access, vol. 2, pp. 652-687, Jul. 2014.
- [7] Z. Xiao and Y. Xiao, "Security and privacy in cloud computing," IEEE Commun. Surveys Tuts., vol. 15, no. 2, pp. 843859, May 2013.
- [8] C. Hongbing, R. Chunming, H. Kai, W. Weihong, and L. Yanyan, "Secure big data storage and sharing scheme for cloud tenants," China Commun., vol. 12, no. 6, pp. 106-115, Jun. 2015.
- [9] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-preserving multikeyword ranked search over encrypted cloud data," IEEE Trans. Parallel Distrib. Syst., vol. 25, no. 1, pp. 222-233, Jan. 2014.
- [10] O. M. Soundararajan, Y. Jenifer, S. Dhivya, and T. K. P. Rajagopal, "Data security and privacy in cloud using RC6 and SHA algorithms," Netw. Commun. Eng., vol. 6, no. 5, pp. 202-205, Jun. 2014.
- [11] S. Singla and J. Singh, "Cloud data security using authentication and encryption technique," Global J. Comput. Sci. Technol., vol. 13, no. 3, pp. 2232-2235, Jul. 2013.
- [12] U. Troppens, R. Erkens, W. Muller-Friedt, R. Wolafka, and N. Haustein, Storage Networks Explained: Basics and Application of Fibre Channel SAN, NAS, iSCSI, InfiniBand and FCoE. New York, NY, USA: Wiley, 2011.
- [13] P. Mell and T. Grance, "The NIST definition of cloud computing," Nat. Inst. Standards Technol., 2011.
- [14] V. Goyal, O. Pandey, A. Sahai, and B. Waters, "Attribute-based encryption for fine-grained access control of encrypted data," in Proc. ACM Conf. Comput. Commun. Secur., Oct. 2006, pp. 89-98.
- [15] J. Bethencourt, A. Sahai, and B. Waters, "Ciphertext-policy attributebased encryption," in Proc. IEEE Int. Conf. Secur. Privacy, May 2007, pp. 321-334.

- [16] K. Yang, X. Jia, K. Ren, B. Zhang, and R. Xie, "DAC-MACS: Effective data access control for multiauthority cloud storage systems," *IEEE Trans. Inf. Forensics Security*, vol. 8, no. 11, pp. 1790-1801, Nov. 2013.
- [17] K. Yang and X. Jia, "Expressive, efficient, and revocable data access control for multi-authority cloud storage," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 7, pp. 1735-1744, Jul. 2014.
- [18] K. Yang, X. Jia, and K. Ren, "Secure and verifiable policy update outsourcing for big data access control in the cloud," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 12, pp. 3461-3470, Dec. 2015.
- [19] X. Boyen and B. Waters, "Anonymous hierarchical identity-based encryption (without random oracles)," in *Proc. Adv. Cryptol. (ASIACRYPT)*, vol. 4117, Aug. 2006, pp. 290-307.
- [20] M. Mambo and E. Okamoto, "Proxy cryptosystems: Delegation of the power to decrypt ciphertexts," *Fundam. Electron., Commun. Comput. Sci.*, vol. E80-A, no. 1, pp. 54-63, 1997.
- [21] M. Blaze, G. Bleumer, and M. Strauss, "Divertible protocols and atomic proxy cryptography," in *Proc. Adv. Cryptol. (ASIACRYPT)*, 1998, pp. 127-144.
- [22] M. Green and G. Ateniese, "Identity-based proxy re-encryption," in *Proc. Int. Conf. Appl. Cryptogr. Netw. Secur.*, 2007, vol. 4521, pp. 288-306.
- [23] J. Shao, "Anonymous ID-based proxy re-encryption," in *Proc. Int. Conf. Inf. Secur. Privacy*, vol. 7372, Jul. 2012, pp. 364-375.
- [24] K. Liang, W. Susilo, and J. K. Liu, "Privacy-preserving ciphertext multisharing control for big data storage," *IEEE Trans. Inf. Forensics Security*, vol. 10, no. 8, pp. 1578-1589, Aug. 2015.
- [25] C. Gentry, "A fully homomorphic encryption scheme," Ph.D. dissertation, Dept. Comput. Sci., Stanford Univ., Stanford, CA, USA, 2009.
- [26] S. Nepal, R. Ranjan, and K.-K. R. Choo, "Trustworthy processing of healthcare big data in hybrid clouds," *IEEE Trans. Cloud Comput.*, vol. 2, no. 2, pp. 78-84, Mar./Apr. 2015.
- [27] X. Huang and X. Du, "Achieving big data privacy via hybrid cloud," in *Proc. Int. Conf. INFOCOM*, Apr. 2014, pp. 512-517.
- [28] C. Liu et al., "Authorized public auditing of dynamic big data storage on cloud with efficient verifiable finegrained updates," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 9, pp. 2234-2244, Sep. 2014.
- [29] C. Liu, R. Ranjan, X. Zhang, C. Yang, D. Georgakopoulos, and J. Chen, "Public auditing for big data storage in cloud computing-A survey," in *Proc. IEEE Int. Conf. Comput. Sci. Eng.*, Dec. 2013, pp. 1128-1135.
- [30] R. C. Merkle, "A digital signature based on a conventional encryption function," in *Proc. Adv. Cryptol. (CRYPTO)*, Jan. 1988, pp. 369-378.
- [31] G. Ateniese et al., "Provable data possession at untrusted stores," in *Proc. Int. Conf. ACM Comput. Commun. Secur.*, 2007, pp. 598-609.
- [32] A. Juels and B. S. Kaliski, Jr., "PORS: Proofs of retrievability for large files," in *Proc. ACM Conf. Comput. Commun. Secur.*, Oct. 2007, pp. 584-597.
- [33] G. Ateniese et al., "Remote data checking using provable data possession," *Trans. Inf. Syst. Secur.*, vol. 14, no. 1, May 2011, Art. no. 12.
- [34] F. Armknecht, J.-M. Bohli, G. O. Karame, Z. Liu, and C. A. Reuter, "Outsourced proofs of retrievability," in *Proc. ACM Conf. Comput. Commun. Secur.*, Nov. 2014, pp. 831-843.
- [35] H. Shacham and B. Waters, "Compact proofs of retrievability," in *Proc. Adv. Cryptol. (ASIACRYPT)*, Dec. 2008, pp. 90-107.
- [36] Q. Wang, C. Wang, K. Ren, W. Lou, and J. Li, "Enabling public auditability and data dynamics for storage security in cloud computing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 22, no. 5, pp. 847-859, May 2011.
- [37] C. Wang, Q. Wang, K. Ren, and W. Lou, "Privacy-preserving public auditing for data storage security in cloud computing," in *Proc. IEEE Int. Conf. INFOCOM*, Mar. 2010, pp. 1-9.
- [38] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Comput. Surv.*, vol. 42, no. 4, Jun. 2010, Art. no. 14.
- [39] R. C. W. Wong and A. W.-C. Fu, "Privacy-preserving data publishing: An overview," *Synth. Lectures Data Manage.*, vol. 2, no. 1, pp. 1-138, 2010.
- [40] L. Sweeney, "k-anonymity: A model for protecting privacy," *Int. J. Uncertainty, Fuzziness Knowl. Based Syst.*, vol. 10, no. 5, pp. 557-570, 2002.

- [41] R. J. Bayardo and R. Agrawal, "Data privacy through optimal k-anonymization," in Proc. Int. Conf. data Eng., Apr. 2005, pp. 217-228.
- [42] K. LeFevre, D. J. Dewitt, and R. Ramakrishnan, "Mondrian multidimensional k-anonymity," in Proc. Int. Conf. data Eng., Apr. 2006, p. 25.
- [43] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W.-C. Fu, "Utility-based anonymization for privacy preservation with less information loss," ACM SIGKDD Explorations Newslett., vol. 8, no. 2, pp. 21-30, Dec. 2006.
- [44] A. Gionis and T. Tassa, "k-anonymization with minimal loss of information," IEEE Trans. Knowl. Data Eng., vol. 21, no. 2, pp. 206-219, Feb. 2009.
- [45] L. Xu, C. Jiang, Y. Chen, Y. Ren, and K. J. Ray Liu, "Privacy or utility in data collection? A contract theoretic approach," IEEE J. Sel. Topics Signal Process., vol. 9, no. 7, pp. 1256-1269, Oct. 2015.
- [46] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review," ACM Comput. Surv., vol. 31, no. 3, pp. 264-323, Sep. 1999.
- [47] R. Xu and D. Wunsch, Clustering. New York, NY, USA: Wiley, 2009.
- [48] A. Fahad et al., "A survey of clustering algorithms for big data: Taxonomy and empirical analysis," IEEE Trans. Emerg. Topics Comput., vol. 2, no. 3, pp. 267-279, Sep. 2014.
- [49] A. S. Shirkhorshidi, S. R. Aghabozorgi, Y. W. Teh, and T. Herawan, "Big data clustering: A review," in Proc. Int. Conf. Comput. Sci. Appl., 2014, pp. 707-720.
- [50] W. Xiao-Dan, Y. Dian-Min, L. Feng-Li, and C. Chao-Hsien, "Distributed model based sampling technique for privacy preserving clustering," in Proc. Int. Conf. Manage. Sci. Eng., Aug. 2007, pp. 192-197.
- [51] H. Xu, Z. Li, S. Guo, and K. Chen, "CloudVista: Interactive and economical visual cluster analysis for big data in the cloud," in Proc. VLDB Endowment, 2012, pp. 1886-1889.
- [52] A. M. Elmisery and H. Fu, "Privacy preserving distributed learning clustering of healthcare data using cryptography protocols," in Proc. IEEE 34th Annu. Comput. Softw. Appl. Conf. Workshops, Jul. 2010, pp. 140145.
- [53] D. Feldman, M. Schmidt, and C. Sohler, "Turning big data into tiny data: Constant-size coresets for k-means, PCA and projective clustering," in Proc. ACM-SIAM Symp. Discrete Algorithms, 2013, pp. 1434-1453.
- [54] S. R. M. Oliveira and O. R. Zarane, "Privacy preserving clustering by data transformation," in Proc. 18th Brazilian Symp. Databases, 2003, pp. 304-318.
- [55] S. R. M. Oliveira and O. R. Zarane, "Privacy-preserving clustering by object similarity-based representation and dimensionality reduction transformation," in Proc. ICDM Workshop Privacy Security Aspects Data Mining, 2004, pp. 40-46.
- [56] P. Shen and C. Li, "Distributed information theoretic clustering," IEEE Trans. Signal Process., vol. 62, no. 13, pp. 3442-3453, Jul. 2014.
- [57] C. Tekin and M. van der Schaar, "Distributed online Big Data classification using context information," in Proc. Int. Conf. Commun., Control, Comput., Oct. 2013, pp. 1435-1442.
- [58] P. Rebertost, M. Mohseni, and S. Lloyd. (2014). Quantum Support Vector Machine for Big Feature and Big Data Classification. [Online]. Available: <http://dblp.uni-trier.de/db/journals/corr/corr1307.html#RebertostML13>
- [59] R. Agrawal and R. Srikant, "Privacy-preserving data mining," in Proc. ACM SIGMOD Conf. Manage. Data, 2000, pp. 439-450.
- [60] A. Evfimievski, J. Gehrke, and R. Srikant, "Limiting privacy breaches in privacy preserving data mining," in Proc. ACM Symp. Principles Database Syst., 2003, pp. 211-222.
- [61] S. Agrawal and J. R. Haritsa, "A framework for high-accuracy privacy preserving mining," in Proc. 21st Int. Conf. Data Eng., Apr. 2005, pp. 193-204.
- [62] G. Weiping, W. Wei, and Z. Haofeng, "Privacy preserving classification mining," J. Comput. Res. Develop., vol. 43, no. 1, pp. 39-45, 2006.
- [63] X. Zhang and H. Bi, "Research on privacy preserving classification data mining based on random perturbation," in Proc. Int. Conf. Inf. Netw. Autom., 2010, pp. 173-178.
- [64] J. Han, J. Pei, and Y. Yin, "Mining frequent patterns without candidate generation," in Proc. ACM SIGMOD Int. Conf. Manage. Data, 2000, pp. 1-12.
- [65] M.-Y. Lin, P.-Y. Lee, and S.-C. Hsueh, "Apriori-based frequent itemset mining algorithms on MapReduce," in Proc. Int. Conf. Ubiquitous Inf. Manage. Commun., 2012, p. 76.

- [66] M. Riondato, J. A. DeBrabant, R. Fonseca, and E. Upfal, "PARMA: A parallel randomized algorithm for approximate association rules mining in MapReduce," in Proc. ACM Int. Conf. Inf. Knowl. Manage., 2012, pp. 85-94.
- [67] C. K.-S. Leung, R. K. MacKinnon, and F. Jiang, "Reducing the search space for big data mining for interesting patterns from uncertain data," in Proc. Int. Conf. Big Data, Jun./Jul. 2014, pp. 315-322.
- [68] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms," in Proc. 20th ACM SIGACT- SIGMOD-SIGART Symp. Principles Database Syst., 2001, pp. 247-255.
- [69] A. Evmievski, R. Srikant, R. Agrawal, and J. Gehrke, "Privacy preserving mining of association rules," in Proc. 8th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, 2002, pp. 217-228.
- [70] S. J. Rizvi and J. R. Haritsa, "Maintaining data privacy in association rule mining," in Proc. 28th Int. Conf. Very Large Databases, 2002, pp. 682-693.
- [71] Y. Lindell and B. Pinkas, "Privacy preserving data mining," in Proc. Adv. Cryptol., 2000, pp. 36-54.
- [72] Z.-H. Zhou, N.-V. Chawla, Y. Jin, and G. J. Williams, "Big data opportunities and challenges: Discussions from data analytics perspectives [discussion forum]," IEEE Comput. Intell. Mag., vol. 9, no. 4, pp. 62-74, Nov. 2014.

An Indirect Approach to Audit Structural Health of Bridges

Rakhi Akhare¹, Monika Mangla^{1*}, Nonita Sharma², Vaishali Mehta³
Lokmanya Tilak College of Engineering, Navi Mumbai¹
Dr. B. R Ambedkar National Institute of Technology, Jalandhar²
Panipat Institute of Engineering & Technology, Panipat³
manglamona@gmail.com*

Abstract: The whole world is witnessing a revolutionary transformation in terms of infrastructure during the past few decades. This transformation has been experienced in all facets of human life including retail industry, businesses, and transportation etc. Now, along with the significant development of new infrastructure, there is a need to continuously monitor the health of existing infrastructures as well. In this context, Timely monitoring and maintenance has emerged as the key aspect related to structural Health Monitoring (SHM). During recent years, various researchers have proposed various methods to audit the structural health of bridges, a vital component of infrastructure development. These methods have been broadly classified into direct and indirect approach. Authors in this paper propose an indirect method for damage detection of bridges to ensure their reliability and safety. Efficient methods for SHM will definitely serve as a boon to the mankind as collapse of bridge is a huge loss to any nation in terms of human life and economy.

Keywords: Structural Health monitoring, VBI Model, Indirect Damage detection, Accelerometer

I. Introduction

During the past few decades, the whole world is going through an unprecedented transformation particularly in terms of infrastructure. This revolutionary advancement in the infrastructure is owing to various factors viz. technological development, increase in population, and economic growth among other factors. Among various components of infrastructure bridges have been accepted to be the key component as it heavily transforms the facet of transportation in the country. As a result, there has been a boom in the construction of new bridges in the developing countries in order to achieve a smooth and fast transportation system. While new bridges are being constructed in the country, the monitoring of existing bridges also becomes a matter of paramount importance. According to the report at <https://www.nytimes.com/topic/subject/bridge-disasters>, it is evident that world has witnessed the worst disasters resulting due to collapsing of a bridge. Such disasters result into a huge loss in terms of human life and national economy thus establishing the requirement to devise an efficient method to audit the health of bridges. The prime reasons that lead to bridge collapse are identified to be old construction and violation of traffic rules. The requirement of devising efficient approaches to audit structural health of bridges is understood and hence numerous researchers have worked in this direction.

SHM is a powerful method that ensures structural integrity and safety of the civil structures. SHM consists of implementing a strategy that detects damage to any infrastructures using various sensing technologies in order to prevent any untoward incident. The prime method employed for SHM is usage of vibration data pertaining to the structure (bridges in current scenario). The motive behind using vibration data is the established fact that any damage to a structure leads to changes in its physical properties leading to noticeable variation in dynamic properties. SHM basically involves diagnosis of the damage followed by identification of its location and severity [1]. Hence, it can be summarized that SHM answers following questions:

- If there is any damage in the structure?
- What is location of the damage?
- What is severity of the damage?
- Prediction of remaining life ahead of possible collapse.

Different types of sensors and algorithms are used to answer different questions related to various levels of SHM.

The traditional method for SHM has been visual inspection which is highly fluctuating and inefficient as it can identify the weakened bridge only if the damage is visible. Hence, it may sometimes fail to identify some prospective bridges which may collapse leading to a huge loss. According to authors in [2], there have been several bridge collapses due to lack of efficient inspection thus establishing advocating the belief that visual

inspection may not be fruitful for SHM. Further, in countries like Japan that are highly prone to natural disasters, it becomes imperative to regularly perform SHM known as continuously Bridge management systems (BMSs). Hence, various authors have developed various models for SHM using various technologies.

The methods suggested by various researchers have been broadly categorized into direct and indirect methods. In direct method of SHM, different types of sensors are mounted at different places on the bridge for evaluation of vibration responses. Although direct method has agreed to be an efficient method, it has some limitations such as high cost, danger involved depending upon location of bridge, maintenance of sensors, lack of spatial information and requirement of huge data processing [3].

To address these challenges of direct method, indirect methods for SHM has gained widespread acceptance. Indirect approach uses instrumented vehicles that moves over the bridge in order to collect data about the bridge such as natural frequencies as shown in Fig. 1. In this interaction among bridge and vehicle termed as vehicle-bridge interaction (VBI), the vehicle works as transmitter as well as receiver. Resultantly, indirect monitoring offers a mobile, cost-effective, and sustainable solution to the traditional direct bridge SHM approach.

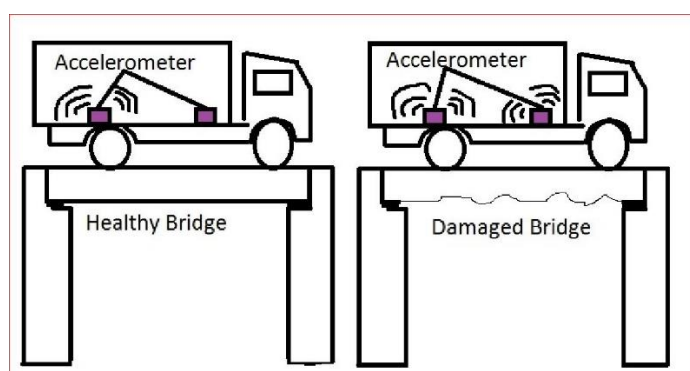


Figure 1: Illustration of Indirect method for SHM

Authors in this paper present an indirect approach for SHM in consideration with transportation bridges. In the proposed method, authors use the Vehicle-Bridge Interaction model (VBI) to identify the weakened bridges that may collapse in near future. VBI model measures the response of moving instrumented vehicles as a moving sensory system that moves over bridge to record the dynamics properties of bridges. Recorded dynamic information indicates the health of bridge which may be used to prevent sudden collapses. The proposed system consists of a prototype model that is used as an instrumented vehicle mounted with accelerometer on its axles. The model also integrates frequency analysis. In this simulation model, simple VBI method is explored in detecting damage from vehicle acceleration. The proposed model avoids direct installation of any type of instruments on the bridge and hence reduces the cost and efforts.

In the paper, authors have proposed a scaled model that employs an indirect method for SHM of bridges. The current model collects the acceleration signals received from vehicle and bridge system. The collected signals are processed for four different bridge scenarios and five speed levels. Further, the signals are classified through a simple short-time Fourier transform that detects shifts in the fundamental frequency of the bridge owing to variation in the bridge condition. Results demonstrate efficient detection of variation in the bridge condition on applying this technique on the signals collected from the bridge thus constituting a direct method of monitoring. This technique outperforms the indirect method where signals from sensors mounted on the vehicle are processed [4].

The paper has been organized into various sections. The concept of SHM and along with direct and indirect method for the same has been discussed in section I. Related work by various authors has been presented in section II. Experimental setup is explained in section III. Results are discussed in section IV and finally conclusion and future work is given in section V.

II. Related Work

The concept of indirect approach of SHM, where the dynamics of bridge are extracted through a moving vehicle, is the brainchild of authors in Yang et al. [5][6]. In their work, authors used a theoretical closed-form solution for vehicle-bridge interaction as discussed earlier. Basically, the indirect methods for SHM are based on systems that are used for monitoring the vibration based conditions of machine and its parts [7]. It is done by utilizing the response of a machine in terms of velocity and displacement in response to vibrations [4].

According to the literature available, the methods for SHM can be classified into multiple categories [8]. Considering the requirement for devising efficient methods for SHM, various countries have employed Bridge management systems (BMS's)[9]. According to authors in [10] [11], indirect method of SHM can provide more information in comparison to direct method. Further, indirect methods can also be extended to design a new range of VBI models that works for greater complexity.

The work of devising efficient approaches for SHM is also carried by Keenahan et al. in [12] by proposing subtraction idea. The proposed method suggests mounting accelerometers over the axles of a trailer which is towed by a heavy vehicle such as truck. In the method proposed by [12], the accuracy depends upon the similarity of the axles as claimed by Yang et al. [13]. Further, it is also noticed that this method gets influenced by environmental noise thus limiting its efficiency.

However, the authors in [14][15] discussed the challenge in indirect method of SHM. For the same, authors claimed that vehicle moves over the bridge for a very short time which poses limitation for conventional signal processing techniques that uses infinite time series. This limitation can be addressed by employing wavelet methods which have established its supremacy over conventional indirect method by demonstrating noticeable results in detecting, quantifying and finally locating the damages in bridge.

Although several researchers have preferred indirect method over direct method of SHM as evident from the literature [16][17][18]. However, this indirect method also has an associated challenge. If the vehicle intended for extracting the vital features of the bridge moves too fast over bridge, it will remain on the bridge for a very limited time. This may lead to probable shortage of vehicle bridge interaction data. This can be addressed by moving the vehicle using slow speed which may lead to another problem of congestion or bottleneck on busy roads. Another associated challenge for indirect methods of monitoring is interference from environmental effects such as temperature. According to work in [10][19], these issues of indirect method can be addressed by using vehicles that repeatedly moves over the bridge. Authors in current paper also present an indirect method for SHM as presented in the subsequent section.

III. Experiment

As the vehicle crosses a bridge, there is a dynamic interaction among vehicle and bridge. The response of vehicle is influenced by the bridge dynamics. From the underlying principle of SHM, it is evident that if the bridge is damaged viz. concrete cracking, stiffness, damping, mass of the bridge change due to damage and its vibration characteristics also changes. The authors in this paper have used this property to design a prototype model to identify the bridge damage detection.

Various components used in design of prototype model are:

Accelerometer MPU-6050 is an integrated 6-axis Motion Tracking device that combines a 3-axis gyroscope, 3-axis accelerometer, and a Digital Motion Processor™ (DMP) all in a small 4x4x0.9mm package

Main Signal Processor ESP32, a 32 bit microcontroller is a series of low-cost, low-power system on a chip microcontrollers with integrated Wi-Fi and dual-mode Bluetooth.

Arduino IDE with ESP-IDF Core for programming the microcontroller

Graphing Software Processing can be considered as a opensource graphical library and integrated development environment (IDE) that is designed for electronic arts and visual design communities etc. The aim is to educate non-programmers with the fundamentals of computer programming using a visual tool.

ESP32 is interfaced with the IMU unit i.e. MPU6050 using I2C protocol as shown in Fig. 2. This IMU unit is placed on the axle of the car. The data received is stored into an array of variables and a software rolling. Average filter is implemented to reduce the noise. The filtered data received is then processed to get FFT. This data is sent over serial to PC and graph is displayed. Processing IDE is used to plot the FFT data received.

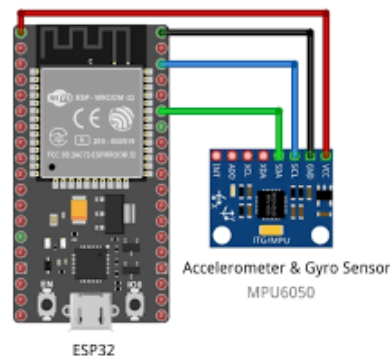


Fig. 2: Circuit diagram for ESP32 and MPU6050 interface

For evaluating the output received, the IMU unit is mounted on a vibrator as shown in Fig. 3. Further, a sine wave of 100Hz is played through a vibrator and the corresponding output of ESP32 is compared. As the vibrator creates sound, the vibration is picked by the sensor. Fig. 4 shows the FFT plot on the Arduino plotter for 100Hz and 200Hz. The peak indicates 100Hz inertial vibration caused by the vibrator. As it is a pure sine wave, the harmonics observed are negligible.



Fig. 3: IMU unit mounted on a vibrator.

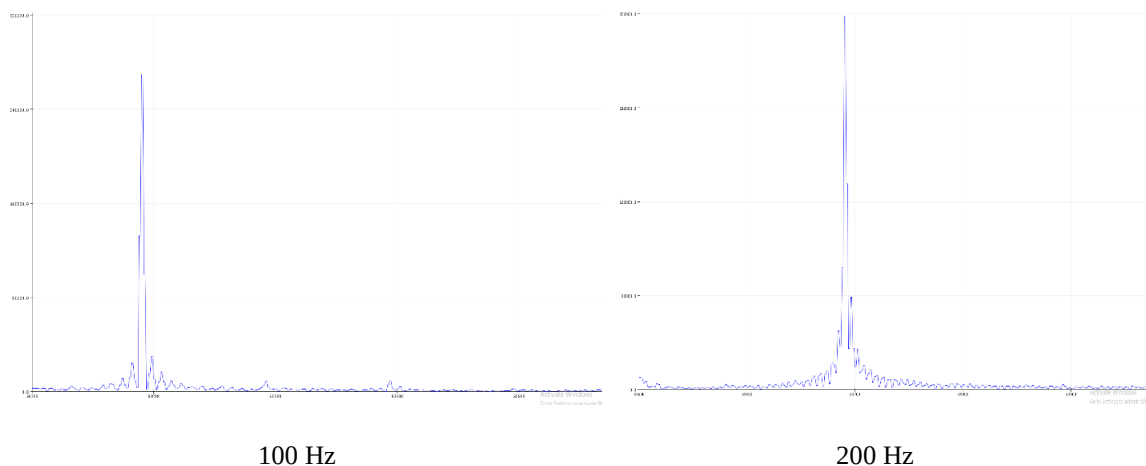


Fig. 4: FFT plot on the Arduino plotter.

Further, 100Hz square wave is played through the vibrator and the corresponding Harmonics are illustrated in Fig. 5.



Fig. 5: Harmonics generated by vibrator due to increased frequency

The setup of the proposed prototype is demonstrated in Fig. 6 where ESP32 is interfaced with accelerometer and is placed on a vehicle.

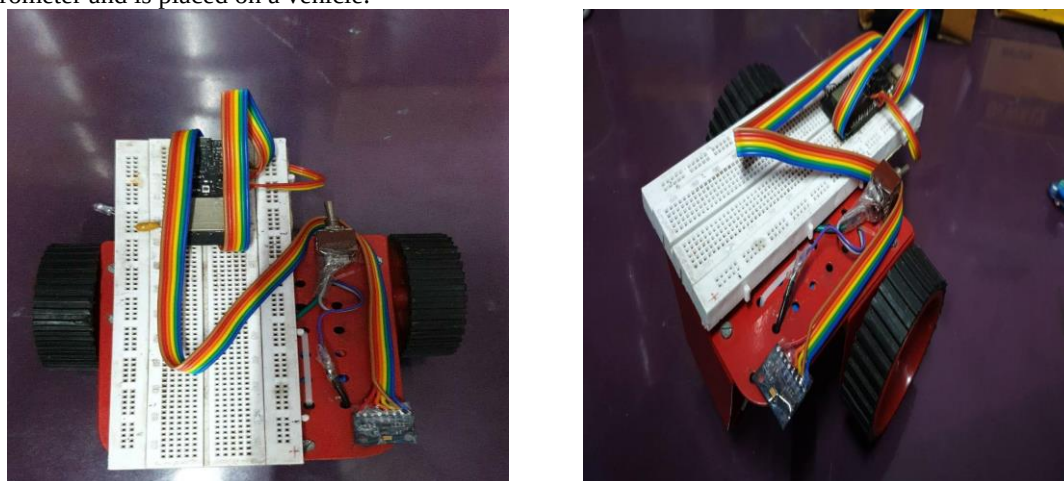


Fig. 6: Illustration of proposed prototype model

IV. Results

The proposed model uses indirect approach to evaluate the health of bridge by mounting accelerometer on the moving vehicle. The vehicle must neither be too fast (which would fail the sampling of lower frequencies) or not too slow. The ideal speed should be around 90 km/h. In this demonstration, authors use stationary vehicle owing to the limitations of size of bridge and the low sampling rate of accelerometer. Authors used a vibrator/speaker to induce vibrations in the proposed system. Frequency would be generated around the natural frequency of bridge so as to emulate the real vibrating condition of bridge. However, it is worth noting that higher frequency components will vary from real scenario. Now, when vibrations are generated, the accelerometer (MPU 6050) sends sample to connected ESP32 microprocessor at certain sampling rate. ESP32 processes these samples and forwards these values to remote workstation/PC. The fundamental frequency and its harmonics are observed in form of a graph.

In order to determine the Boundary condition, some threshold values are predefined. In real life situations, threshold is determined based on factors like construction of bridge and the parameters of its endurance. Further, LED is connected to ESP 32 as an indicator to presence of damage or not. If the vibrations generated by vibrator crosses the determined threshold value, it will be displayed on graph and it triggers the LED to ON state as shown in Fig. 7. This is generally indicated by red LED signaling the warning. Thus, the proposed model may identify whether the bridge is damaged or not. Vibrations within the threshold value signals the safe state.

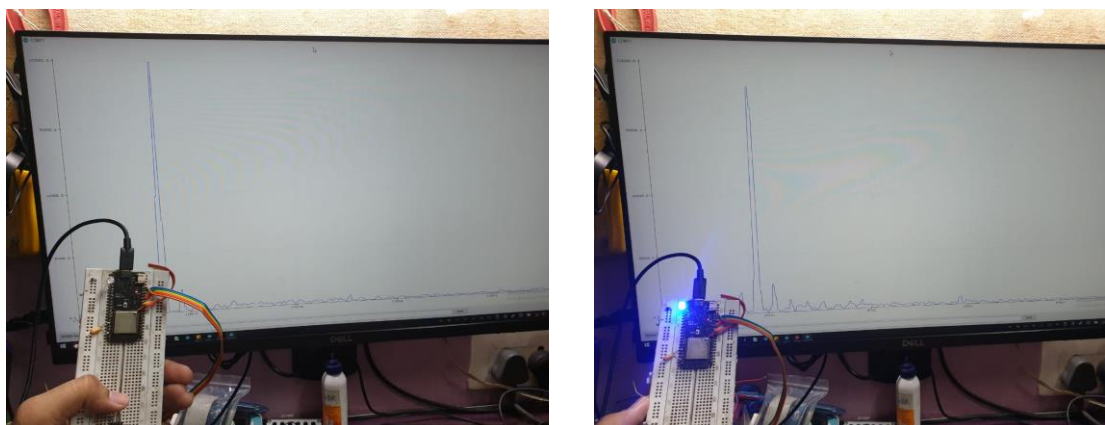


Fig. 7: Illustration of LED in proposed prototype model

Now as evident from Fig. 7, if the bridge is damaged, its physical properties (vibration in this case) gets changed. When an instrumented vehicle passes from a damaged bridge, the built in system observes the vibration frequency of bridge. If vibration crosses the threshold value, it indicates that the bridge is damaged. It then sends a warning signal to the remote PC. The proposed VBI prototype model is easy to implement and does not require any direct placement of sensors on bridge. Hence, the proposed model can be efficiently used for SHM of bridges at large.

V. Conclusion

In this paper, authors presented a simple indirect approach for SHM of bridges, the most crucial concept related with safety of people during their journey. The paper has presented a simple VBI model for testing the structural behavior of bridges. Here, system places an accelerometer as a sensor at the axel of a vehicle instead of bridge to measure the vibration between vehicle and bridge. The prime advantages of proposed model are cost effectiveness, ease in implementation and reliability. However, it has an associated challenge in terms of low measurement time due to short length of bridge. The work can be extended to consider additional parameters like mode shapes, modal parameter etc. for damage detection.

References

- [1] Farrar, C.R., Worden, K.: Structural Health Monitoring. A Machine Learning Perspective. Wiley, New York (2013)
- [2] Chupanit P, Phromsorn C. The Importance of Bridge Health Monitoring. International Science Index. 2012;6:135-8.
- [3] Malekjafarian, A., McGetrick, P. J., & OBrien, E. J. (2015). A review of indirect bridge monitoring using passing vehicles. Shock and vibration, 2015.
- [4] Cerda F, Garrett J, Bielak J, Barrera J, Zhuang Z, Chen S et al. Indirect structural health monitoring in bridges: scale experiments. Bridge Maintenance, Safety, Management, Resilience and Sustainability. 2012:346-53.
- [5] Yang YB, Lin CW, Yau JD. Extracting bridge frequencies from the dynamic response of a passing vehicle. Journal of Sound and Vibration. 2004;272:471-93.
- [6] Yang YB, Lin CW. Vehicle-bridge interaction dynamics and potential applications. Journal of Sound and Vibration. 2005;284:205-26.
- [7] Kan, M. S., Tan, A. C., & Mathew, J. (2015). A review on prognostic techniques for non-stationary and non-linear rotating systems. Mechanical Systems and Signal Processing, 62, 1-20.
- [8] Carden EP, Fanning P. Vibration based condition monitoring: A review. Struct Health Monit. 2004;3:355-77.
- [9] Akgul F. A BMS development project with an integrated inspection program. Bridge Maintenance, Safety, Management, Resilience and Sustainability. 2012:2659-65.
- [10] Kong X, Cai CS, Kong B. Damage Detection Based on Transmissibility of a Vehicle and Bridge Coupled System. J Eng Mech-Asce. 2014.
- [11] Khorram A, Bakhtiari-Nejad F, Rezaeian M. Comparison studies between two wavelet based crack detection methods of a beam subjected to a moving load. Int J Eng Sci. 2012;51:204-15

- [12] Keenahan J, OBrien EJ, McGetrick PJ, Gonzalez A. The use of a dynamic truck-trailer drive-by system to monitor bridge damping. *Struct Health Monit.* 2014;13:143-57.
- [13] Yang YB, Li YC, Chang KC. Using two connected vehicles to measure the frequencies of bridges with rough surface: a theoretical study. *Acta Mech.* 2012;223:1851-61.
- [14] Rytter A. *Vibration based inspection of civil engineering structures.* Denmark: Aalborg University; 1993.
- [15] Carden EP, Fanning P. Vibration based condition monitoring: A review. *Struct Health Monit.* 2004;3:355-77.
- [16] Gonzalez A, OBrien EJ, McGetrick PJ. Identification of damping in a bridge using a moving instrumented vehicle. *Journal of Sound and Vibration.* 2012;331:4115-31.
- [17] McGetrick PJ, Kim CW. A Parametric Study of a Drive by Bridge Inspection System Based on the Morlet Wavelet. *Key Eng Mater.* 2013;569-570:262-9.
- [18] McGetrick PJ, Kim CW. An indirect bridge inspection method incorporating a wavelet-based damage indicator and pattern recognition. 9th International Conference on Structural Dynamics, EUROLYN 2014. Porto, Portugal 2014.
- [19] Carden EP, Fanning P. Vibration based condition monitoring: A review. *Struct Health Monit.* 2004;3:355-77.

Recognition of Pneumonia from the X-Ray image pattern using Convolutional Neural Networks

Sanjay Santhanam¹, Dr. Selvi S², Sathya S³, Thyagaraj T⁴, Yanamadala Vasu Deva Sai⁵, Kalaivani P⁶

^{1,3,4,5} Student, Department of Computer Science, RMK Engineering College, Chennai, India

² Professor, Department of Computer Science, RMK Engineering College, Chennai, India

⁶ Professor, St. Josephs Engineering College, Chennai, India

sanjays2402@gmail.com

Abstract: Pneumonia is an infection in the lungs that inflames the air sacs in one or both lungs. It is caused by a bacteria called Streptococcus pneumonia that usually lives in the upper respiratory tract. More than 10 million people in India are affected by this annually. Doctors use a chest X-ray to confirm the infection and its location. Thus developing an automatic system for detecting the disease can help to treat the patients quickly. With the powerful Artificial Intelligence algorithms, the medical field has evolved in a very vast manner. Deep Learning Convolutional Neural Networks (CNN) is the go-to algorithm when it comes to image classification. CNN models are so powerful that they are already in use for many disease detections. Moreover, CNN models can be used to analyze and find patterns in the images. We build a CNN model to analyze the given X-ray image patterns and predict whether the patient is infected or not.

Keywords: Artificial Intelligence, Machine Learning, Deep Learning, Convolutional Neural Networks, Image Classification, Pneumonia Detection, lung disease

I. Introduction

The recent outbreak of Covid-19 caused by the virus SARS-CoV-2 has taken us back to the 20th century when Spanish influenza spread. The coronavirus contained in the droplets moves down the respiratory tract. The virus starts to multiply, and the infection progresses to the lungs. This can also develop Pneumonia which is not usual Pneumonia caused by bacteria. Therefore, it becomes crucial for the doctor to diagnose the symptoms and identify the disease. The primary diagnosis for Pneumonia is by taking a Chest X-ray of the patient. However, examining chest X-rays is not a relaxed and easy-going task. Pneumonia can be hazy and misapprehended with other diagnoses like lung scarring, congestive heart failure, and many more in the images. This can lead to misclassification or false detection of Pneumonia in a patient or, in the worst case, not be diagnosed as Pneumonia itself when it is present.

Furthermore, with more doctor's safety and precautionary measures, the time spent attending a patient becomes less. In the less time available for the physicians to attend to the patients, the former has to check the reports and scans and diagnose the disease or disorder. Wrong diagnoses can lead to loss of life. Thus, detecting Pneumonia from X-ray images is challenging, and modeling an algorithm that can detect the presence of thoracic disease like Pneumonia will save much time, and doctors can quickly start to treat the patients accordingly.

Deep learning is an AI technique that makes computers learn and do things, like how humans complete their tasks with their intelligence. It imitates how the human brain processes the data to detect objects, speech recognition, translate speech, and make decisions. For example, Deep learning is a critical technique in developing self-driving automobiles, which makes the vehicle recognize traffic signals to distinguish a pedestrian from a lamp-post, increasing or decreasing speed concerning the nearby cars. It is the leading technology used in voice control in consumer devices like phones, tablets, smart TVs, and hands-free speakers, and it is used in Apple's Siri, Windows Cortana, etc. Moreover, there are many more applications of deep learning: image recognition, Visual art processing, suggesting medicines based on disease, Customer relationship and management, Recommendation systems, money laundering, face recognition, etc. Now it is achieving the outcomes which were not possible in previous years.

A Convolutional Neural Network is a class of Deep Learning algorithms that takes an image or set of images as input and gives a prediction with the help of parameters provided such as weight, number of nodes. Compared to other image processing algorithms, the amount of pre-processing in CNN is much lesser and faster than different algorithms. ConvNets can quickly identify the characteristics with enough data and training. The visual region's layout prompted the creation of a ConvNet, which is comparable to the neuronal connection network in the human brain.

II. Literature Survey

The Computer Vision problems that have existed for the past five decades are now getting solved with the Convolutional Neural Networks of Deep Learning algorithms. Image classification models built with CNN are performing effectively and have gained trust that they have replaced the labor and time in analyzing the images. For example, doctors diagnose Pneumonia by getting a report from radiology which they can use to crucial pneumonia regions. The latest AI system that can report all key pneumonia findings is called Chexpert. This system uses Convolution Neural Networks (CNN) as a deep learning algorithm to analyze chest x-ray images.

Paper [1], [17] sets up a diagnostic system based on a deep-learning framework for detecting whether the patient has Pneumonia or not. The only classification algorithm used is the Neural Network, and there is no comparison between the accuracies of several algorithms. In [8], [12], Automatic disease identification from chest X-rays at the level of experienced radiologists would be beneficial, not just in clinical situations. It would, however, help provide health care to communities in remote regions that do not have access to diagnostic imaging specialists.

Paper [2], [13], [24] proposed a convolutional neural network (CNN) model that can help the doctor to predict the presence of Pneumonia using the set of images that the model was trained with and the collection of the previous X-Ray helps to prepare the model much better. They built a CNN model that helped figure characteristics from a chest x-ray picture and predicted whether a patient was sick of Pneumonia from scratch and did not use any pre-trained models.

In [3], The research investigates the variations in findings obtained when neural networks are employed to analyze the trends of different age groups. [25] Furthermore, the document displays the accuracy generated from various neural network techniques and the accuracy derived from various SVM kernels. [23] SVM kernels beat Neural Network methods by a substantial margin in this case. As a result, this study demonstrates that SVM outperforms neural network methods in terms of system performance.

In [4], Based on a deep-learning architecture, the research offers a diagnostic tool for screening individuals with blinding retinal diseases. [14][18] The weights and biases of a pre-trained model, on the other hand, have a substantial influence on its performance. As a result, the performance of this model is expected to increase when tested on a larger ImageNet dataset with more complicated deep-learning methods.

In [5], the suggested study proposes a convolutional neural network and residual network-based deep neural network (ResNet), as well as cosine annealing and stochastic gradient descent with restarts methods for discovering optimum differential rates to create an efficient and accurate network that can detect and forecast the existence of Pneumonia using chest x-rays. Paper [6], features from pre-trained models and combinations of SVM and CNN models were used to predict the presence of pneumonia disease. Finally, in [7], the famous CNN model VGG16 with transfer learning and fine-tuning was used to diagnose Pneumonia.

Paper [8][10][17], Mask R-CNN, a faster R-CNN, was used to predict the presence of Pneumonia. In [9],[26], to generalize the results, the underlying model RetinaNet was trained with the test set from Kaggle on ImageNet, and the dataset was supplemented.

The motive for creating a system that already has replacements in the market that have had success is because this new system may provide fresh insight or perception. This insight or perception might be helpful when building more systems in the future.

III. Proposed Solution

• Dataset

The dataset is taken from Kaggle[34], which Paul Mooney submitted. It contains 5863 chest X-ray images divided into three folders (i.e., train, test, and validation), with two subfolders having regular and Pneumonia chest X-ray images.

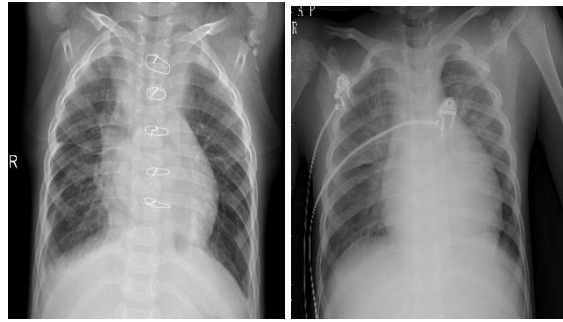


Fig. 1. Sample images with pneumonia

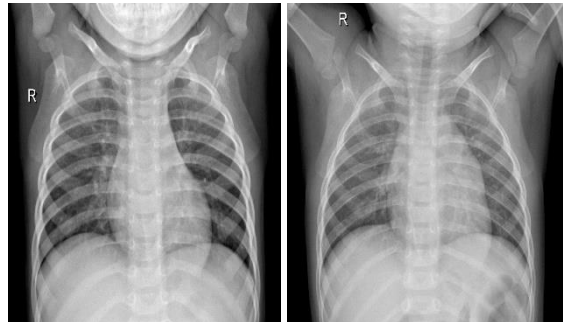


Fig. 2. Sample images without Pneumonia

IV. Image Processing

Image processing is a way of converting images to digital form. The images obtained may differ in size and shape, so it must be taken care of to accept them in the appropriate condition. The following steps are done:

- The dataset is loaded to do image processing.
- The images are converted to the correct size.
- The images are augmented by flipping, zooming, shearing the images to introduce variance in the pictures.
- Using a Feature detector, crucial features like edges are detected from the photographs and are used to produce feature maps.
- The model can use the feature maps to predict a new image easily.

V. Convolutional Neural Network

Convolutional Neural Network is the deep learning technique we used to diagnose Pneumonia from the chest X-ray pictures. Convolutional Neural Networks (CNNs) is a type of Deep Learning algorithm that receives input, calculates values from one layer with suitable weights and outputs. The basic idea is that CNN tries to identify features and patterns from the input and predicts the result. CNN strives to imitate the neuron connectivity of the human brain.

We initialize the CNN model and give the augmented image dataset as the input to it. The Convolution layer detects features from the images and produces feature maps. The feature maps are then fed to the pooling layer, where the convolved features are furthermore reduced in dimension for faster computation. The pooling can be done in two ways, Max pooling, where the maximum value is obtained from the feature map, and Average Pooling, where the average is obtained. The Flatten layer then reduces the high dimensional data to a column vector. The column vector is then given to the hidden layers where the layers' units process the data and produce output by an activation function. At last, the output layer predicts the result.

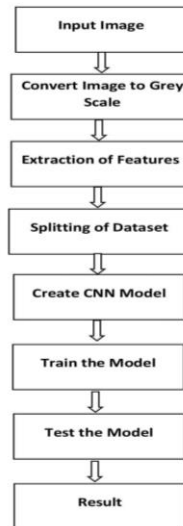


Fig. 3. System Flow Chart

VI. Result

The model was deployed in a local server for testing purposes. The datasets were collected from Kaggle[34], which Paul Mooney submitted. The training set in the dataset was fed into the CNN model. After the training, the Test dataset was used to test the model adequacy.

The model utilized three layers which are the Conv2d layer, max-pooling layer, and flatten layer. These three dense layers had units 512, 256, 1, respectively.

The model possessed an accuracy of 94.57% and validation accuracy of 85.53% for every ten epochs (Number of passes of training dataset). The output layer used the sigmoid activation function for valid prediction. sss

VII. Conclusion

The CNN model thus built has an accuracy of 94.57%. The model is saved in the local storage for future predictions. Using Python's Flask framework, we made a web page to upload chest X-ray images. The image is stored locally and then given to the saved model, predicting and returning the output. The output is then displayed on the webpage with the uploaded x-ray image.

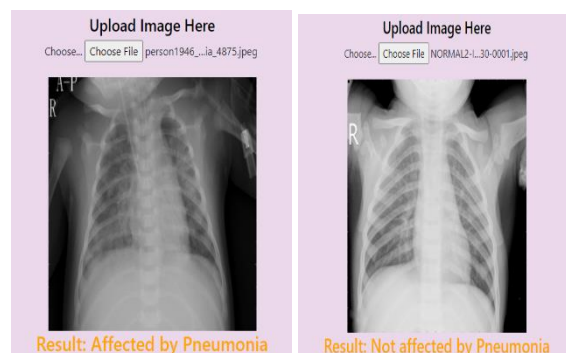


Fig. 4. Output displayed on the web interface

VIII. Future Scope

The image size used was 64 x 64, which can be increased for better accuracy and prediction results that require increased GPU power. In addition, the number of layers of the CNN model can be altered to improve the model's

efficiency and performance.

References

- [1] Rajpurkar P, Irvin J, Ball RL, Zhu K, Yang B, Mehta H, et al. (2018) Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNet algorithm to practicing radiologists. *PLoS Med* 15(11): e1002686. <https://doi.org/10.1371/journal.pmed.1002686>
- [2] Okeke Stephen, Mangal Sain, Uchenna Joseph Maduh, Do-Un Jeong, "An Efficient Deep Learning Approach to Pneumonia Classification in Healthcare," *Journal of Healthcare Engineering*, vol. 2019, Article ID 4180949, 7 pages, 2019. <https://doi.org/10.1155/2019/4180949>
- [3] Begg, R.' A Comparison of Neural Networks and Support Vector Machines for Recognizing Young-Old Gait Patterns' ISBN: 0-7803-8162-9 Published in TENCON 2003. <https://ieeexplore.ieee.org/document/1273344>
- [4] Daniel S. Kermany, Michael Goldbaum, Wenjia Cai' Identifying Medical Diagnoses and Treatable Diseases by Image-Based Deep Learning' *Cell Volume 172, Issue 5, P1122-1131.E9*, February 22, 2018. <https://doi.org/10.1016/j.cell.2018.02.010>
- [5] Kalyani Kadam, Dr.Swati Ahirrao, Harbir Kaur ,Dr. Shraddha Phansalkar ,Dr.Ambika Pawar' Deep Learning Approach For Prediction Of Pneumonia' *International Journal Of Scientific & Technology Research* Volume 8, Issue 10, October 2019 Issn 2277-8616
- [6] D. Varshni, K. Thakral, L. Agarwal, R. Nijhawan and A. Mittal, "Pneumonia Detection Using CNN based Feature Extraction," 2019 IEEE International Conference on Electrical, Computer and Communication Technologies (ICECCT), 2019, pp. 1-7, DOI: 10.1109/ICECCT.2019.8869364.
- [7] Thakur S., Goplani Y., Arora S., Upadhyay R., Sharma G. (2021) Chest X-Ray Images Based Automated Detection of Pneumonia Using Transfer Learning and CNN. In: Bansal P., Tushir M., Balas V., Srivastava R. (eds) *Proceedings of International Conference on Artificial Intelligence and Applications. Advances in Intelligent Systems and Computing*, vol 1164. Springer, Singapore. First Online:02 July 2020 https://doi.org/10.1007/978-981-15-4992-2_31
- [8] S. Roshini, R. Sanjana, K. B. Ranjith, and M. Karthick, "Pneumonia Detection from Chest X-ray Using Mask R-CNN," *IJRESM*, vol. 4, no. 3, pp. 82–83, Mar. 2021.
- [9] T. Gabruseva, D. Poplavskiy, and A. Kalinin, "Deep Learning for Automatic Pneumonia Detection," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2020, pp. 1436-1443, DOI: 10.1109/CVPRW50498.2020.00183.
- [10] H. Sharma, J. S. Jain, P. Bansal, and S. Gupta, "Feature Extraction and Classification of Chest X-Ray Images Using CNN to Detect Pneumonia," 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2020, pp. 227-231, DOI: 10.1109/Confluence47617.2020.9057809.
- [11] Yan Han, Chongyan Chen, Ahmed H Tewfik, Ying Ding, Yifan Peng, "Pneumonia Detection on Chest X-ray using Radiomic Features and Contrastive Learning" DOI: <https://arxiv.org/abs/2101.04269>
- [12] Luka Račić, Tomo Popović, Senior Member, IEEE, Stevan Čakić, Stevan Šandi, Pneumonia Detection Using Deep Learning Based on Convolutional Neural Network, DOI: <http://dx.doi.org/10.1109/IT51528.2021.9390137>
- [13] Jianpeng Zhang;Yutong Xie;Guansong Pang;Zhibin Liao;Johan Verjans;Wenxing Li;Zongji Sun;Jian He;Yi Li;Chunhua Shen;Yong Xia, "Viral Pneumonia Screening on Chest X-Rays Using Confidence-Aware Anomaly Detection" DOI: <https://doi.org/10.1109/TMI.2020.3040950>
- [14] Zong-Ye Yang;Qiangfu Zhao, A Multiple Deep Learner Approach for X-Ray Image-Based Pneumonia Detection, 2020 International Conference on Machine Learning and Cybernetics (ICMLC), DOI : <https://doi.org/10.1109/ICMLC51923.2020.9469043>
- [15] Anand Nayyar; Rachna Jain; Yash Upadhyay, "Object detection based approach for Automatic detection of Pneumonia" 2020 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD), DOI: <https://doi.org/10.1109/icABCD49160.2020.9183876>
- [16] Muhammed Talo, "Pneumonia Detection from Radiography Images using Convolutional Neural Networks", 2019 27th Signal Processing and Communications Applications Conference (SIU), DOI: <https://doi.org/10.1109/SIU.2019.8806614>
- [17] Songtai Wan;Chih-Yu Hsu;Jie Li;Ming Zhao, "Depth-Wise Convolution with Attention Neural Network (DWA) for Pneumonia Detection", 2020 International Conference on Intelligent Computing, Automation and Systems (ICICAS), DOI: <https://doi.org/10.1109/ICICAS51530.2020.00035>
- [18] Sheikh Rafiul Islam;Santi P. Maity;Ajoy Kumar Ray;Mrinal Mandal, "Automatic Detection of Pneumonia on Compressed Sensing Images using Deep Learning". 2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE), DOI: <https://doi.org/10.1109/CCECE.2019.8861969>
- [19] Abdullah Faqih Al Mubarak;Jeffrey A. M. Dominique;Ahmad Habbie Thias, "Pneumonia Detection with Deep Convolutional Architecture", 2019 International Conference of Artificial Intelligence and Information Technology (ICAII), DOI: <https://doi.org/10.1109/ICAII.2019.8834476>
- [20] S. L. Contreras-Ojeda;C. Sierra-Pardo;J. A. Dominguez-Jimenez;J. Lopez-Bueno;S. H. Contreras-Ortiz, "Texture Analysis of Ultrasound Images for Pneumonia Detection in Pediatric Patients", 2019 XXII Symposium on Image, Signal Processing and Artificial Vision (STSIVA), DOI: <https://doi.org/10.1109/STSIVA.2019.8730238>

- [21] Gaurav Labhane;Rutuja Pansare;Saumil Maheshwari;Ritu Tiwari;Anupam Shukla, "Detection of Pediatric Pneumonia from Chest X-Ray Images using CNN and Transfer Learning", 2020 3rd International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things (ICETCE), DOI: <https://doi.org/10.1109/ICETCE48199.2020.9091755>
- [22] Devansh Srivastav;Akansha Bajpai;Prakash Srivastava, "Improved Classification for Pneumonia Detection using Transfer Learning with GAN based Synthetic Image Augmentation", 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), DOI: <https://doi.org/10.1109/Confluence51648.2021.9377062>
- [23] Abhishek Sharma;Daniel Raju;Sutapa Ranjan, Detection of pneumonia clouds in chest X-ray using image processing approach, 2017 Nirma University International Conference on Engineering (NUiCONE), DOI: <https://doi.org/10.1109/NUiCONE.2017.8325607>
- [24] Xiaochun Wei;Yishuang Chen;Zihao Zhang, "Comparative Experiment of Convolutional Neural Network (CNN) Models Based on Pneumonia X-ray Images Detection", 2020 2nd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI), DOI: <https://doi.org/10.1109/MLBDBI51377.2020.00095>
- [25] Sonain Jamil;Muhammad Sohail Abbas;Fawad;Muhammad Faisal Zia;Muhib Ur Rahman, "A Deep Convolutional Neural Network Based Framework for Pneumonia Detection", 2021 International Conference on Digital Futures and Transformative Technologies (ICoDT2), DOI: <https://doi.org/10.1109/ICoDT252288.2021.9441539>
- [26] Xin Li;Fan Chen;Haijiang Hao;Mengting Li, A Pneumonia Detection Method Based on Improved Convolutional Neural Network, 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), DOI: <https://doi.org/10.1109/ITNEC48623.2020.9084734>
- [27] Jakub Garstka;Michał Strzelecki, "Pneumonia detection in X-ray chest images based on convolutional neural networks and data augmentation methods", 2020 Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA), DOI: <https://doi.org/10.23919/SPA50552.2020.9241305>
- [28] Shanay Shah;Heeket Mehta;Pankaj Sonawane, "Pneumonia Detection Using Convolutional Neural Networks", 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), DOI: <https://doi.org/10.1109/ICSSIT48917.2020.9214289>
- [29] Shailender Kumar;Pankaj Singh;Mohit Ranjan, "A review on deep learning based pneumonia detection systems", 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS) DOI: <https://doi.org/10.1109/ICAIS50930.2021.9395868>
- [30] Harsh Sharma;Jai Sethia Jain;Priti Bansal;Sumit Gupta, "Feature Extraction and Classification of Chest X-Ray Images Using CNN to Detect Pneumonia", 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), DOI: <https://doi.org/10.1109/Confluence47617.2020.9057809>
- [31] Rabia Emhamed Al Mamlook;Shengfeng Chen;Hanin Fawzi Bzizi, "Investigation of the performance of Machine Learning Classifiers for Pneumonia Detection in Chest X-ray Images", 2020 IEEE International Conference on Electro Information Technology (EIT), DOI: <https://doi.org/10.1109/EIT48999.2020.9208232>
- [32] TALIBI ALAOUI Youssef;BERRAHOU Aissam;DOUGE Khalid;BELABED Imane;JAARA El Miloud, "Classification of chest pneumonia from x-ray images using new architecture based on ResNet", 2020 IEEE 2nd International Conference on Electronics, Control, Optimization and Computer Science (ICECOCS), DOI: <https://doi.org/10.1109/ICECOCS50124.2020.9314567>
- [33] Meldon Dcunha;Nimesh Naik;Ivan Francis;Shainila Mulla, "Pneumonia Detection in X-rays Using OpenCV and Deep Learning", 2021 International Conference on Communication information and Computing Technology (ICCICT), DOI: <https://doi.org/10.1109/ICCICT50803.2021.9510180>
- [34] Kaggle.com. 2021. Chest X-Ray Images (Pneumonia). [online] Available at: <<https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia>> [Accessed 10 November 2021].

Search for Alzheimer's Type Dementia Pattern from Spontaneous Speech: Signal Processing Approach

Akta Rani¹, Deepak Kumar², Suresh Chand Gupta³

¹⁻²⁻³ Department of Computer Science and Engineering, Panipat Institute of Engineering and Technology,
Panipat, Haryana, India

Email: ektabajaj15@gmail.com, wadhwa123deepak@gmail.com, sureshguptaji@yahoo.com

Abstract: Alzheimer's disease, popularly known by the term AD, is a disease that affects the cognition and memory of a person. The affected person becomes dependent on others even for basic tasks. Due to the unavailability of treatment for this disease, it becomes necessary to catch the symptoms in its early stage so that proper steps can be taken to stop or delay its deterioration further. Our study develops a model for Alzheimer's diagnosis where the biomarkers are found from the spontaneous speech of subjects. The model is prepared by applying the machine learning classifiers on acoustic parameters of speech which are extracted by applying the signal processing approach to the subject's speech. The model was able to diagnose the disease with 77% accuracy using Logistic regression and 76% using an SVM classifier.

Keywords: Alzheimer's detection, Machine learning, classification, speech analysis, feature selection, signal processing, feature extraction, automatic analysis of spontaneous speech (ASSA).

1. Introduction

As a person gets older, his ability to remember things starts getting down [6]. We notice many odd things in them like forgetting the date or time, forgetting words while speaking, etc. But if it's too often, the person is unable to make small decisions, losing the track of season and time; it may be a sign of Alzheimer's [9].

Alzheimer's is a type of dementia that affects brain-related functions. It is a disorder that slowly destroys the memory, thinking skills, cognitive abilities of human beings. The progression of the disease is gradual. In the beginning, the person faces difficulty retaining new information, expressing his/her thoughts. This is the early stage. In the next stage, also called the moderate stage, the person is unable to recall his personal details, complete complex tasks, etc. The damage may go on to the severe stage where the person can't even do his basic tasks and gets limited to his bed. The reason for this is the death of brain cell connections and cells themselves [13].

The name has been kept on the name of Dr. Alois Alzheimer. In 1906, He observed a woman who was going through a mental disorder [3] [16]. The woman showed symptoms like memory weakness, navigation problems, and many more. The disorder went on to severe stage and she lost her life. After the death of the woman, Dr. Alzheimer examined the woman's brain and found many abnormal clumps and tangled fibers in various positions.

Also, the breaking of connections in neurons can be found in Alzheimer's patients which help in transmitting messages between the brain and other organs of the body. The disorder initially attacks through the hippocampus and the entorhinal cortex. With time, additional parts start getting affected. By the end, the disorder spreads all over. Brain tissues get shrunk considerably.

The change in behavior increases with the increase in damage to the brain. Although brain changes start taking place even before the symptoms appear, but still it is better to detect these symptoms as early as possible. We may not cure the damage done so far but we can stop or delay the deterioration further. To detect these symptoms earliest is the motive of our research.

Different types of Biomarkers can be noticed in Alzheimer's patients. They may be the physical ones like plaques, tau, hippocampal volume, etc. which are monitored by brain scanning.

Speech is also considered a good biomarker in the detection of disease. Speech is a human motor activity [14]. According to researches, an adult can produce 14 different sounds in a second via harmonized action of

around 100 muscles. It is believed that fluency to speak and lexicon quality are not good in Alzheimer's patients [8]. So our study works on finding the biomarkers related to these two speech characteristics.

Our study proposes such a system that helps in the automatic distinction between subjects with or without Alzheimer's. The system depends on acoustic parameters of speech which are evaluated using the Signal Processing strategy on the speech signal. The system helps us predict whether the subject shows any symptoms related to AD or not.

This paper comprises 5 sections. Section 2 gives an insight into the related work done in the detection of Alzheimer's biomarkers using speech analysis. In section 3, the proposed approach is described. The experimental setup and results are exposed in sections 4 and 5 respectively. In the end, Section 6 gives the remarks of our study and also an insight into the future study.

2. Related Work

Various studies have been read which make us believe that Alzheimer's patients develop language dysfunctionality as the level of brain cell damage increases. The dysfunctionality can be observed in the words of a person's speech as well as the way of speaking.

According to Kato [15], speech has three elements: prosody, timbre, and phoneme. The prosodic component has important information such as emotional expressions. His study develops a data mining approach for distinguishing the patients based on prosodic features extracted from elderly speech. The prediction accuracy was 75.9 %.

The study by K. Lopez [11] introduces two new nonlinear features i.e. Castiglione fractal dimension and Permutation entropy for distinguishing the categories of patients. The model developed by him was dependent on linear features such as duration, voice quality, pitch, etc., and above two nonlinear features. The categorization task was performed by the WEKA software suite. The limitation is that he could achieve an accuracy just above 50 %

The study by Toth [17] tried to build its own ASR (Automatic Speech Recognition) tool which considered some acoustic features related to breathing, pauses, laughter, etc. He wanted to focus on these acoustic parameters to classify patients. They were able to diagnose MCI (Mild Cognitive Impairment) patients with an accuracy of 75 %. The cons to this study were that the dataset was not large enough, which gives a more general result, and also it didn't consider Alzheimer's patients in his study.

According to Ammar's [3] study, various factors affect the accuracy of prediction like used classifiers, feature extraction, and several attributes for feature selection. The study by Ammar prepares the machine learning model based on linguistic features of speech such as syntactic, semantic, and pragmatic features. It included feature extraction, imperative feature selection followed by classification. Features were related to transcriptions of speech. He could prove the fact with 79% accuracy. The drawback of this study was that it completely depends on transcriptions of speech.

Our study is close to the study of Luz [12]. The study by Saturnino Luz [12] tries to prove the fact that instead of depending on just transcriptions of speech, vocalizations can also be used as significant biomarkers in Alzheimer's diagnosis. Several features were used for the categorization task. Although the model obtained an accuracy of just 68%, it gave a practical solution for the development of the system.

All these studies have been done on different aspects of speech. The aspects include transcription, fluency, prosody, etc. From these different studies, we observed that speech samples can be used as an efficient attribute towards finding the indicators of AD.

Our study is close to the study of Saturnino luz where we have tried to prove that acoustic features are significant biomarkers in Alzheimer's diagnosis. Signal processing equipped with machine learning is applied in our approach.

3. The Proposed approach

Our study's goal was to develop such a system that could help us in the detection of AD symptoms by analyzing the person's speech. This section describes the proposed approach for this system. The approach is divided into the following steps :

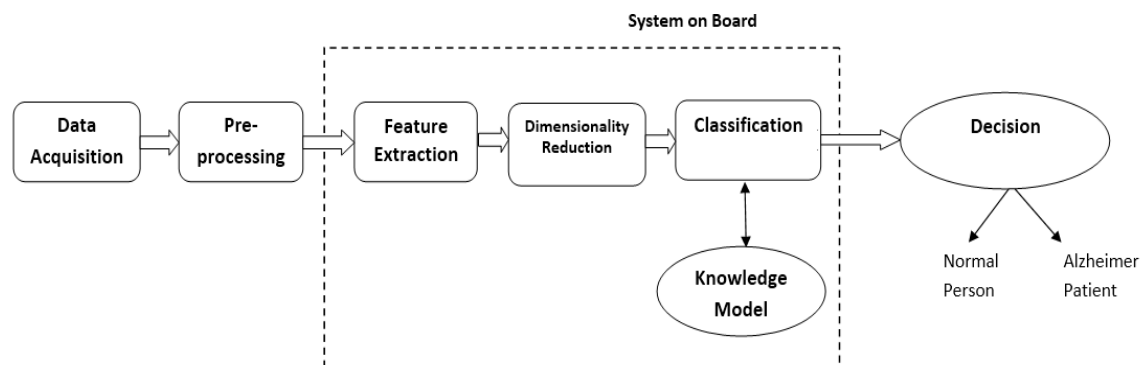


Fig. 1. The Framework of Proposed Approach

3.1. Data Acquisition:

Machine learning is a technology that learns from past data and gives decisions on new data. It finds the patterns or structures in the past data.

Data acquisition, the starting step for any system, refers to the process of collecting this data from relevant sources for further processing. The data collected here is in raw form. We have collected voice data of persons in our approach.

3.2. Data Preprocessing:

Before applying the approach, the raw data is always cleaned and preprocessed, so that it does not contain any bad or missing data values. These values may harm the system in giving correct predictions

3.3. Feature Extraction:

The first step of the detection system approach is feature extraction. This is an important step of analyzing. As the voice data can't be understood by the classification model but only organized statistical features, we apply signal processing on each voice signal to get these statistical features.

To get these features, we take the sample data at a rate of 44100 Hz (44100 samples per second) from each voice signal and apply signal processing to derive the following features.

Acoustic features: Different acoustic features are used in the approach as given in the table [1].

Name	Description
Duration	The total duration of speech
Short Pauses (Pauses of less than 1.5 seconds)	This category further includes three features: Number of short Pauses, duration of short pauses, the average duration of short pauses
Long Pauses (Pauses of more than 3 seconds)	This category further includes three features: No of long Pauses, duration of long pauses, the average duration of long pauses
Filled Pauses (hesitation sounds like um)	This category further includes three features: No of filled Pauses, duration of filled pauses, the average duration of filled pauses
Duration of voice speech	No of seconds, the subject spoke
Speech rate	The rate at which person speaks; duration of voice segments/ total duration
Articulation rate	The rate of speech excluding hesitation
Pause rate	The total duration of pauses/ total duration of speech

Pause rate without hesitation	The total duration of pauses excluding filled pauses/duration of speech

Table 1: The Proposed Acoustic features

- **Time-domain features:** short term energy, amplitude envelope, zero-crossing rate, and root mean square value
- **Frequency domain features:** Spectral centroid, spectral bandwidth
- **Mel frequency cepstral coefficients (12)**
- **Other features:** Gender, age education, MMS score ('Mini-Mental State Exam'), CDR score ('Clinical Dementia Rating Scale').

3.4. Dimensionality Reduction:

The second step is dimensionality reduction. This step transforms the data from high dimension to low dimension.

The principal component analysis (PCA) method is used for dimensionality reduction. This is also called a feature projection method where old features are projected to give a new feature set that is less in dimension but preserves the information and relationships of old features [1][7]. There is minimum information loss. This method can result in a simple and general classifier model which can perform better on a new data set.

We considered 25 principal components (the most significant features) for further processing of classification.

3.5. Classification

The next step is classification which classifies the data into systematic groups or categories.

Machine learning technology is known for many applications. One of them is Classification using Pattern Recognition [5][10]. Different Classifiers are tried and tested on dataset features for the categorization task of Alzheimer and non-Alzheimer patients. They are random forest, decision tree, Logistic regression with gradient descent, support vector machine (SVM) with Gaussian kernel, SVM with linear kernel etc.

The 6 fold cross-validation method is also chosen to find the best classifier model for the categorization task [4].

3.6. Decision:

The machine learning classification produces the knowledge model, also known as the classifier model. This model can be used on new data set to make decisions whether the person is suffering from the disease or not.

4. Experimental Setup

The study has been experimented on Pitt data set available through Dementia bank [2]. This dataset has transcripts and audio files speech recordings of subjects that were gathered as a part of a larger protocol administered by the Alzheimer and Related Dementias Study at the University Of Pittsburgh School Of Medicine. Subjects belong to different categories such as Alzheimer's disease, vascular dementia, Parkinson's disease, and healthy persons, etc. The recordings were recorded for different tasks such as word fluency task, picture description task, story recall task, and sentence construction.

In our experiment, we have considered the recordings of cookie theft picture where a picture of 'Cookie Theft stimulus photo' is shown to the subjects, and subjects are supposed to describe the scene depicted there.

The data set goes to preprocessing before applying the proposed approach. In the preprocessing, the recordings which are of our interest are considered whereas others are excluded. For example, the recordings which have a voice signal hardly audible or recordings which do not belong to the category of Alzheimer's or healthy individual. These types of recordings may make the classifier do wrong predictions.

After preprocessing, the dataset consisted of a total of 290 audio files on which the proposed approach was applied. 155 files belonged to Alzheimer's patients while the remaining 135 belonged to healthy individuals.

5. Results

Applying the proposed approach on the data set available after pre-processing, we were able to classify the subjects with an accuracy of 77% using Logistic regression and 76% using a support vector machine algorithm with a Gaussian kernel.

The resulting metrics of Logistic Regression, SVM for Alzheimer's patients are as shown in the below table [2].

Classifier/Metric	Accuracy	Precision	Recall	F1 Score
Logistic Regression	0.77	0.72	0.84	0.75
SVM (Guassian)	0.76	0.77	0.82	0.77

Table 2: Results of 10 fold Classification

For further insights, figure 2 shows the Receiver Operating Characteristics (ROC) curves of Logistic Regression and Support Vector Machine model. Logistic Regression shows the Area under Curve (AUC) score of 0.84 for mean of all folds of cross validation while SVM model shows the score of 0.87.

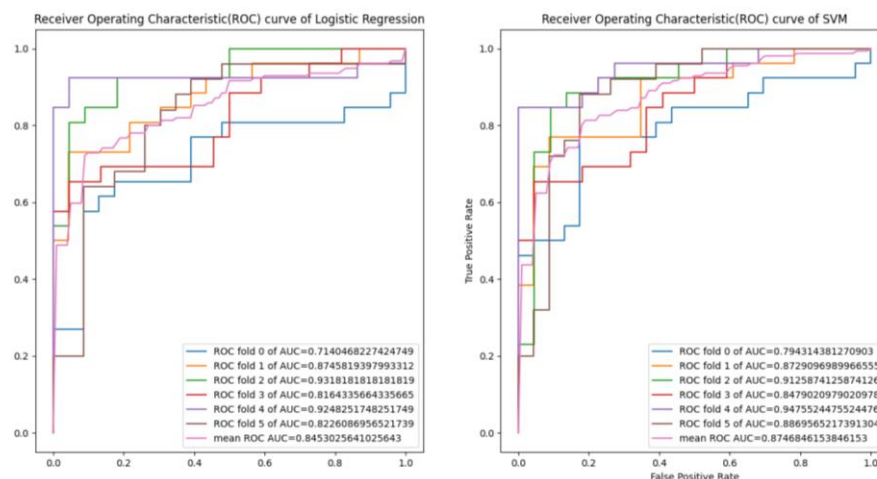


Fig. 2. ROC curves of Logistic Regression and Support Vector Machine Algorithms

6. Conclusion

The work on automatic analysis of spontaneous speech (ASSA) for Alzheimer's diagnosis is a wide-ranging study. There are different kinds of features and different ways to find those features which can be used as detectable indicators of AD. We just focused mainly on vocalization features related to the fluency of speech for the analysis and we were able to diagnose 77% accurately. Although the results need to be improved for accurate prediction, still they surpass the results of other studies related to these features. Future work may focus on more such features which can give us an accurate diagnosis.

References

- [1] ASonawale, Swati & Ade, Roshani. (2015). Dimensionality Reduction: An Effective Technique for Feature Selection. International Journal of Computer Applications. 117. 18-23. 10.5120/20535-2893.

- [2] Becker, J. T., Boller, F., Lopez, O. L., Saxton, J., & McGonigle, K. L. (1994). The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. *Archives of Neurology*, 51(6), 585-594.
- [3] Ben Ammar, R., & Ben Ayed, Y. (2018). Speech Processing for Early Alzheimer's Disease Diagnosis: Machine Learning-Based Approach. 2018 IEEE/ACS 15th International Conference on Computer Systems and Applications (AICCSA).
- [4] Berrar, Daniel. (2018). Cross-Validation. 10.1016/B978-0-12-809633-8.20349-X.
- [5] Chaki, J., Dey, N., Shi, F., & Sherratt, R. S. (2019). Pattern Mining Approaches used in Sensor-Based Biometric Recognition: A Review. *IEEE Sensors Journal*, 1–1.
- [6] Christy, A., Naveen Renold, J., Jesudoss, A., Mercy Theresa, M., Anto Praveena, M. D., & Sujihelen, L. (2021). Improving the Mental Health of Senior Citizens using Pygame. 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS)
- [7] Cunningham, P. (n.d.). Dimension Reduction. *Cognitive Technologies*, 91–112. doi:10.1007/978-3-540-75171-7_4
- [8] Damián Solís Rosas ; Saúl Tovar Arriaga ; Marco Antonio Aceves Fernández – ‘Search for Dementia Patterns in Transcribed Conversations using Natural Language Processing’
- [9] Fernandez Montenegro, J. M., & Argyriou, V. (2015). Cognitive tests for the diagnosis of Alzheimer's disease based on virtual environments. *IET International Conference on Technologies for Active and Assisted Living (TechAAL)*.
- [10] Jayashree Rajesh Prasad, "Pattern Recognition: Possible Research Areas and Issues" *IJCSN International Journal of Computer Science and Network*, Volume 3, Issue 5, October 2014
- [11] K. López-de-Ipiña et al., "Automatic analysis of Categorical Verbal Fluency for Mild Cognitive impairment detection: A non-linear language-independent approach," 2015 4th International Work-Conference on Bioinspired Intelligence (IWOBI), 2015, pp. 101-104, DOI: 10.1109/IWOBI.2015.7160151.
- [12] Luz, S. (2017). Longitudinal Monitoring and Detection of Alzheimer's Type Dementia from Spontaneous Speech Data. 2017 IEEE 30th International Symposium on Computer-Based Medical Systems (CBMS).
- [13] Patel, R., Liu, J., Kewei Chen, Reiman, E., Alexander, G., & Jieping Ye. (2009). Sparse Inverse Covariance Analysis of human brain for Alzheimer's disease study. 2009 ICME International Conference on Complex Medical Engineering.
- [14] Robin J, Harrison J, E, Kaufman L, D, Rudzicz F, Simpson W, Yancheva M: Evaluation of Speech-Based Digital Biomarkers: Review and Recommendations. *Digit Biomark* 2020;4:99-108. DOI: 10.1159/000510820
- [15] S. Kato, A. Homma, T. Sakuma, and M. Nakamura, "Detection of mild Alzheimer's disease and mild cognitive impairment from the elderly speech: Binary discrimination using logistic regression," 2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2015, pp. 5569-5572, DOI: 10.1109/EMBC.2015.7319654.
- [16] Supe, Ujjwala & Supe, Jayant. (2018). Alzheimer: A disease of the brain.
- [17] Toth, L., Hoffmann, I., Gosztolya, G., Vincze, V., Szatloczki, G., Banreti, Z., ... Kalman, J. (2018). A Speech Recognition-based Solution for the Automatic Detection of Mild Cognitive Impairment from Spontaneous Speech. *Current Alzheimer Research*, 15(2)

An Empirical Study on the Impact of IT Services on End-User Satisfaction for University of Petroleum and Energy Studies

Jatendar Sharma¹

Senior Manager IT, Office of Information Technology, University of Petroleum & Energy Studies, Village-Bidholi, Dehradun – 248007, Uttarakhand, India

jatendar@dd.upes.ac.in

Gagan Deep Singh², Sunil Kumar²

School of Computer Science, University of Petroleum & Energy Studies, Village-Bidholi, Dehradun – 248007, Uttarakhand, India

gagan@ddn.upes.ac.in, skumar@ddn.upes.ac.in

Sorabh Gupta³

Panipat Institute of Engineering and Technology, Grand Trunk Rd, Samalkha, Haryana 132102,
sorabh.it@piet.co.in

Abstract : In this paper, the authors emphasize the IT Services that are provided to the end-users in the University of Petroleum and Energy Studies, Dehradun. The author gathered the data through online surveys. The data were collected from more than 775 end users. The survey was conducted biannually. After analyses of the first survey results, the data is reviewed for the weak areas of IT services and then the IT services were improved afterwards. Then the next survey was conducted for end-users of UPES. Then analysis was done on the data gathered from the second survey and compared then it with the first survey results. The user satisfaction was improved in the second survey. The satisfaction index is computed for both surveys. The improvement done in IT services reflects the satisfaction level of end-users and it is improved by a satisfaction index of 0.13. The result was satisfactory and it can be concluded that end users are very much dependent on IT services for their routine works. The IT services have played a vital role to increase the outcome of the end-users performance and this reflects through the satisfaction indexed computed through the data gathered from biannual surveys.

Keywords: IT Services, Online surveys, UPES, user satisfaction, satisfaction index.

1. Introduction

University of Petroleum and Energy Studies is a leading university with an ambitious vision and unique approach. UPES offers its students a world-class academic infrastructure, multimedia-enabled classrooms, modern research lab facilities for curricular and extracurricular activities.

UPES has diversified courses in the fields of Engineering, Management, Law, Design, Health Science and Modern media.

UPES has more than 1000 plus faculty and staff who are day in day out working to achieve the best results in the industry. IT is offering its services to all UPES users. Common services of IT that users are leveraging includes Internet, Intranet, Mailing and support. A robust helpdesk <https://upes-hd.ddn.upes.ac.in/upeshd> [1] is deployed here which act as a single point of contact for all the queries raised by users. Calls are categorized into service and incident based on their severity. Every call has the response and resolution time associated with it. Users can reach the helpdesk by the following means:

1. Users can call the helpdesk from their extensions.
2. Users can visit the helpdesk.
3. Users can log a call through the portal. <https://upes-hd.ddn.upes.ac.in/upeshd>.

As soon as the call is logged by the user, then they will get the reference number which will help the user to trace the ticket status till its satisfactory closure.

The main goal of IT services and infrastructure deployment is to maximize the profit of its stakeholders which will be directly proportional to organizational growth. IT adoption will help in increasing throughput by IT

staff revitalization, resources mobilization, process restructuring to serve users better and thus performance gets better.

2. Literature Review

The IT service is considered the backbone of any organization. The downtime of IT services directly affects its business and these days standard of 99.999 of uptime has been made as to the benchmark for it. The adoption of IT and its dependency requires raising the profit and growth of the organization and all its stakeholders. From the previous studies, it has been observed that IT made a major changeover through total end-user satisfaction and brought a radical improvement in organizational performance [2].

The leading industries of any country rely on the availability of skilled professionals and depend on the successful implementation of advanced technologies and software applications. But a quantitative and qualitative imbalance of trained professionals varies according to business cycles, rapid advancements in technology and demographic factors [3].

Organizations and Indian firms urgently require to hire professionals fluent in adopting IT digital work environment to keep pace with and an ever-changing business needs for introducing their next-generation products [4].

The delivery of services for the IT sector is the collaborative approach of specific proficiencies in the IT and business domains that are defined by this customer-centric perspective. Now, IT needs frequent updates through a digital environment for any business where the dynamic needs of the business demand IT-enabled solutions from their service providers to work along changing environments. In this context, IT solutions become increasingly “servitized” [5].

The capital relational impact on the satisfaction level of IT users correlates from both the perspectives i.e. users and IT departments. When users and IT support have a significant amount of cooperation and respect for one another, then both IT professionals and users may have a great desire to share information among others. As a result, they will be able to earn a lot of respect and will be able to expect similar returns in the future [6].

However, much researches in IT services signifies the importance of building trust between IT professionals and end-users [7]. Periodical training sessions for end-users improves IT usability, perhaps deliver the best IT services, and surely meet the expectations of user desires.

The IT department must conduct time to time study for the evaluation of the satisfaction of its end-users. Through this type of study, the improvement in the IT services can be enhance and hence performance will be increased.

3. Methodology and process followed

To conduct the study both qualitative and quantitative approaches were used. In the qualitative approach, theoretical literature from different sources is read which helps to make a framework for questions [8]. Perception meetings with users were also conducted to find out the pain areas. A set of 10 questions were prepared to keep all the points into consideration. After this quantitative approach was applied and is performed by survey questionnaire. The methodology adopted is shown in figure 1.

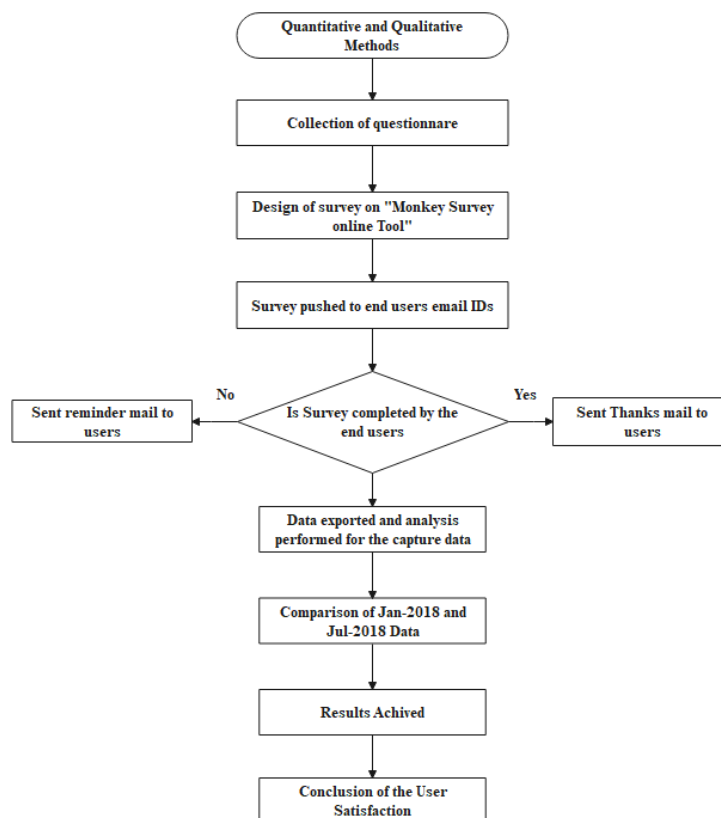


Fig. 1. The methodology adopted and the process followed.

- The survey was administered in online mode (Survey Monkey). www.surveymonkey.com [9].
- Data of end-users were collected from HR. HR extracted the data from the HRM module of the SAP application. <https://sapep.delhi.upes.ac.in:50401/irj/portal> [10]. The First name, Last Name and Email id of users are taken from the HRM module. University of Petroleum and Energy Studies is using five modules of SAP: (Human Resource Management) HRM, (Student Lifecycle Management) SLCM, (Finance and Controlling Module) FICO, (Finance and Contract Accounting) FICA, and (Material Management) MM.
- This data is uploaded to the tool (Survey Monkey) in the .csv format.
- Each end-user was sent an individual mail to participate in the survey. Office365 is the Exchange server that was used to send the emails. <https://www.office.com/> [11].
- Automated thanks mail was sent to all the users who have completed their survey.
- All those users who haven't participated in reminder mail were sent on the specified date and time configured on the survey tool.
- The reports are extracted from the tool in ".csv" format.
- Results have been evaluated independently for each section from the ".csv" format.
- Statistical Mode, Average & Standard Deviation from the Mean have been used to derive the overall satisfaction levels these were calculated through excel.
- The choice "Neutral" has been considered as an indicator of "No Relevance" or "No Awareness" about the service. However, the score of "Neutral" is used to derive the overall satisfaction index.
- The satisfaction levels range between 1 to 7 where 1 stands for "Very Dissatisfied", 2 for "Dissatisfied", 3 for "Somewhat Dissatisfied", 5 for "Somewhat Satisfied", and 6 for Satisfied "Satisfied and 7 for "Very Satisfied" as shown in figure 2.
- The survey scores have been evaluated vertically and horizontally. A vertical evaluation refers to a question wise scrutiny of all responses received and a horizontal evaluation refers to a user wise scrutiny of responses. As shown in figure 3.

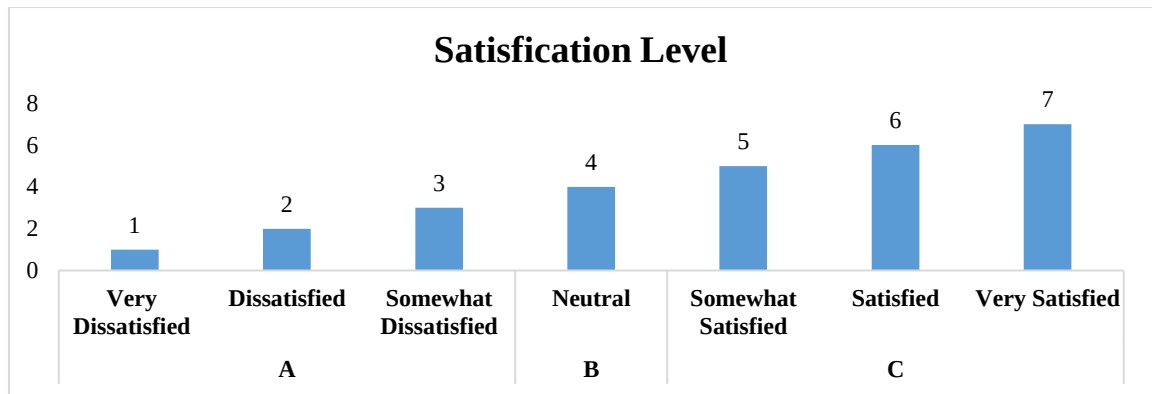


Fig. 2. Satisfaction Levels

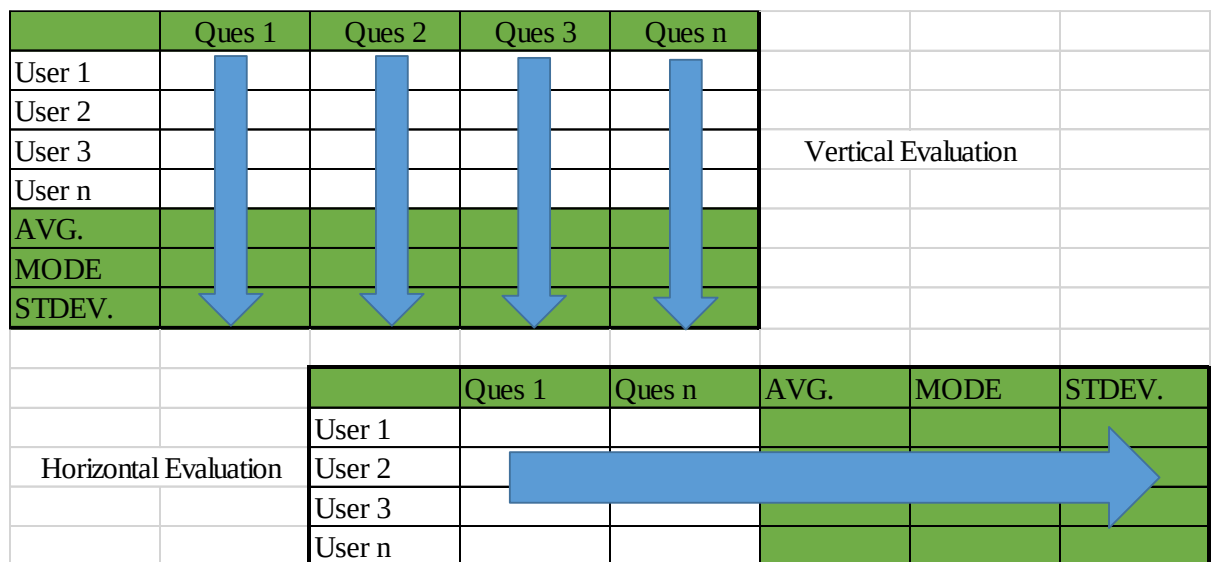


Fig. 3. Vertical and Horizontal Evaluations

- Horizontal evaluation is based only on the averages of scores on the 10 questions answered by the users. The scores of 1 through 7 have been substituted by -3 to 3 with 0 as the “Neutral” choice.
- All the questions which were skipped by users are considered neutral.
- Crucial Parameters are taken as per table 1.
- Derivation for these survey results is mentioned in the conclusion.

Responses	Indicates the total number of responses considered to calculate statistical figures
Mode	Indicates the response appearing the maximum number of times.
Average	An average of the responses given for a particular question.
Standard Deviation	Indicates the degree of deviation from the average.
Satisfaction (SR- +) Range	Derived by subtracting and adding standard deviation from average.

Satisfaction Index	An Intuitively derived whole number ranging between 1 and 7. Depending on the satisfaction range.

Table 1. Parameter details.

4. Analysis of the Surveys conducted

The analysis of both the surveys conducted in January 2018 and July 2018 was performed using Vertical and Horizontal Evaluation.

4.1 Vertical Evaluation

The analysis of the first survey was done in Jan 2018 using the vertical evaluation technique and is shown in Table 2.

Q.NO	Responses	Mode	AVG	STDEV	SR-	SR+	SI
1	819	6	6.15	0.96	5.18	7.11	6.15
2	819	6	6.22	0.89	5.33	7.11	6.22
3	819	6	6.13	0.97	5.15	7.10	6.13
4	819	6	6.01	1.09	4.93	7.10	6.01
5	819	6	6.04	1.02	5.02	7.06	6.04
6	819	6	5.94	1.09	4.85	7.02	5.94
7	819	6	6.13	0.97	5.16	7.10	6.13
8	819	6	5.54	1.36	4.19	6.90	5.54
9	819	6	5.97	1.04	4.93	7.00	5.97
10	819	6	6.03	0.97	5.06	7.01	6.03
Jan-18						SI	6.02

Table 2. Vertical Evaluation January-2018.

The outcomes observed after vertical evaluation of the January 2018 survey are presented as below:

- More than 800 end-users participated in the survey of January 2018.
- The Mode was computed for all ten questions.
- The result of mode is 6 which indicates that maximum users are satisfied with the IT services.
- The average response of seven out of 10 questions is more than or equal to 6 but three questions are having less than 6.
- Standard Deviation of as high a value of 1 and above indicate a good level of dispersion of the data about the average and have led to a high SI score in a few cases as it is 6.22.
- SI for nine questions is calculated as “*Satisfied*” and for one question, it is “*somewhat satisfied*”.

The analysis of the survey conducted in July 2018 is shown in Table 3

Q.NO	Responses	Mode	AVG	STDEV	SR-	SR+	SI

1	783	6	6.30	0.86	5.45	7.16	6.30
2	783	7	6.35	0.81	5.54	7.16	6.35
3	783	6	6.23	0.90	5.33	7.13	6.23
4	783	6	6.12	1.02	5.1	7.14	6.12
5	783	6	6.15	0.91	5.23	7.06	6.15
6	783	6	5.98	1.1	4.89	7.08	6.03
7	783	7	6.22	0.96	5.26	7.18	6.25
8	783	6	5.80	1.18	4.62	6.97	5.80
9	783	6	6.10	0.95	5.14	7.05	6.13
10	783	6	6.14	0.88	5.26	7.02	6.14
Jul-18						SI	6.15

Table 3. Vertical Evaluation July-2018

The outcomes observed after vertical evaluation of the July 2018 survey are presented as below:

- More than 775 end-users participated in the July 2018 survey.
- The Mode of response to eight questions out of ten is given as 6 which indicates that users are satisfied.
- The remaining two questions out of ten are rated as very satisfied as shown in table 3.
- The average response for eight questions is more than or equal to 6 but two questions are having less than 6.
- Standard Deviation of as high a value of 1 and above indicate a good level of dispersion of the data about the average and have led to a high SI score in a few cases it is 6.35.
- SI for all ten questions calculated as “Satisfied”.

4.2 Horizontal Evaluation

The average of responses to all the questions answered by a user is used to evaluate the overall satisfaction of the user. Only considering the averages, a little above 92.79% of user falls in the satisfied category. Figure 4 shows the number of users with their average and their satisfaction level. Figure 4 also shows that only 7 users are completely dissatisfied with the IT services. Figure 5 shows Horizontal Evaluation for July 2018 that 94.24% of user's falls in the satisfied category and only 3 users are dissatisfied with the IT services.

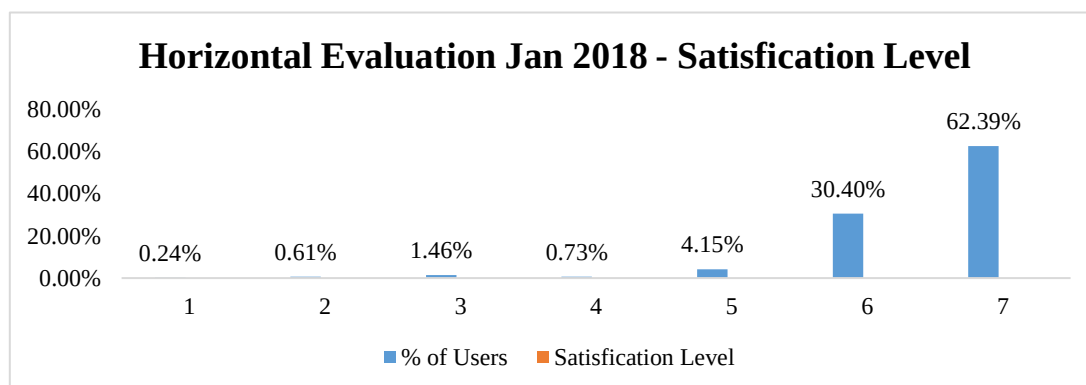


Fig. 4. Horizontal Evaluation for Survey Data conducted in January-2018.

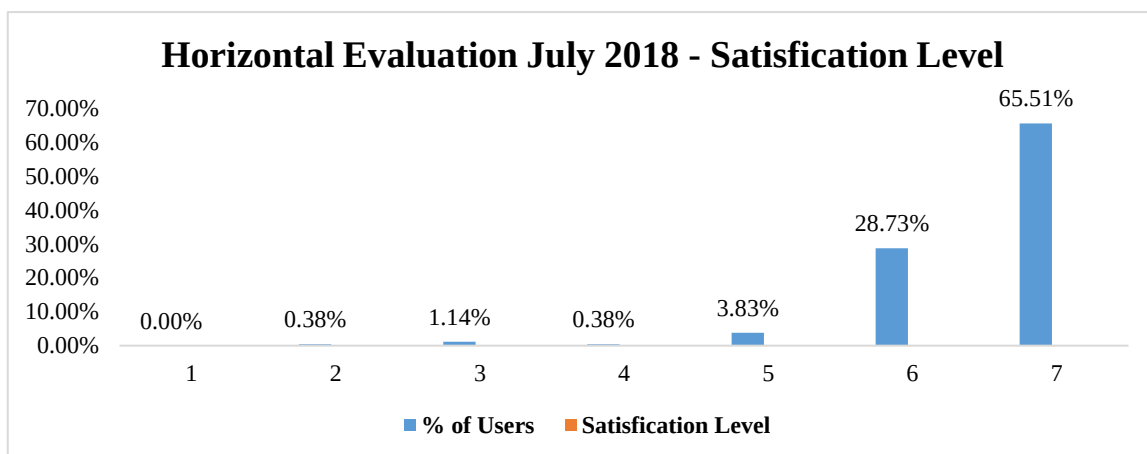


Fig. 5. Horizontal Evaluation for Survey Data conducted in July-2018.

5. Results

Equation 1 clearly depicts that the satisfaction of the users was increased in the second survey. The Satisfaction Index (SI) of the first survey was 6.02 and in the next survey, it increases to 6.15. The FS^i represents a satisfaction index for the survey conducted in January 2018 and SS^i represents the satisfaction index conducted for the survey in July 2018.

$$FS^i = 6.02$$

$$SS^i = 6.15$$

$$SI = SS^i > FS^i$$

Table 4 and Figure 6 shows the comparison of STDEV and SI for the January 2018 and July 2018 user satisfaction index. The graph shows that in the second survey Satisfaction Index is higher as compared to the first survey.

Months	Jan-18		Jul-18	
Q.No	STDEV	SI	STDEV	SI
1	0.96	6.15	0.86	6.30
2	0.89	6.22	0.81	6.35
3	0.97	6.13	0.90	6.23
4	1.09	6.01	1.02	6.12
5	1.02	6.04	0.91	6.15
6	1.09	5.94	1.1	6.03
7	0.97	6.13	0.96	6.25
8	1.36	5.54	1.18	5.8

9	1.04	5.97	0.95	6.13
10	0.97	6.03	0.88	6.14

Table 4. STDEV and SI comparative for Jan-2018 and Jul-2018

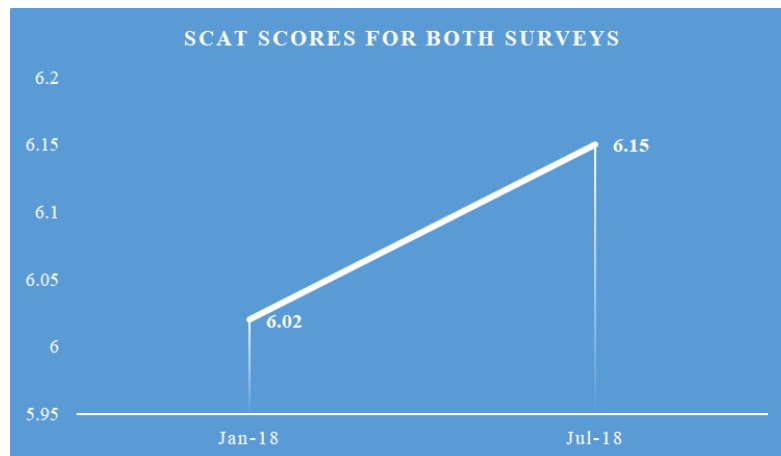


Fig. 6. STDEV & SI comparison with CSAT score.

The outcome of improved user satisfaction was achieved because of the initiatives done by IT after reviewing the January 2018 survey.

i. Firewall:

- Few rules were created in Firewall for dedicated bandwidth for users.
- Changes were done on content filtering categories which help end-users to access the internet smoothly.

ii. Systems:

- Systems were patched with the latest patches.
- RAM of systems was upgraded to 8GB. The system performance has increased.

iii. Helpdesk:

- SOP of helpdesk was prepared and shared with users which help them as a guide to log a call.
- Training and motivation were also given to end-users.

iv. Internet:

- One more internet connection was taken. Now the University has a primary and secondary Internet connection. If one goes down the load is shifted to another connection.
- The bandwidth of the Internet was increased from 1024Mbps to 1334Mbps.
- Now there was two ISP's in UPES.
 - NKN: 1024Mbps
 - Vodafone: 310Mbps.

6. Conclusion

The research paper focuses on the end-user satisfaction who are availing IT services at UPES. From equation 1 it can be concluded that the satisfaction index of July 2018 is greater than the SI of the January 2018 survey. So, the authors are stating that Satisfaction Index can be increased by achieving the end-user satisfaction level through improvement in IT services. In this study, the rise of 0.13 satisfaction index has been achieved. Hence, it is very important to review and improve the IT services of the organization from time to time because it's a continual process of improvement.

This research paper suggests to IT departments of all organizations that such types of surveys and their review are very necessary for ensuring the pace with ever-changing business expectations and to increase the growth of the organization.

As the satisfaction index of end-users will increase towards IT services, the users will focus more on their core duties which will have a direct positive impact on the organization's goal.

References

- [1] "GLPI - Authentication." <http://upes-hd.ddn.upes.ac.in/upeshd/> (accessed Sep. 09, 2021).
- [2] N. S., B. S., Kumar V., and C. T. S., "State of the art literature review on performance measurement," *Comput. Ind. Eng.*, vol. 60, no. 2, pp. 279–290, Mar. 2011, doi: 10.1016/J.CIE.2010.11.010.
- [3] P. Kapetanopoulou and A. Kouroutzi, "An Empirical Study of Drivers, Barriers, and Cost Efficiency of Information Systems in Greek Industry," *Appl. Sci.* 2021, Vol. 11, Page 3475, vol. 11, no. 8, p. 3475, Apr. 2021, doi: 10.3390/APP11083475.
- [4] C. C. Cheng and D. Krumwiede, "The role of service innovation in the market orientation—new service performance linkage," *Technovation*, vol. 32, no. 7–8, pp. 487–497, Jul. 2012, doi: 10.1016/J.TECHNOVATION.2012.03.006.
- [5] G. Barreto, P. D. Battaglin, and S. Varga, "Ensuring Efficient IT Service Management to Increase Information Systems Availability," *J. Inf. Syst. Eng. Manag.*, vol. 4, no. 4, p. em0108, Dec. 2019, doi: 10.29333/JISEM/6352.
- [6] T. Almeida, J. B. De Vasconcelos, and G. Pestana, "A knowledge management architecture for information technology services delivery," *Iber. Conf. Inf. Syst. Technol. Cist.*, vol. 2018-June, pp. 1–4, Jun. 2018, doi: 10.23919/CISTI.2018.8399202.
- [7] R. Yandri, Suharjito, D. N. Utama, and A. Zahra, "Evaluation Model for the Implementation of Information Technology Service Management using Fuzzy ITIL," *Procedia Comput. Sci.*, vol. 157, pp. 290–297, Jan. 2019, doi: 10.1016/J.PROCS.2019.08.169.
- [8] K. R. Pillai, P. Upadhyaya, A. V. Prakash, B. S. Ramaprasad, H. V. Mukesh, and Y. Pai, "End-user satisfaction of technology-enabled assessment in higher education: A coping theory perspective," *Educ. Inf. Technol.* 2021 264, vol. 26, no. 4, pp. 3677–3698, Jan. 2021, doi: 10.1007/S10639-020-10401-2.
- [9] "SurveyMonkey: The World's Most Popular Free Online Survey Tool." <https://www.surveymonkey.com/> (accessed Sep. 09, 2021).
- [10] "Overview - SAP NetWeaver Portal." <https://sapep.delhi.upes.ac.in:50401/irj/portal> (accessed Sep. 09, 2021).
- [11] "Microsoft 365 | Combines Office 365, Windows 10, Security." <https://www.microsoft.com/en-in/microsoft-365>.

Machine Learning Techniques for Heart Disease Prediction

¹Ashima Arya, ²Mitu Sehgal, ³Neha Bhatia

^{1,2,3}Assistant Professor, Department of IT PIET ,Samalkha
ashiarya@gmail.com

Abstract: A body is essential to everyone. It consists of several organs out of which Heart is counted as one of the major part of body. It helps in pumping the blood and supplies it in the whole body. Heart disease is one of the major threats to any human being. It may cause anyone life's to death. Predicting cardiac disease is a tough task in medical science. Data Analytics is the approach which is used to predict the disease from the related information. Huge amount of data has been recorded and stored by the medical centers. The collected data helps in prediction of future outcomes. Artificial Neural Network (ANN), Decision Tree, Fuzzy Logic, K-Nearest Neighbour (KNN), Logistic Regression, Linear Regression, and Support Vector Machine are some of the Data Mining and Machine Learning approaches used to predict heart disease (SVM). This paper provides sight towards existing algorithm giving overall summary of the existing work. Naive Bayes, Logistics Regression, and Random Forest are some of the algorithms that have been considered. The accuracy of model has been calculated using each algorithm. When it comes to predicting heart disease, the algorithm with the utmost precision is used.

Keywords: Machine learning, Prediction of Heart Disease, learning classifiers.

1 Introduction

In a normal daily routine, everyone is undergoing the busy schedule which results into stress and anxiety. This leads to multiple diseases within a human body. Heart disease, cancer, and brain issues are only a few of the major ailments. Heart disease is one of the most common situations that reduce public's lives [6]. Heart disease claims the lives of 17.5 million people each year. It is a kind of disease that affects the whole body functioning. The problem exists when it comes to correctly predict the cause of disease and to know how it can be treated [7]. Multivariate regression analysis can be used to create a prediction model. As we all know, digital technology is rapidly evolving. As a result, healthcare facilities retain a large quantity of data in their databases, which can be difficult to analyse. Due to the Data Mining and Machine Learning techniques it is made really easy to analyze and predict the data. A dataset can be collected from the huge database for creating the model [8].

The types of cardiac disorders are as follows: The word "cardio" is derived from the word "heart." All cardiac disorders fall under the umbrella of cardiovascular diseases. Coronary heart disease, Angina pectoris, Congestive heart failure, Cardiomyopathy, Congenital heart disorders, and coronary artery diseases (narrowing of the coronary arteries) are the numerous types of heart disease. The coronary arteries supply the heart with blood and oxygen. A high number of persons fall ill or die as a result of it. It is well-known as the most common type of heart disease [9]. High blood glucose disturbs the blood vessels and nerves that run the heart and its working as a result of diabetes. If a person has diabetes for a long time, there is a considerable possibility that they will develop heart disease in the future. Additional reasons that add to heart disease in people with diabetes consist of smoking, which increases the risk of heart disease, and high blood pressure, which makes it challenging for the heart to pump blood and sets stress on the heart, fading blood vessels [10]. Heart disease and obesity are also interrelated to irregular cholesterol levels. Family history can also be the reason to have a heart disease in someone. The other risk factors include age, gender, stress and unhealthy diet. Chance of having a heart disease is more in old age. Men's according to the study have a greater risk of heart disease than women.

We can anticipate the possibility of heart disease based on the factors presented earlier in this research. Several Data Mining and Machine Learning approaches can be used to make the prediction. The chosen algorithms have been taken into account and implemented onto the dataset according to its requirements. The aim of this study is to increase precision and create the system well-organized in foretelling outcomes.

2 Literature Review

The Health care industry is the one the vast application area in which there is a huge number of resources that holds a tedious job to be handled and operated manually. A great amount of research and study has been done on the heart disease prediction systems using various non invasive methods such as algorithms based on machine learning.

Sahaya Arthy et.al (2018) [1] focuses on data mining techniques to predict various heart disease that are prevailing in the world. The main motive behind this study is to extract and analyze the unknown patterns in the data and to focus on the enhancement of accuracy status of any dataset. They also have provided descriptive information about the databases, tools used such as Rapid Miner, Weka, Data melt, Apache Mahout etc. Based on the results, it proves that combination of two or more algorithm produces a better and accurate output than a single algorithm use.

R. Sharmila et al, (2018) [2] proposed an efficient heart disease system that predicts the heart related issues using data mining. Nowadays, it plays a major role in predicting various diseases in comparatively reduced number of tests. It recommended using big data tools like SVM, HDFS for a better and improved result. A comparative study is done on parallel and sequential SVM which yielded that better computation time is achieved in parallel SVM.

Chala Beyene et al(2018), [3] suggested predicting the prevalence of any heart related disease using mining algorithms. The major objective of this research is to sense and treat heart problems initial on and in a relatively short period of time. Various attributes are taken into account for better evaluation of the person extent of sufferings from the heart disease. .Dataset analysis has been done using Weka Software.

Amandeep Kaur et.al(2018) [4] emphases on approaches that are suitable and helpful in forecasting disease by taking out old data records of patients make use of several algorithms such as K-nearest neighbour, artificial neural network, Nave Bayes, and Support vector machine. From this we were able to reach to a conclusion that SVM provided a highest accuracy rate as compared to other mining algorithms.

Sushmita Roy Tithi(2019) [5] proved that mining techniques can be used to predict cardiac disease. This aids the medical practitioner's analysis and decision-making. ECG is a medical diagnosis of the activity of the heartbeat of a human being as the conventional methods are very time consuming and tedious to work on. The apt algorithm can be used to classify ECG signals. This recommended approach increases ECG classification accurateness, letting for extra particular finding.

3 Proposed Methodology

Application of any algorithm depends on the nature of dataset. The dataset which has been used consists of 14 attributes which are as below:

TABLE I. ATTRIBUTE TAKEN IN DATASET FOR PREDICTION

1. Age: the number of years between one's birth and one's death.
2. Gender: Male = 1, Female = 0)
3. Cp: Types of Chest Pain a) Type 1: typical angina b) Type 2: atypical angina d) Type4: asymptomatic c) Type3: nonanginal pain
4. Trestbps: Vital sign at rest (in torr on admission to the hospital)
5. Chol: cholesterol serum concentration in milligrammes per deciliter
6. Fbs: (1 = true; 0 = false) (fasting blood glucose > 120 mg/dl)
7. Thalach: reached maximum heart rate

8. Restecg: electrocardiographic data at rest
Type 0: normal; Type 1: aberrant STT wave (T wave inversions and/or ST elevation or depression of > 0.05 mV)
- c) Type 2: left ventricular hypertrophy is suspected or confirmed.
9. Exang: exercise-induced angina (yes or no).
10. Oldpeak: Exercise-induced ST depression compared to rest
11. Slope: the slope of the ST segment of the height exercise.
i. Type 1 is upsloping; ii. Type 2 is flat; and iii. Type 3 is downsloping.
12. Ca: number of large vessels (03) that are flourosopy coloured
13. Thal: 3 denotes normal, 6 denotes a permanent defect, and 7 denotes a reversible defect.
14. Num: a cardiac problem diagnosis (angiographic disease status)
Type 0 (zero): a 50 percent reduction in diameter
Type 1: a diameter narrowing of more than 50%

4 Machine Learning Classifier

1. Logistics Regression

A Machine Learning classification methodology called logistic regression is used to guess the probability of a clear-cut dependent variable. The dependent variable in logistic regression is a binary variable that involves data coded as 1 (yes, success, etc.) or 0 (no) (no, failure). In other words, as a function of X, the logistic regression model predicts $P(Y=1)$.

In Data Modeling, logistic regression is one of the best used methods for suitable models for categorical data, mostly for binary response data. It goes to a class of models identified as generalized linear models, and it is the most essential (and possibly most commonly used) of them. The minimal possibilities of the training data are conserved using logistic regression. The model's coefficients also provide some vision into the relative relevance of each input variable. For example,

To conclude if an email is spam (1) or not (0)

Whether or not the tumor is cancerous (1) (0)

Types of Logistic Regression

- 1) Logistic Regression (Binary): There are just two possible results for the categorical response. For example: Is it Spam or Not?
- 2) Logistic Multinomial Regression: Without any sorting, there are three or more categories. Predicting which food is preferred more, for example (Veg, Non-Veg, Vegan)
- 3) Logistic Ordinal Regression: Three or more categories, each having its own order.

Advantages of Logistics Regression

1. When the dataset is split by a linear line, Logistic Regression performs well.
2. While logistic regression is less prone to overfitting, it can overfit in datasets with a high number of dimensions. In these cases, regularisation (L1 and L2) approaches should be used to minimise over-fitting.
3. Logistic regression provides not only a measure of a predictor's relevance (coefficient size), but also the direction of relationship (positive or negative).

4. Logistic regression is more straightforward to implement, analyse, and train.

Limitations of Logistics Regression

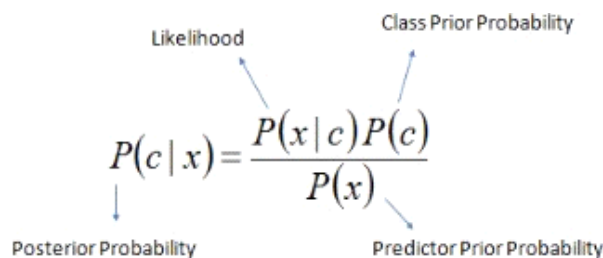
1. The assumption of linearity between the dependent and independent variables is the main drawback of Logistic Regression. Data is rarely linearly separable in the actual world. The majority of the time, data is a muddled mess.
2. Logistic Regression should not be employed if the number of observations is less than the number of features; otherwise, it may result in overfitting. Only discrete functions may be predicted using logistic regression. As a result, Logistic Regression's dependent variable is limited to the discrete number set. This limitation is troublesome in and of itself, as it makes continuous data prediction impossible.

2 Naive Bayes

The Bayes Theorem is used to create a collection of classification algorithms known as Naive Bayes classifiers. It is a family of algorithms that share a similar idea, namely that each pair of features being classified is independent of the others. It's a dependable and adaptable tool that has endured the test of time. The Naive Bayes Classifier approach is based on the Bayesian theorem and is especially useful when the inputs have a high dimensionality.

Everything in statistics relies around hypotheses. We establish a hypothesis (an educated guess) about how the world works and then set about gathering evidence to test it. Assuming an occurrence and predicting the corresponding outcome is how Naive Bayes works.

The Bayes Theorem calculates the likelihood of an event occurring given the probability of a previous event. The following equation expresses Bayes' theorem mathematically.



The diagram shows the Bayes' Theorem formula:
$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$
 with arrows pointing from labels to parts of the formula:

- Likelihood** points to $P(x|c)$
- Class Prior Probability** points to $P(c)$
- Posterior Probability** points to $P(c|x)$
- Predictor Prior Probability** points to $P(x)$

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

Fig1. Bayes Formula

Here,

- The posterior probability of class (target) given predictor is $P(c|x)$ (attribute).
- $P(c)$ is the class prior possibility.
- The likelihood is $P(x|c)$, which is the probability of a predictor given a class.
- $P(x)$ is the predictor's prior possibility.

Advantages of Naive Bayes

1. A Naive Bayes classifier outperforms other models when the assumption of independent predictors holds true.
2. To estimate the test data, Naive Bayes requires a little amount of training data. As a result, the training duration is shorter.

3. Naive Bayes is similarly simple to use.

Limitations of Naive Bayes

1. The assumption of independent predictors is the main premise of Naive Bayes. All of the qualities in Naive Bayes are assumed to be mutually independent. In actual life, getting a collection of predictors that are totally independent is nearly impossible.
2. If a categorical variable in the test data set has a category that was not included in the training data set, the model will assign a probability of 0 (zero) and will be unable to make a prediction. This is commonly mentioned to as Zero Frequency. We can apply the smoothing approach to remedy this. Laplace estimation is one of the best simple smoothing techniques.

3. Random Forest

An ensemble of decision trees is referred to as a Random Forest. We have a pool of decision trees in Random Forest. Each tree provides a classification for a new object based on attributes, which we refer to as the tree - votes for that class. The classification with the highest votes is chosen by the forest (over all the trees in the forest).

As the name implies, a random forest is made up of a vast number of single decision trees that work together as an group. Every tree in the random forest creates a class prediction, and the class with the maximum votes turn into the likelihood of our model.

The following is how each tree is planted and grown:

1. If the number of cases in the training set is N , a random sample of N cases with replacement is taken. This sample will serve as the tree's training set.
2. If there are M input variables, a number m is supplied so that m variables are randomly chosen from the M at each node, and the best split on this m is used to split the node. Throughout the development of the forest, the value of m is kept constant.
3. Every single tree is full-grown to its supreme potential. Snipping is not an option.

Scikit-Learn is used to design a decision tree and train (fit) it on data. We feed the model both the features and the labels during training so that it may learn to classify points using the features. On the training data, we can check the accuracy of our model. Because we gave it the answers (y) for training and did not limit the depth of the tree, we notice that it gets 100 percent accuracy, which is what we expected. It turns out that a decision tree's ability to completely understand the training data can be a drawback because it can lead to overfitting.

Advantages of Random Forest

1. The Random Forest algorithm is created on the bagging algorithm and employs Ensemble Learning. It grows as many trees as possible on a subset of the data and then merges the results of all of the trees. As a result, the overfitting problem in decision trees is reduced, as is the variance, which increases accuracy.
2. Random Forest can be used to address difficulties comprising classification and regression.
3. Both categorical and continuous variables function well with Random Forest.
4. Random Forest can handle missing values automatically.

5. Random Forest does not require feature scaling (standardisation and normalisation) because it uses a rule-based method rather than distance calculation.

Limitations of Random Forest

1. Complexity: Random Forest generates a large number of trees (as opposed to a single tree in a decision tree) and then mixes their outputs. In the Python sklearn library, it builds 100 trees by default. This approach necessitates a significant increase in processing power and resources. On the other hand, a decision tree is straightforward and does not necessitate a large amount of computer power.

2. Longer Training Period: Random Forest takes significantly longer to train than decision trees since it generates several trees (instead of just one) and makes decisions based on the majority of votes.

5 Result

TABLE II. DETAILS OF PATIENTS HAVING HIGH CHANCES OF HEART DISEASE

> 35 years old Man = Gender Pain in the chest = 0 Blood Pressure is greater than 100. Cholesterol in the blood > 131 Fasting 0 for blood sugar Result of electrocardiography = 0 The maximum heart rate is more than 71 beats per minute. Depression about 1.6 Slope >= 1 Vessels about 1.1 Thal > 2.5		
S.No	Algorithms Used	Accuracy Achieved
1.	Logistics Regression	74.8%
2.	Naïve Bayes	74.6%
3.	Random Forest	80.2%

6 Conclusion

This paper includes an overview of the present method as well as a synopsis of previous work. In the above paper, we have explored numerous classification algorithms. Also we have studied different technologies like naive bayes, Log1istic Regression, SVM, KNN, Random Forest and so on to compare the graph of accuracies obtained by each one of them. We have worked on 3 different datasets to analyze and experience a better result with more precise and accurate results.

From our study, we chose some combinations of techniques to study the datasets. Our approach leads us to the result that random forest provided the highest accuracy of all the approaches we have applied individually on the dataset.

Therefore in a nutshell, it seems very useful and beneficial to adopt this approach that will help in diagnosing the disease quickly and easily and lessen the medical costs of the patient.

References

- [1] A.Sahaya Arthy, G. Murugeswari” A survey on heart disease prediction using data mining techniques “(April 2018)
- [2] R. Sharmila, S. Chellammal, “A conceptual method to enhance the prediction of heart diseases using the data techniques”, International Journal of Computer Science and Engineering,(May 2018)
- [3] Mr. Chala Beyene, Prof. Pooja Kamat, “Survey on Prediction and Analysis the Occurrence of Heart Disease Using Data Mining Techniques”, International Journal of Pure and Applied Mathematics (2018)
- [4] Amandeep Kaur, Jyoti Arora “Heart disease prediction using data mining techniques: asurvey” (2018)
- [5] Sushmita Roy ,Tithi, Afifa Akhtar Dept “ECG data analysis and heart disease prediction using machine learning algorithms”.(2019)
- [6] Mohan, Senthilkumar, Chandrasegar Thirumalai, and Gautam Srivastava. “Effective heart disease prediction using hybrid machine learning techniques.” (2019)
- [7] Khourdifi, Youness, and Mohamed Bahaj. “Heart disease prediction and classification using machine learning algorithms optimized by particle swarm optimization and ant colony optimization.”(2019)
- [8] Rajdhan, Apurb, et al. “Heart disease prediction using machine learning.”(2020)
- [9] Krishnan, Santhana, and S. Geetha. “Prediction of Heart Disease Using Machine Learning Algorithms.” (2019).
- [10]Barik, Shekharesh, et al. “Heart Disease Prediction Using Machine Learning Techniques.” (2020).

IoT enabled obstruction evasion Robots for enhancing the security of the Systems

¹Ashima Arya, ²Shivam Mudgil, ³ Dr.Sapna Juneja, ⁴ Dr. Abhinav Juneja

¹Assistant professor, PIET, Samalkha, ³Professor, IMS, Ghaziabad

⁴Professor, KIET, Ghaziabad

ashiarya@gmail.com

Abstract: The surveillance of homes and other places like homes, restaurants, clubs, cafes, schools etc have become a major task in this technological era. There are various reasons required for this surveillance. One being the increasing crimes like that of theft . The major reason being the unattended homes by parents, either alone or under the care of babysitter or nanny. We don't know how these house service keeps on working and what they do at our homes, while we are away, leaving our homes completely unattended. So, to keep a check on it, we came on an idea of Mobile Controlled Surveillance Robot. This paper is based on the concept of Internet of Things, creatively called as IoT. The aim of this paper is to establish the communication between the owner and is/her unattended belongings, also to keep an eye on the surroundings. Thus, this research work will act as a game changer, i.e. nothing would go un-noticed now.

Keywords—Raspberry Pi, DTMF, IoT, Arduino, IR Sensor

1 Introduction

The recent advancements in the field of technology and the communication area, includes IoT. The robot is considered as a game changer due to the features it provides like advanced connectivity of systems, devices which go beyond the machine-to-machine communications. This project makes use of IoT in a secure way. It uses SSH (secure Shell Networking) tunneling Protocol, which provides the end-to-end encryption. As, only the designated person can see the online live streaming of robot, through using the Internet Protocol (IP) address of the Raspberry Pi, the brain of the robot. This project is mainly to provide live surveillance and act as a communication between the two places. Moreover, as it can be controlled by a mobile phone wirelessly, its ease of access becomes top notch.

Most importantly, the wildlife enthusiasts, travelers etc. can use it as a perfect medium for observing closely the life of the wild animals like Lions, Tigers, and Jackals etc., which are very deadly to be dealt with. Thus, without causing any loss to the human life, we can also keep a vigilant survey on them. Also, in Science Laboratory, where certain reactions take place, which may be harmful for the humans to be present there, surveying them personally, that can be easily surveyed through this Surveillance robot.

2 Literature Review

Internet of things has been felicitating our lives in countless ways with the rapid evolution of smart sensors and computing devices. T. Saravanakumar et al. [1] illustrated in his paper about the usage of Raspberry Pi as the brain of the robot and using it for the surveillance purpose. Paper contained detailed information for controlling a robotic vehicle, which is guided by the internet. Here, HD (High Definition) surveillance camera was used to monitor the video containing the detected motion which gets stored in the server and the live video can be accessed from anywhere just by entering static IP assigned to the system in the web browser of the same. It used HTML (Hyper Text Markup Language) as its source of streaming the video online.

Raj Hakani [2], in his paper explained the working of a robot cart using DTMF (Dual Tone Multi Frequency) tones. The DTMF tones are generated by pressing the different keys of telephone. There are 12 different tones that are mixture of High and Low frequency. Here, the movement of the robot is getting controlled by the mobile phone. One mobile phone is attached to the robot using 3.5mm AUX cable. The other phone is then made to call on that attached device, and then the keys are pressed from the mobile, and the robot moves

according to the tones it sensed. 2- Forward Movement, 8- Backward Movement, 4- Left Movement, 6- Right Movement, and 5- Stop.

Shahajada Mahmudul Hasan [3] in his paper detailed the working of a cart robot wirelessly. It has infinite working range as it is being controlled through the Internet. It has a teleoperation system which uses RF (Radio Frequency) technology for the task of processing the signals. An interesting component used here is the GPS (global Positioning System) tracker which is mounted to the robot for tracking the path of it and getting the satellite image of the trajectory motion of the robot and then, a track map is created. Here, 4 number of cameras have been attached to the body of the robot (one in each direction) for the purpose of accomplishing surveil of the surroundings. Various softwares have been used for this purpose. For the operation (movement) of the robot, the Visual Basic software has been used; for real-time image processing, vocal software has been taken use of; and for connecting every component together and to the real-time viewing on the monitor, VNC (Virtual Network Computing) software is used.

Amruta Nikam[4]elaborated the working of a 'Obstacle-Avoider Robot'. Movement of the robot is accomplished by using the DC motors and the L293D motor driver circuit to power the wheels. Here, the Pi-camera, which a camera of HD (High Definition) quality, is used for the same. This Pi-camera offers a great clarity of video streaming options like 1080p at 30fps and 720p at 60fps. The robot uses the obstacle detection algorithm, which it detects through the camera only and thus, gets saved from collision. Python is used as the core language for. Also in the Raspberry Pi, Open CV (Open source Computer Vision) has been used as the library. Open CV offers real time programming functions.

Kolapo Sulaimon Alli [5] discussed, the obstacle avoider Robot/ Unmanned Vehicle is made using the micro controller as Arduino UNO. The body of the robot is made using lightweight metal. It uses two ultrasonic infrared object sensors, which are put facing forward, thus, as soon as it detects any obstacle, it changes its path. The ultrasonic sensors transmits a sound at 37 kHz and then it waits to receive the same sound which has been sent. Then the distance pf the object is calculated from the robot. Afterwards, if the distance is less than 40 m, it will turn right 90 degrees and then transmits the echo again. This process repeats. If the distance is greater than 40m, it will continue to move forward till it becomes less than 40 m.

3 Hardware Design

Major components used in the design of circuit are listed below :

A. Raspberry Pi 3 Model B+

Raspberry Pi is latest model in Version 3. It is popularly known as a credit-card sized single board computer. It has 40 GPIO pins, 1 GB of LPDDR2 SDRAM (Static Dynamic Random Access Memory) and it supports micro SD card for the purpose of ROM. It is having Broadcom BCM2837BC0 SoC (System on Chip) and Cortex A53 (ARM v8) which also supports ARM on the Central Processing Unit and on Graphics Processing Unit. For wireless connection, it has 802.11ac Wireless LAN (2.4GHz and 5 GHz) support and Bluetooth 4.2. For Internet connectivity, it also supports Gigabit Ethernet over USB 2.0 (max 300 Mbps). Having 4 USB (Universal Serial Bus) 2.0 ports, makes connecting devices easier. It also has a HDMI port and a 5V/2.5A DC power input. Model 3B+ works on the 64-bit architecture and offers clock speed of unto 1.4GHz. The amazing fact, it being having a nice thermal (heat) management.

B. Arduino UNO R3

Arduino UNO R3 is a micro-controller having ATmega 328 IC which is designed and manufactured by Arduino. It can operate on tan external supply of the voltage ranging from 6-20V whereas the Input recommended voltage is 7-12V. It has 14 Digital IO pins, out of which 6 provides PWM output and 6 Analog Input Pins. Flash memory in Arduino is of 32Kb, of which 0.5Kb is used by the Bootloader. It has 2 Kb SRAM (Static RAM) and 1 Kb EEPROM and the operating clock speed is 16MHz. The DC current required per IO pin is 40 mA and for 3.3V pin is 50 mA.

C. USB Camera

The USB camera used here, can be attached directly to one of the USB ports of the Raspberry Pi model 3B+. It does not require any special configuration as the USB cameras available have mostly pre-installled softwares, which makes them ready-to-use by simple mechanism known as plug-and-play. Here, the QHMP

HD camera is used. It offers various features like 25MP Resolution (Software Interpolated), 6 LEDs (giving clear video even in dark), special visual effects and most importantly, Auto-focusing.

D. DC motors [6]

DC motors is a type of rotatory electrical motor, which helps in converting the DC electric current energy into the Mechanical energy. It has a shaft at the centre which rotates at the particular rpm. (rotations per minute) based on the voltage supply to it. Here, Brushed DC motors are used. Brushed Dc motors generates torque directly from the DC power supply by using rotating electro-magnets, stationary magnets and internal commutation. Here, we have used Single Shaft BO1 DC motor having 20 rpm (lesser the rpm of motor used, more is the video streaming quality) and wheels are attached to it through shaft. DC motors takes power from the 9V battery.

E. Motor Driver (L293D IC)

The micro-controller and the Arduino UNO board operated at a voltage of 5 Volts but the motors needs 9 Volts to run. Motors also need a higher current, to what is available from the arduino. That's why motor driver circuit is used. It acts as an interface in between the Micro controller and motor. L293D is a 16 pin IC. It uses H-bridge, thus, it can power two DC motors simultaneously.

F. DTMF Decoder (HT 9170 IC)

DTMF (Dual tone Multi Frequency) Decoder is a circuit, which is used to decode the DTMF tones produced by the telephone keypad while pressing keys. DTMF tone consists of two signal frequencies overlapping each other. It uses HT 9170 IC which is an 18 pin IC. It detects which key is pressed by knowing tone. It is connected to the device by using AUX cable as DTMF decoder module has a 3.5mm AUX input jack. Thus, the audio tone is received and it gets converted into the binary form. For eg. when key 4 is pressed, it detects it as it gets the information in binary code as 0100. For powering up the module, it also has a AC/DC connector, by which we can power it up using the DC connecting wire.

G. IR Object Sensor

The IR (InfraRed) object sensor module consists of an IR transmitter and receiver, an output LED, potentiometer, few resistors and IC LM 358 OPAMP. Consists of 3/4 pins usually. 3 pins are used at a time. The VCC pin, used for Input power supply to the module, the GND pin, used for Power supply ground and the OUT pin, which is active high Output.

The potentiometer is used to set the distance and the colour code, mostly black, to which the IR should detect and glow its Output LED. The maximum range of the IR Object Sensor is 20cm.

H. Sketch IDE [7]

As Arduino designs and manufactures the ICs and micro-controllers, so it makes software too. The coding for Arduino is done in Arduino Sketch IDE (Integrated Development Environment). It is a cross-platform application (i.e. compatible with MacOS, Windows and Linux). Its functions are mostly written in C and C++. This software is used to write an upload programs to the Arduino Compatible boards. It supplies a software library from the Wiring (open source electronics prototyping platform composed of programming language and an IDE project, started in 2003) which provides many common input and output procedures. It stores the files with .ino extension.

4 Design and Fabrication of the Robot

For the designing of the robot, the base of the robot we have chosen is wood. As wood being light in weight and it provides high rigidity. For front side, instead of wheels, a castor wheel is used. A castor wheel, being a light wheel, singular, all-direction wheel. In the rear side, two wheels are attached to the motors. Between both the bases, melamine cylindrical structures are used to provide the strength to the robot. Melamine is chosen as its light and super strong plastic, also cost efficient. On the top, the aluminium rod is used to place the camera as, the aluminium being lightest in weight. We always have to make our robot as light as possible, because various other components would be adding up some of the more weight to it.

Now, for accessing, we have used a software, VNC Viewer. It offers point-to-point transmission which is safe and secure. And by knowing the IP address of the Raspberry Pi, we can connect it to the host accordingly.

The block diagram exactly shows the working of the robot in the most exact manner. The Raspberry Pi 3B+ acts as the brain of the Robot. It is getting it's power from the Powerbank. As it supplies DC Voltage of 2A. For the surveillance purpose, USB Camera is used. The USB camera is connected to the Raspberry Pi using the USB cable through the USB port. The camera is also being powered up by the Powerbank. We have used the power bank here, as the power bank supplies for max 7-8 hours of power to the Raspberry Pi. For the connection, i.e. surveillance to the desktop, we have connected the raspberry to the system through the Wi-Fi, which is available in Raspberry. It comes with a pre-equipped Broadcom Wi-Fi chip. Thus, as the raspberry gets connected to the wifi, and for the desktop also, It must be connected to Wi-Fi. Now, for accessing, we have used a software, VNC Viewer. It offers point-to-point transmission which is safe and secure. And by knowing the IP address of the Raspberry Pi, we can connect it to the host accordingly.

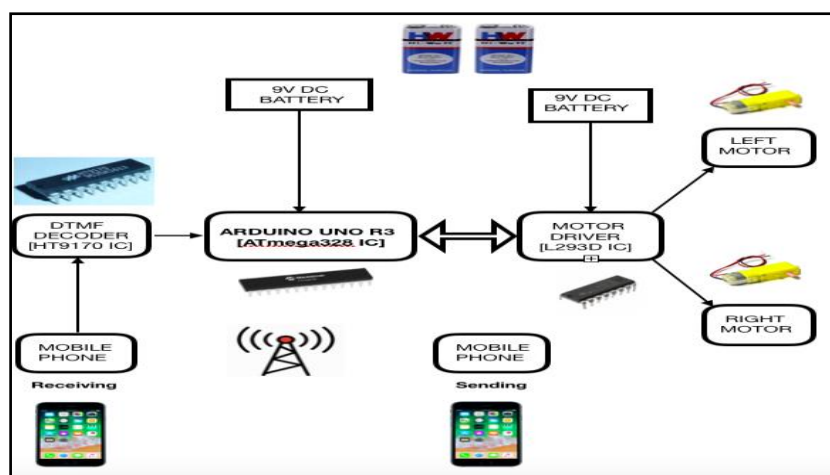


Fig 1: Block Diagram of Arduino UNO with various other components.

The arduino is used up as the mini brain here. It is used for managing and controlling the motors as well as the DTMF Decoder, which decodes the tones produced by pressing the keys of the mobile phone, which is used to control the motion of the Robot. The motors are being powered up by 9V DC power source. For this purpose, we have used a 500MAh Rechargeable battery, which is connected to the Motor Driver Circuit Board, which gets recharged from the Powerbank, if necessary, using a Micro USB cable. For the controlling purpose, two mobile phones are used. One is connected to the DTMF Decoder using a 3.5mm AUX Cable. And they both are connected through the cell phone tower, i.e. the Base Station of the Service provider. And that DTMF Decoder is connected to the Arduino. The motors are connected to the motor driver circuit board, which has the capability to power up two DC motors at a single time. The arduino also need power to run. For this purpose, it is connected to a single 9V Battery using a DC snap connector.

5 Result

a. Outputs from the Raspberry Pi - showing the surveillance

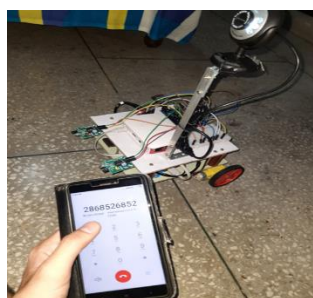


Fig 2: Snapshot of the Robotic System

The pressing of keys and movement of the robot accordingly. Here, it is moving forward, as the last command given it is to move forward. Also, the LED's on the DTMF Decoder module changes as we press the different keys.

6 Conclusion

In the recent times of where the Information Technology is at the peak, also the Artificial Intelligence is taking up a step forward, this obstacle avoider - mobile controlled Surveillance Robot is a stepping stone toward this field. This robot is based on the Live Surveillance using the microprocessor Raspberry Pi. Also, it uses microcontroller Arduino UNO for the purpose of its movement, which is commanded by the DTMF tones provided through a mobile phone, also it is used for its obstacle detection as well as avoidance mechanism. Using this Robot, the individual could easily keep an eye on his unattended place or things. Also, its various features like mobile controlled makes it easily accessible through any corner of the world and the Obstacle-Avoidance feature makes it collision free, thus improving its rigidity.

7 Future Scope

We further need to add some more features to make it work for a bit longer time. Now, it's time of working continuous ranges from 20 hours till a day. Further battery efficient changes needs to be done so that at least, it works for about a week on a single charge. Also, its wheels would be made as a caterpillar ones, so that allowing it to work even on sand, thus, making it all-terrain Robot.

References:

- [1] Mr. T. Saravanakumar , D. Keerthana , D. Santhiya , J. Sneha , D. Sowmiya , “Surveillance Robot Using Raspberry Pi-IoT”, International Journal of Electronics, Electrical and Computational System IJEECS ISSN 2348-117X Volume 7, Issue 3 March 2018.
- [2] Raj Hakani, “DTMF based Controlled Robot Vehicle”, IJSRD - International Journal for Scientific Research & Development | Vol. 2, Issue 12, 2015 | ISSN (online): 2321-0613
- [3] Shahajada Mahmudul Hasan, Syed Mamun R Rasid, Avijit Mallik, Md. Rokunuzzaman, “Development of a Wireless Surveillance Robot for Controlling from Long Distance”, International Journal of Engineering Research and Management (IJERM) | ISSN: 2349- 2058, Volume - 05, Issue - 09, September 2018
- [4] Amruta Nikam, Akshata Doddamani, Divya Deshpande, Shrinivas Manjramkar, “RASPBERRY Pi BASED OBSTACLE AVOIDING ROBOT”, International Research Journal of Engineering and Technology (IRJET)| Volume: 04 Issue: 02 | Feb -2017 , e-ISSN: 2395 -0056, p-ISSN: 2395-0072
- [5] Kolapo Sulaimon Alli, Moses Oluwafemi Onibonoje, Akinola S. Oluwole, Michael Adegoke Ogunlade, Anthony C. Mmonyi, Oladimeji Ayamolowo and Samuel Olushola Dada, “DEVELOPMENT OF AN ARDUINO BASED OBSTACLE AVOIDANCE ROBOTIC SYSTEM FOR AN UNMANNED VEHICLE”, ARPN Journal of Engineering and Applied Sciences| VOL. 13, NO. 3, FEBRUARY 2018| ISSN 1819-6608
- [6] https://en.wikipedia.org/wiki/DC_motor
- [7] https://en.wikipedia.org/wiki/Arduino_IDE
- [8] <https://opencv.org/about/>

Readiness towards Cashless Payments: India

¹ Atul Gautam, ²Sanjeev Kumar , ³ Kunal Aggarwal

¹Research Scholar, Haryana School of Business, Guru Jambheshwar University of Science and Technology, Hisar, Haryana, India,

²Professor, Haryana School of Business, Guru Jambheshwar University of Science and Technology, Hisar, Haryana, India

³Student, Panipat Institute of Engineering & Technology, Panipat Haryana, India.
Atulgautam27@gmail.com

Abstract: India is changing to a digital economy from a currency presiding economy through the arising new digital technologies namely mobile wallet in addition towards the electronic money (Routray et al., 2019). The cashless payment system was started by the Indian government in the leading program of Digital India (Sagayarani, 2017) The program of Digital India is a flagship program by the Indian government which has an aim to empower India digitally. Cashless means digital transaction economy or less cash society, minimizing the use of physical cash for transactions. Digital payments are initiated to make the payments in place of cash (Mohd. & Pal, 2020). This study explores aspects that affects the readiness of consumers towards cashless payments. This study gives two inputs to practice and theory. Initially, the factors that inhibit and drive the readiness of digital payments with a restructured context of the TRI [Technology Readiness Index] (Parasuraman, 2000) with the purpose of matching with earlier international cashless payment readiness studies, along with developing an understanding of Indian context. The outcomes of the study should support the cashless service providers competently to devote the proper effort, money and time in the necessities, marketing and developments of cashless transaction services.

Keywords: Cashless payments; Readiness; Digital India.

1 Introduction

India is changing to a digital economy from a currency presiding economy through the arising new digital technologies namely mobile wallet in addition to the electronic money (Routray et al., 2019). The cashless payment system was started by the Indian government in the leading program of Digital India (Sagayarani, 2017) The program of Digital India is a flagship program by the Indian government which has an aim to empower India digitally. Cashless means digital transaction economy or less cash society, minimizing the use of physical cash for transactions. Digital payments are initiated to make the payments in place of cash (Mohd. & Pal, 2020)

A non-cash transaction states to a financial activity where services and goods are executed deprived of cash via cheque payment or digital transfer (Paul & Friday, 2012). Digital payment is stated by means of a payment conducted through digital transfer of details of direct credit, credit cards or some other digital means except payment through cheque and money [Agimo, 2004]. In line with (Cabanillas et al., 2014), any mobile payment is well-defined as any kind of business or individual action by means of the help of a digital device linking through to a mobile network, allowing the fruitful fulfilment of a money related transaction. (Mallat, 2007) describes digital payment transactions as the usage of a digital device or gadget to make a transaction, wherein funds or money are transferred or transacted from payer to the remote or near receiver through an intermediate, directly, or without the help of an intermediary. (Slade et al., 2014) defines cashless payment by means of the potential of a mobile phone operator to authorize, start, and finish a economical transaction, in which funds or money are moved through network of mobiles, or through wireless communications knowhows to the beneficiary with the help of a digital device.

This study explores aspects that affects the readiness of consumers towards cashless payments. This study gives two inputs to practice and theory. Initially, the factors that inhibit and drive the readiness of digital payments with a restructured context of the TRI [Technology Readiness Index] (Parasuraman, 2000) with the purpose of matching with earlier international cashless payment readiness studies, along with developing an understanding of Indian context. The outcomes of the study should extend support to the service providers

more efficiently to devote the proper effort, money and time in the provisions, marketing and developments of cashless transaction services.

Objectives Of The Study

- To explore factors influencing Readiness of people towards cashless Payments.
- To explore cashless Indian economy.

2 Cashless Indian Economy

An economy without cash is one where all transfers are made through cards or automated means. The distribution of the actual currency would become marginal and diminished tax evasion cases, as trade traces remain within financial institutions. It would also suppress black money generation and decrease real property prices due to black-money curbs, though most of the hidden black money is deposited in real property values that inflate real estate prices. It paves the way for universal bank service access, given that no physical infrastructure other than digital infrastructure is required. Payments can be tracked and processed quickly, and corruption drops naturally meaning that consumers do not have to pay much to collect what they are entitled to.

On the part of the Government, the reliance on cash is reduced at different stages. The Indian Payment Corporation of India was set up to provide centralized infrastructure for different retail payment structures in India, including card payments, under the aegis of the Indian banks' Association and directed by RBI. NPCI would increase the reliability of retail payment uniformity and standardization and improve and expand the scope of both current and new payment products to enhance consumer comfort.

According to the Population Totals of Census 2011(Provisional) conducted by the Indian government, the total population calculation of India was 1210.193 million in 2011 out of which the male literacy rate was found to be 82.14% and literacy rate related to females was found at 65.46% (censusindia.gov.in). Conferring to a report by (Statista.com), India had the second largest internet populace in world at some 749 million handlers in 2020. Of these, 744 million operators retrieved the internet using digital devices. According to a study in 2020 by (Statista.com), most of India's internet users were in rural areas and male. About 42 percent of rural internet users were female that year with 58 percent males. However, urban areas female count of internet users was more, it was 43 percent for females while there were 57 percent male users in urban areas.

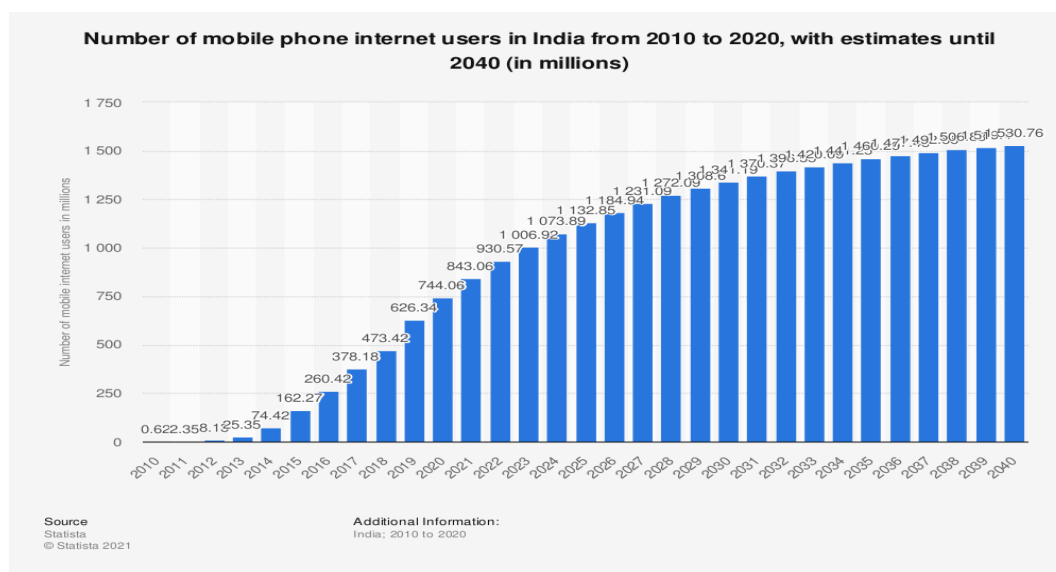
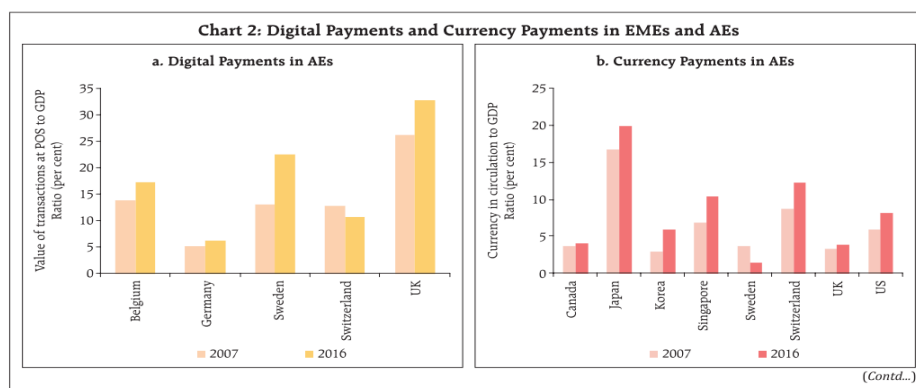


Fig 1: Chart Representing number of mobile phone internet users

▪ Digital Payments Trends: A Cross-Country Analysis

According to (world payments report, 2018), the proportions of electronic transactions grew worldwide by a CAGR (compound annual growth rate) of 9.8% whilst 2012-16 directed by evolving Asia and Central Europe,

Middle East and Africa. It is showed that advanced economies (AEs) account for a major share of cashless transactions in respect of volume, however, emerging market economies (EMEs) are also not lagging. They are also catching up quickly [Fig1]. Even though EMEs are captivating extensive leaps in cashless transactions, figures from discrete countries advocate some variances. First, it suggests that intensity and spread of use of electronic transactions does not always have a direct association with GDP level of the country. Some countries like Germany (AE) which has small digital payment to GDP ratio, the ratio is pretty indistinguishable in some other EMEs and AEs. (for example- ratio of an AE like Belgium is identical with ratio of an EME like Brazil.) As opposed to, a high level of GDP does not inevitably have an inverse relationship with low usage of currency (like- Japan which has one of the chief currencies to GDP ratio and one among lowest digital/cashless payments to world's ratio of whole world GDP.) Secondly, increase in digital payments to ratio of GDP completed in a time not looking to result in the reduction in the use of currency like- In Mexico, Saudi Arabia, Brazil and India the digital payments to ratio of GDP rose expressively throughout the period ending 2016 nonetheless the currency to ratio of GDP remained almost stable or grew (chart2). Hence, against these abnormalities, various behavioral and economic factors are required to be examined to decode the causes of growth in cashless transactions.



² Measured in terms of value of transactions at POS terminals to GDP ratio.

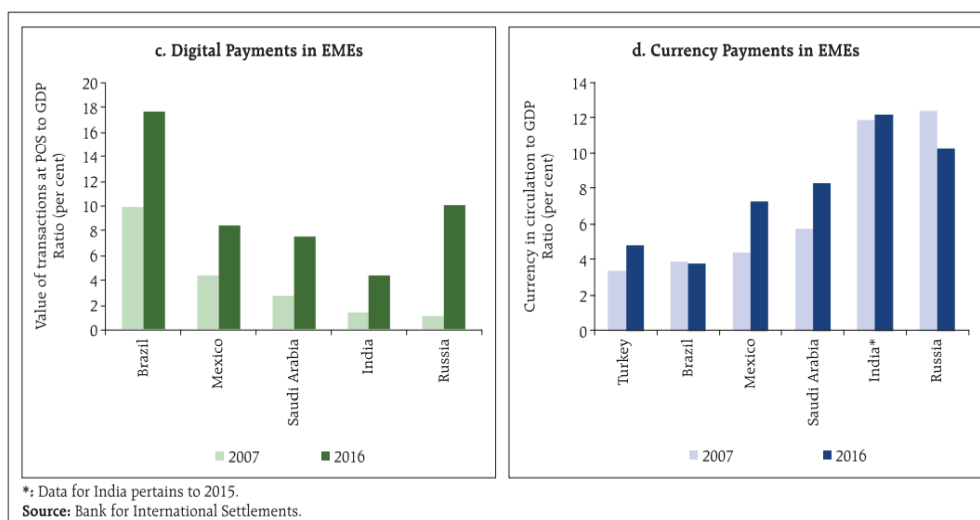
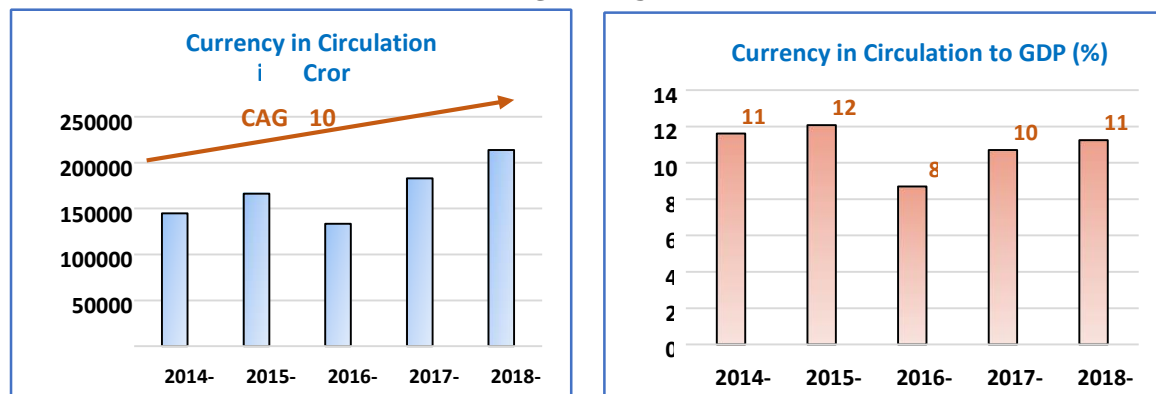


Fig 2:Chart representing digital and currency payment

2.1 Currency In Circulation

The Currency in Circulation (CIC) amount relates to the use of cash for payment determinations as the major demand for currency arises out of the payment needs i.e. for initiating payments. The CIC escalated at Compounded Annual Growth rate of 10.2 percent across the country around the last 5 years i.e. from 2014-19.

CHART-3



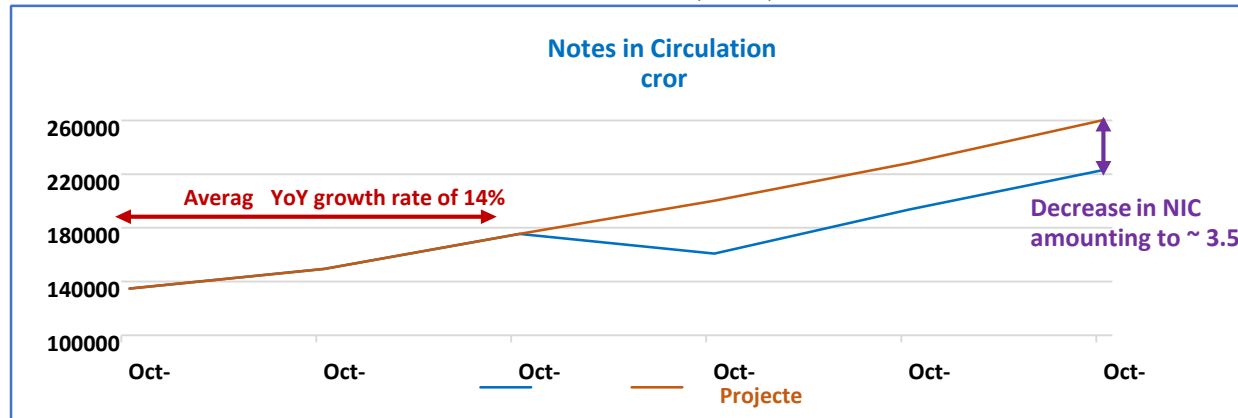
Source: RBI Data

Fig 3 : Represents currently circulation

Based on the assumption that high CIC relative to GDP shows currency is a majorly favored payment instrument, India has a robust influence for currency transactions. The sporty growth in GDP and demonetization reduced the CIC as a proportion of GDP to 8.70% in 2016-2017 and it improved to 10.70% and 11.2% in 2017-18 and 2018-19 though which was lesser against the pre-demonetization level of 12.1% in year 2015-16. The degree of growth is comparatively low showing a move far away from cash.

The (NIC) notes in flow improved to average rate of 14% from October 2014-16. Taking the same rate of growth in October 2019, NIC should have been Rs.26,04,953 crore but was only 22,31,090 crores showing that decrease in cash usage and digitization reduced NIC by some Rs.3.5 lakh crore.

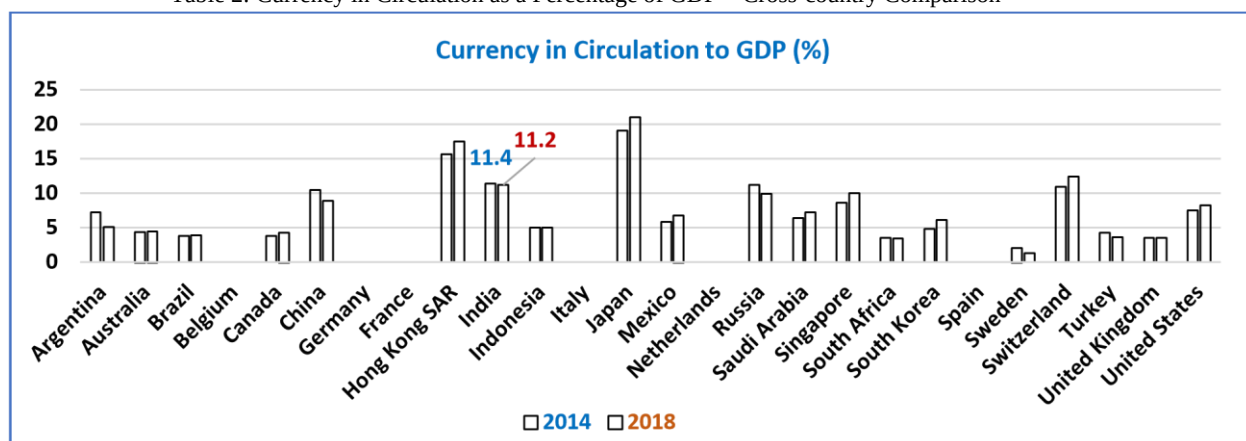
Table 1: Notes in Circulation (₹ crore)



Source: RBI Data

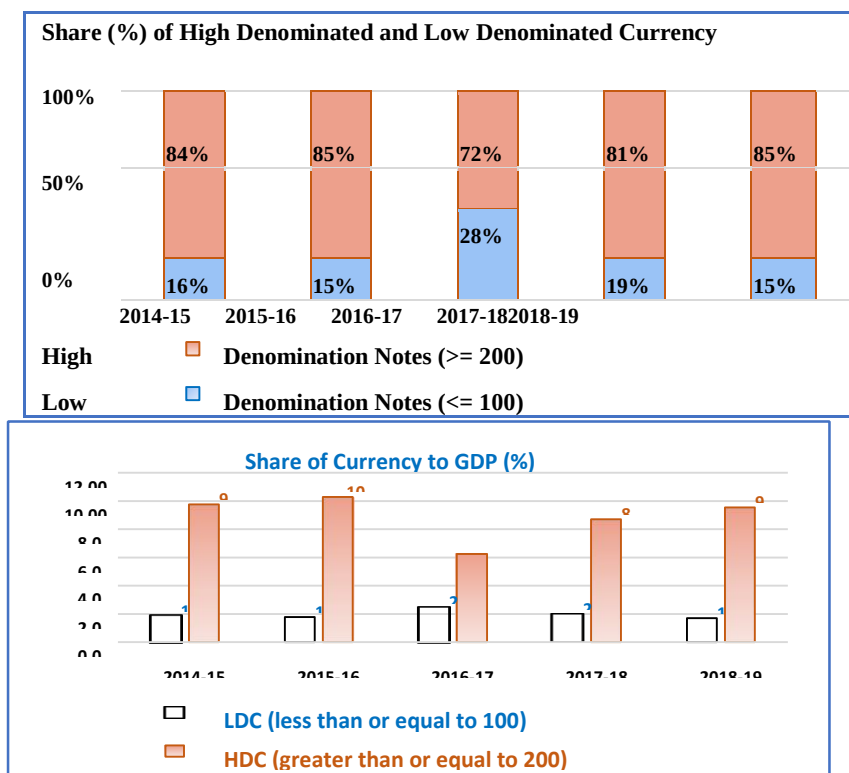
Though levels of CICs in India decreased in 2018 in comparison to 2014, other countries, except China, Turkey, Sweden, South Africa, Argentina, Indonesia and Russia had growing levels of cash. Though cash is extremely placed in Indian transaction systems, intentional efforts post-demonetization showed a noticeable move to cashless payments.

Table 2: Currency in Circulation as a Percentage of GDP – Cross-country Comparison



Source: BIS Red Book 'Country Tables' compiled by the Bank for International Settlements. CIC figures were not available for Euro area countries.

Table 3: High and Low Value Denominated Currency – India



Source: RBI Data

Cash is utilized together as and as a store of value and a means for payment. It is assumed that denomination notes of high value are majorly detained as a store of value and denomination notes of low value for transactions to try and unravel the dual types of cash demands (Amromin & Chakravorti, 2007). From the last 5 years, the demand for denominated currency of more or high value has overtook the currency of less value denomination which indicates that cash is progressively used less for making payments and more for store of value.

3. Literature Review

Cashless transactions are based on technology and in the frame work of banking technology also readiness acts as an important consumers' level technology readiness remains an important gauge of technology usage.

The outcomes about digital environment's optimism, innovation, discomfort and consumer insecurity within the digital atmosphere offered by this study shows that attitude, experience, innovation and involvement affect's consumers' purchasing decisions (Monteiro et al., 2019). The attitude of user for technology usage is evaluated based on four variables: optimism, innovativeness, discomfort and insecurity constructed on (Parasuraman & Colby, 2001).

Optimism

Optimism is a very important and key factor directing the micro-merchants' flowing intention to use cashless payment modes having a positive effect while discomfort is the main inhibitor (Priyanka et al., 2020). However, it showed that innovativeness and optimism are not important for affecting consumers towards the adoption of cashless payment services, whereas discomfort is not a main inhibitor (Humbani & Wiese, 2017). Optimism and innovativeness deliver favorable goals and intentions, but discomfort and insecurity convey unfavorable intentions that can prevent the usage of QR code-based mobile transactions (Humbani, 2019). Enthusiasm in consumers pertaining to new technologies like mobile cashless payment has a direct and positive impact on the usage of mobile payments (Guhra et al., 2013). Using digital payments rather than cash would provide a comprehensive record pertaining to all the transactions happening in the society which permits added transparency in money transfers and business operations which saves from money laundering and tax invasion and avoidance by people (Garg & Panchal, 2016). Using cashless payments methods increases transparency in the economy which would further lead to rise in GDP of the country which would lead to increasing country's credibility and the investment level (Agarwal & Khatri, 2021). Paying through new payment methods like mobile phones decreases the pain of paying which results in a better consumer experience (Pisani & Atalay, 2018). Cashless transactions improve scalability, accountability and enhances transparency as the written record helps in keeping an eye over spending resulting in better budgeting (Ramya et al., 2017). A digital transaction is a safe payment for customers, which increases revenue and enhances operational efficiency for sellers (Kumari & Khanna, 2017). Electronic card payments and boost consumption of the economy and increases efficiency (Tee & Ong, 2016). Using, Digital payment methods, provides more economic freedom to people i.e. allowing people to trade with each other in different countries. Digital currencies deliver more economic freedom for business transactions of their users or consumers present anywhere in the world (Singhaur & Garwal, 2018).

Innovativeness

Cashless transactions are promoted by innovation, awareness about environment and entrepreneur mindset while conservatism, orthodox mindset and tradition may pose a hindrance (Glennow & Granstrom, 2019). Attitude, innovativeness, cost and mobility are substantial antecedents of the intention of consumer for adopting digital (m-payment) systems (Patil et al., 2019). Innovativeness also influences the use of NFC Mobile Credit Card (MCC) (Anthony & Wong, 2018). Personal innovativeness along with perceived usefulness as measured by personal innovativeness in information technology serve as the primary determinants for user's continuous intention for the usage of mobile payments. Perceived ease of using mobile payments is largely affected by user's personal innovativeness influencing user's perceptions (June lu, 2014). To facilitate the acceptance of digital payment systems, difference in adopters' personal innovativeness should also be taken care of. (Yang et al., 2011). Personal innovativeness effects the ease of use for mobile payment services to a major extent. Mobile payments are more likely to be used by men as compared to women as they are more easily effected by their social environment as compared to potential risks involved while personal innovativeness affects women (Kalinic et al., 2019).

Discomfort

Financial, performance and psychological risks are the most important apprehensions that consumers feel while conducting online shopping (Marriott & Williams, 2018). Customers often find themselves with a

feeling of discomfort, skepticism and apprehension while searching the internet or products or services when the actual monetary exchange is to take place. This happens due to lack of trust when it comes to sharing personal and financial information (Sahney et al., 2013). B-school teachers and students try to avoid online mobile shopping and still prefer for offline shopping by reason of a sense of occasional discomfort they feel with online shopping (Shaikh & Daddikar, 2017). Respondents show a feeling of discomfort corresponding to online risks (Acuesta & Tamayo, 2018). lack of trust and security of the technology were found to be major reasons for non-adoption of online banking (Abid & Noreen, 2015). The findings lend backing to the fact that shopping through internet is viewed as dangerous proportion of risk despite its significant benefits (Wani & Malik, 2013). Another drawback of digital payment system is that only technology savvy people who are aware as how to use technology can only use this method, causing discomfort for unaware people (Garg & Devi, 2019).

Insecurity

The major challenges that electronic payments face are fraud prevention and security issues which are expected to rise with the widespread of e-commerce technologies (Benjamin et al., 2006). On the expression of a survey report, nearby 62% people are anxious regarding the security of their systems. According to another survey in 2015, only 16% customers preferred online payment while 67% customers prefer to pay using hard cash among the existing systems of customer electronic payment. Furthermore, overcoming security and privacy issues, concerns and using the newest technology may well encourage quick adoption of online-payment methods (Khan et al., 2017). Cyber security issues pose an escalating challenge in retail sector in electronic payments, with which cashless transactions happening after demonetization, the risk of virus attacks, cyber-crimes, malware, leakage of information and confidential Data and online fraud has been rising (Gade, 2020). It is discovered that online payment gateways show some security issues in spite of their large-scale use and growing popularity (Masihuddin et al., 2017). The increasing digitization of online payment technologies of money unlocks a new set of apprehensions, with increasing risk of hacking and largescale frauds. Electronic money can result in the formation of huge volumes of data potentially leading to breach of individual privacy (Thomas et al., 2016).

Theoretical Framework

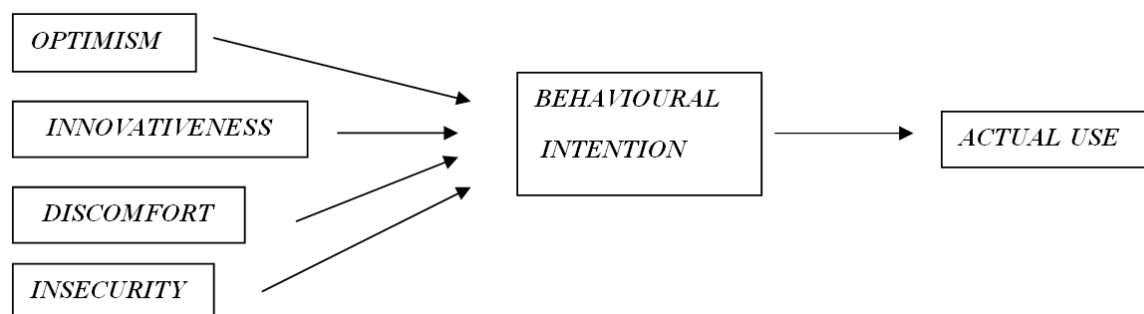


Fig 3: Theoretical Framework

4 Conclusion

Indian economy is rapidly moving towards a more cashless economy. Infrastructure provided by the Indian Government is eyed upon achieving the dreamy state of cashless economy. Intention to the use of cashless payments is mostly affected by Perceived Usefulness, and Optimism had a great influence on Perceived Usefulness indirectly affecting the intention to usage of mobile payments (Martens et al., 2017). Innovativeness's influence on the usage intention is different amid non-users and users representing a very important share of innovativeness in the cashless payment modes intention to use (Thakur & Srivastava, 2014). Optimism and innovativeness have a good positive influence on the aim and intention for the services of online food delivery ordering (OFDO), while discomfort and insecurity negatively influence OFDO service

intentions. Also, the actual performance behavior concerning. Optimism and innovativeness have sturdier effects in demographic variables, relating to males, a good or high income, young, a good or high education, etc. on the contrary, discomfort and insecurity have more impacts related to older or aged, illiterates and low education, female, low income, etc. (Ali et al., 2020). Consistent with previous literature, it is found that there is a positive impact of perceived risk on the intention for using digital wallet services in India (Singh et al., 2020). Thus, factors like optimism, innovativeness, discomfort and insecurity influences the behavioral intention to use and marketers may take advantage from this study for the marketing of cashless payment services.

References

- [1] Abid, H., & Noreen, U. (2006). Ready to E-bank: An exploratory research on adoption of e-banking and e-readiness in customers among commercial banks in Pakistan. *Spider*, 31(2), 1-31.
- [2] Agarwal, M., & Khatri, M. (2021). Moving Towards Cashless Society–Literature Review. *International Journal of Modern Agriculture*, 10(2), 4419-4428.
- [3] Agimo, T. (2004). Better Practice Checklist for ePayment. *Australia government information management office*. Retrieved on, 26(5), 2015.
- [4] Akhalumeh, P. B., and F. Ohiokha. "Nigeria's cashless economy: The imperatives." *International Journal of Management and Business Studies* 2.2 (2012): 31-36.
- [5] Ali, S., Khalid, N., Javed, H. M. U., & Islam, D. M. (2021). Consumer adoption of online food delivery ordering (OFDO) services in Pakistan: the impact of the COVID-19 pandemic situation. *Journal of Open Innovation: Technology, Market, and Complexity*, 7(1), 10.
- [6] Amromin, G., & Chakravorti, S. (2007). Debit card and cash usage: a cross-country analysis.
- [7] Capgemini, B. N. P. Paribas (2018). World Payments Report, 2018, available at <https://worldpaymentsreport.com/wp-content/uploads/sites/5/2018/10/World-Payments-Report-2018.pdf>.
- [8] Colby, C. L., & Parasuraman, A. (2001). *Techno-ready marketing: how and why customers adopt technology*. Simon and Schuster.
- [9] Cuesta, R. M., & Tamayo, A. (2018). Risk Assessment of Online Transactions: An Exploratory Study. *Available at SSRN 3159932*.
- [10] Gade, A. S. ROLE OF CASHLESS TRANSACTION AND ONLINE TRADE DURING THE COVID-19 LOCKDOWN PERIOD.
- [11] Garg, K., & Devi, S. DIGITAL PAYMENT SYSTEMS IN INDIA. *Business Management and Social Innovations*, 349.
- [12] Garg, P., & Panchal, M. (2017). Study on Introduction of Cashless Economy in India 2016: Benefits & Challenge's. *IOSR Journal of business and management*, 19(4), 116-120.
- [13] Glennow, E., & Granström, A. (2019). Cashless society: Is there a relationship between innovation and cash circulation in economy?
- [14] Guhr, N., Loi, T., Wiegard, R., & Breitner, M. H. (2013, February). Technology Readiness in Customers' Perception and Acceptance of M (obile)-Payment: An Empirical Study in Finland, Germany, the USA and Japan. In *Wirtschaftsinformatik* (p. 8)
- [15] <http://cashlessindia.gov.in/>
- [16] https://censusindia.gov.in/2011-prov-results/data_files/india/Final_PPT_2011_progresstables.pdf
- [17] https://m.rbi.org.in/scripts/BS_ViewBulletin.aspx?Id=18409#S2
- [18] <https://www.statista.com/statistics/558610/number-of-mobile-internet-user-in-india/>
- [19] <https://www.statista.com/statistics/751167/india-urban-and-rural-internet-users-by-gender/>
- [20] Humbani, M. (2019, September). GENERATION Y CONSUMER READINESS TO ADOPT QUICK RESPONSE (QR) CODE-BASED MOBILE PAYMENTS. In *31st ANNUAL CONFERENCE OF THE SOUTHERN AFRICAN INSTITUTE FOR MANAGEMENT SCIENTISTS (SAIMS)* (p. 519).
- [21] Humbani, M., & Wiese, M. (2018). A cashless society for all: Determining consumers' readiness to adopt mobile payment services. *Journal of African Business*, 19(3), 409-429.
- [22] Kalinić, Z., Liébana-Cabanillas, F. J., Muñoz-Leiva, F., & Marinković, V. (2019). The moderating impact of gender on the acceptance of peer-to-peer mobile payment systems. *International Journal of Bank Marketing*.
- [23] Khan, B. U. I., Olanrewaju, R. F., Baba, A. M., Langoo, A. A., & Assad, S. (2017). A compendious study of online payment systems: Past developments, present impact, and future considerations. *International journal of advanced computer science and applications*, 8(5), 256-271.
- [24] Kumari, N., & Khanna, J. (2017). Cashless payment: A behavioural change to economic growth. *Qualitative and Quantitative Research Review*, 2(2).
- [25] Liébana-Cabanillas, F., Sánchez-Fernández, J., & Muñoz-Leiva, F. (2014). Antecedents of the adoption of the new mobile payment systems: The moderating effect of age. *Computers in Human Behavior*, 35, 464-478.
- [26] Lowry, P. B., Wells, T. M., Moody, G. D., Humphreys, S., & Kettles, D. (2006). Online payment gateways used to facilitate e-commerce transactions and improve risk management. *Communications of the Association for Information Systems (CAIS)*, 17(6), 1-48.
- [27] Lu, J. (2014). Are personal innovativeness and social influence critical to continue with mobile commerce?. *Internet Research*.

- [28] Mallat, N. (2007). Exploring consumer adoption of mobile payments—A qualitative study. *The Journal of Strategic Information Systems*, 16(4), 413-432.
- [29] Marriott, H. R., & Williams, M. D. (2018). Exploring consumers perceived risk and trust for mobile shopping: A theoretical framework and empirical study. *Journal of retailing and consumer services*, 42, 133-146.
- [30] Martens, M., Roll, O., & Elliott, R. (2017). Testing the technology readiness and acceptance model for mobile payments across Germany and South Africa. *International Journal of Innovation and Technology Management*, 14(06), 1750033.
- [31] Masihuddin, M., Khan, B. U. I., Mattoo, M. M. U. I., & Olanrewaju, R. F. (2017). A survey on e-payment systems: elements, adoption, architecture, challenges and security concepts. *Indian Journal of Science and Technology*, 10(20), 1-19.
- [32] Monteiro, T. A., Giuliani, A. C., Pizzinatto, N. K., & Larios-Gomez, E. (2019). Managing the digital consumer: insights from Brazil and Spain. *Journal of Manufacturing Technology Management*.
- [33] Parasuraman, A. (2000). Technology Readiness Index (TRI) a multiple-item scale to measure readiness to embrace new technologies. *Journal of service research*, 2(4), 307-320.
- [34] Patil, P. P., Rana, N. P., & Dwivedi, Y. K. (2018, June). Digital payments adoption research: a meta-analysis for generalising the effects of attitude, cost, innovativeness, mobility and price value on behavioural intention. In *International Working Conference on Transfer and Diffusion of IT* (pp. 194-206). Springer, Cham.
- [35] Pisani, F., & Atalay, S. (2018). Cashless Payments, Pain of Paying and the Role of Attachment. *ACR European Advances*.
- [36] Priyandana, I., Stevani, M., Sutanto, T. I., & Mariani, M. (2020). Grassroots Economy Towards Cashless Society: An Empirical Analysis Of Micro-Merchant's Readiness In Continuing The Usage Of Cashless Payment System. *International Journal of Scientific & Technology Research*, 9(3), 929-928.
- [37] Ramya, N., Sivasakthi, D., & Nandhini, M. (2017). Cashless transaction: Modes, advantages and disadvantages. *International Journal of Applied Research*, 3(1), 122-125.
- [38] Routray, S., Khurana, R., Payal, R., & Gupta, R. (2019). A move towards cashless economy: A case of continuous usage of mobile wallets in India. *Theoretical Economics Letters*, 9(04), 1152.
- [39] Sagayarani, D. (2017). Digital Payments In India IOSR Journal of Business and Management (IOSR-JBM) e-ISSN: 2278-487X, p-ISSN: 2319-7668 PP 28-33 www.iosrjournals.org Name of Conference: International Conference on—Paradigm Shift in Taxation.
- [40] Sahney, S., Ghosh, K., & Shrivastava, A. (2013). Conceptualizing consumer “trust” in online buying behaviour: An empirical inquiry and model development in Indian context. *Journal of Asia Business Studies*.
- [41] Shaikh, A. H., & Daddikar, P. V. (2017). Online Shopping Habits Among B-School Students and Teachers: A Study of Belagavi City. *International Journal of Marketing & Business Communication*, 6(3).
- [42] Sharif, M., & Pal, R. (2020). Moving from cash to cashless: A study of consumer perception towards digital transactions. *PRAGATI: Journal of Indian Economy*, 7(1), 1-13.
- [43] Singh, N., Sinha, N., & Liébana-Cabanillas, F. J. (2020). Determining factors in the adoption and recommendation of mobile wallet services in India: Analysis of the effect of innovativeness, stress to use and social influence. *International Journal of Information Management*, 50, 191-205.
- [44] Singhraul, B. P., & Garwal, Y. S. (2018). Cashless Economy—Challenges and Opportunities in India. *Pacific Business Review International*, 10(9), 54-63.
- [45] Slade, E. L., Williams, M. D., & Dwivedi, Y. K. (2014). Devising a research model to examine adoption of mobile payments: An extension of UTAUT2. *The Marketing Review*, 14(3), 310-335.
- [46] Tee, H. H., & Ong, H. B. (2016). Cashless payment and economic growth. *Financial innovation*, 2(1), 1-9.
- [47] Thakur, R., & Srivastava, M. (2014). Adoption readiness, personal innovativeness, perceived risk and usage intention across customer groups for mobile payment services in India. *Internet Research*.
- [48] Thomas, L. D., Vernet, A., & Gann, D. M. (2016). Adoption readiness in service innovation: the case of digital money. *Industry and Innovation*, 23(4), 353-381.
- [49] Wani, S. N., & Malik, S. (2013). A Comparative Study of Online Shopping Behaviour: Effects of Perceived Risks and Benefits. *International Journal of Marketing & Business Communication*, 2(4).
- [50] Wong, A. T. T. (2018). A study of consumer acceptance of mobile payment services in Hong Kong. *Journal of Economics, Management and Trade*, 1-14.
- [51] Yang, S., Lu, Y., Gupta, S., Cao, Y., & Zhang, R. (2012). Mobile payment services adoption across time: An empirical study of the effects of behavioral beliefs, social influences, and personal traits. *Computers in Human Behavior*, 28(1), 129-142.

Emotion Based Music Recommender System In Cloud

¹Pavan Kumar, ²Pradeep Arya, ³Atul Mishra, ⁴Sunil Gupta

*BML Munjal University, #University of Petroleum and Energy studies
bellamkonda.kumar.19cse@bmu.edu.in, Pradeep.arya@bmu.edu.in, atul.mishra@bmu.edu.in,
s.gupta@ddn.upes.ac.in

ABSTRACT. While listening to the music it is will be difficult for a person to select the songs that suits their present emotion/mood. Most of us spend hours of time for manually selecting songs creating a playlist that we prefer to listen. The main motto of this Emotion Based Music Recommender System is it suggests songs to the end-user which suits to their emotion/mood. Suggesting music based on the person's emotion and their music preference will be the best approach to develop person's listening experience. This music system very efficient which recommends a playlist of songs based on the present emotion/mood of person using it. The person doesn't have to search the songs and create a custom-playlist that suits their present emotion/mood. The image capture of the end-user is done by using webcam of their system. Then, emotion/mood of the end-user is predicted from the snapshot of the person's picture and according to the selected playlist, the songs are classified, and a new playlist will be created according to the songs that suits to the end-user's current emotion/mood.

1 Introduction

People show their emotion/mood through their facial expressions. Listening song is very important in every person's life. It takes a significant part in improving an individual's life. Most of the people will be in a situation where they are unable to discover melodies that suits their present state of mind. Music listening altogether affects human sentiments, considerations and subsequently, it impacts mental and actual wellbeing. Audio players have gotten varied capabilities including fast forward, variable playback speed, local playback, and streaming playback with a multicast feed because of advances in technology and software. However, the quest for tunes that fit the listener's current mood/emotion continues.

Music recommendations can be applied in various fields such as relaxing, music treatment, stress and tiredness destruction, physical work, studying, sports, and many others. Recommendation Systems will find relationships in the dataset by understanding user's choices then produces results which are associated to the user's requirements. Most of the recommender systems don't think about end-user's current emotions or feelings.

To better understand their customers' musical preferences, several online music stores and streaming services (such as Spotify, iTunes, Grooveshark, and others) employ customised music recommendation algorithms. The approach to personalised music recommendations can take one of three forms: content-based filtering analyses the content of previously enjoyed music to offer like related music to the listener. Collaboration filtering recommends music to a user based on what other people with similar likes have been listening to recently, according to the user. The other strategy is a hybrid approach that makes use of both content and user-generated data in order to deliver music suggestions to consumers.

Here, Emotion Based Music Recommender System would be able to understand end-user's preference and their emotion/mood and select suitable list of songs according to the user's emotional context. In the event where users wish to locate music that are appropriate for their present feelings, this technique is quite beneficial. There are two requirements this system needs to solve which are to detect the end-user's mood and to provide a list of songs to influence the user's emotion.

The goal of this paper is to train the system to recognize the user's emotion, and to propose an idea of evaluating a song using the person's emotion/mood to enhance the person's listening experience by recommending appropriate songs that suits the current state of an individual. This can be performed by capturing the image in real-time by using a webcam. The first step is to detect face of the person and extract

the features. The next step is the classification of emotions. Now according to the classified emotion, the appropriate song from the playlist is recommended.

2 Literature Review

The emotion/mood of a person can be predicted by taking the snapshot of person's image through webcam. The person's image is then preprocessed using dimensional reduction algorithms to get the primary data from the captured image. The face detection is done using Fisher Face and Haarcascade methods. [1]

Fisher Face Algorithm is utilized for dimensional reduction of the facial dimensions utilizing the Principal Component Analysis(PCA) technique and afterward it applies Fisher's Linear Discriminant(FDL), or the Linear Discriminant Analysis(LDA) technique classify different features of the image, we use this technique because it amplifies the division between classes in the training process. The Fisher Face Algorithm can help us in image recognition and the face detection which helps to classify the user's emotion/mood.

Haarcascade algorithm is mainly used for object recognition. It is an algorithm that can detect different objects in an image. To detect different objects from the person's image like the nose, eyes, ears, lips in the face, we can use haarcascade_frontalface_default.xml. In OpenCV a haarcascade classifier is designed which is used for frontal face detection from an image and can also be used to classify different objects from the image. In order to train cascading classifiers, several hundred "positive" sample views of a certain item are used in conjunction with random "negative" photos of the same size. Positive samples include only the items that are required for our classifier to correctly identify them as positive. Negative samples contain the other objects, which are needed for the classifier to distinguish.

Fang-Fei Kuo et al and Suh-Yin Lee et al. [2] presented that creation of music suggestions is beneficial for users because of the increase of digital music. Users' liking for music is taken into account in current recommendation methods. However, there are occasions when it's necessary to suggest music based on the user's emotional state. A new model for emotional music suggestion is presented in this research and is based on the detection of associations in cinema music. We looked at how to extract musical aspects and tweaked the affinity graph to help us identify emotional connections between musical elements. According to the results of the experiments, the proposed technique has an average accuracy of 85%.

Bruce Ferwerda et al and Markus Schedl et al. [3] claimed that the basic study assumptions may be improved by taking personality and emotional states into account while making music choices. Including these psychological considerations, they reasoned, would improve the recommendation's accuracy. Attentiveness is paid to how people manage their emotional states through musical stimulation and how that stimulation relates to their unique personalities.

Ivana Andjelkovic et al and John O'Donovan et al. [4] asserted that a significant amount of work in recommender systems is devoted to improving prediction and ranking accuracy. In the meantime, new research has revealed the significance of other important aspects like control and user experience in general. MoodPlay, a hybrid music recommender system that combines content and mood-based filtering in an interactive interface, builds on these characteristics of music recommendation. Here, we demonstrate how MoodPlay's latent emotive dimensions let users explore their music library, and we describe how to combine user input at recommendation time with predictions based on a user profile that already exists.

James, et.al. [5] They proposed utilising a camera to take input from the user's face. This input may be used for a variety of things, including determining a person's mood by extracting information from their body language. As a result of this information, you can obtain a list of songs that fit with the mood you set with your first input. Using Facial Expression Based Music Player, a playlist may be generated by scanning and analysing facial expression data. In order to develop emotion-based music players, our proposed system focuses on recognising human emotions. This includes identifying the ways now utilised by music players to detect emotions, as well as the methodology we want to apply in our music player to detect human emotions.

Shalini et.al. [6] proposes using a user's emotions to provide music recommendations. Human faces are used as input, with facial expression identified and music produced automatically as a result of those feelings. We employed a point detection technique for feature extraction, and OpenCV for machine learning training to

get encouraging results. As a result, real-life face photos may be recognised with high accuracy using the suggested technique.

Several Research Papers & Patents giving a brief about the idea are [7] :

Title	Year	Description
“Emotion Based Music Recommendation System” [8]	2019	The research paper describes various features and facial parts that can be used to detect emotion and proposes a method for generating music based on detected emotion.
“Emotion Based Music Player” [9]	2017	The project focuses on an android application that would capture an image of the user and detect 4 emotions and develops a simple algorithm for generating a playlist and detecting emotion further also lets user add a song and skip a song.
“Emotion Based Music Player” [10]	2015	This module uses the intensity of speech to detect emotion and further uses artificial neural, confusion matrix for generation of the playlist.
“Smart Music Player Integrating facial Emotion Recognition and Music Mood Recommendation” [11]	2017	This study basically describes the process through 3 modules, emotion module for capturing the user’s details, Music classification module that uses deep learning for recognizing emotion and further Recommendation module for suggesting the playlist.
“Method and Apparatus for Recognizing an Emotion Of An Individual On Facial Action Units” [12]	2014	The invention basically describes the facial action units that can be used and algorithm for combining 2 facial action units for accurate detection of the emotion.
“Generating Music Playlist Based on Facial Expression” [13]	2012	This patent describes how to capture the area of interest for emotion detection and database for storing playlist assigned to particular mood.
“Music Playing System and Music Playing Method Based on Speech Emotion Recognition” [14]	2017	This invention describes how using speech and emotion coordinates using different algorithm for playing the music playlist according to the emotion detected.
“Real Time Emotion Recognition from Audio Signal” [15]	2018	This invention focuses on detecting audio signal and rapid audio fingerprint on computing devices. Features are detected and extracted to determine degree of similarity.
“Apparatus and Methods for The Detection Of Emotions In Audio Interaction” [16]	2008	This invention provides method and apparatus for detecting emotional states of a speaker in an audio recording. There is no requirement of predefined voice sample of the speaker.

3 Proposed System & System Architecture

a) Proposed System

Here I propose an Emotion Based Music Recommender System. Its main objective is to provide the user a playlist of songs that suits the present facial emotion/mood.

Firstly, this system will capture the live image of the user using webcam and predicts the emotion of the user(happy/sad) using the Convolutional Neural Network(CNN) Model trained using the Fer2013 [22] dataset. Then, the user can select the playlist he/she want to listen, the system will extract all the features of the songs present in that playlist and classifies them to happy or sad using the Random Forest Classifier Model trained using the Spotify Music data to identify the moods [23] dataset. Finally, the System will create and display a playlist of songs according to the user’s emotion.

b) System Architecture

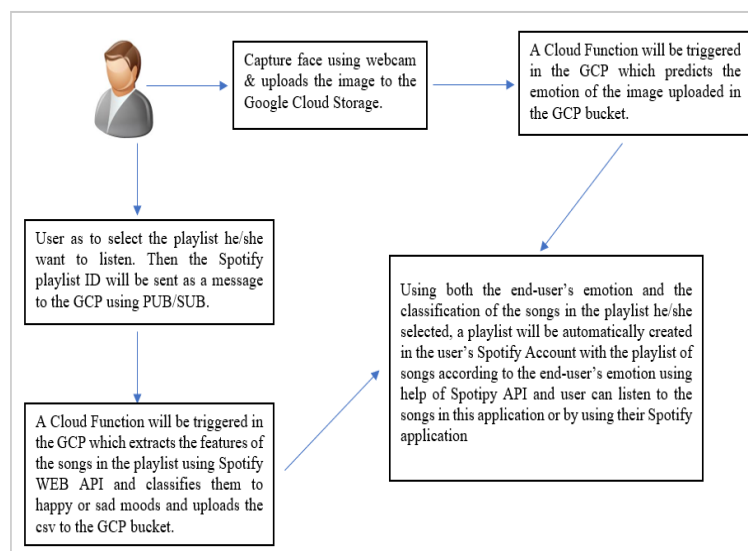


Fig. 19. System architecture of emotion-based music recommender

To use this application, the user must first sign in using their Spotify account, after which the user must record his or her live image with their camera using the application. Meanwhile, the taken image will be posted to the Google Cloud Bucket [20], and then the emotion of the uploaded image will be predicted using the Google Cloud Function [19] using the Google Cloud Function [20]. Following that, the user will be prompted to choose between a Spotify playlist or the list of recently played songs. In the background, the selected playlist Id will be published to Google Cloud Pub/Sub [19], and then the songs and their audio features that are present in the playlist will be extracted using the Spotify WEB API [18], and the songs will be classified using another Google Cloud Function [19] based on the audio features. Finally, using both the predicted emotion and the classified songs, a playlist will be created in the Spotify Library that contains a list of songs that are appropriate for the current user's emotion/mood. This will be done using the spotipy [17] python library and will be accessible via the Spotify website.

c) Implementations

We used Flask [21], Google Cloud Platform [19, 20] (Cloud Storage, Cloud Functions, Pub/Sub) to implement the above application. We used webcam.js [24] library to access the webcam and to capture the live image in the application. We used Spotify WEB API [18] and spotipy [17] python library to access the playlists and the Spotify user info to create a playlist of recommended songs by the system in the user's Spotify Library.

We used the Keras Sequential model to build a facial emotion classifier of 2 classes (happy, sad) and trained the using FER 2013 [22] dataset using 2 classes(happy, sad). It is used by the above application to detect face using haarcascade_frontalface_default.xml [1] file and predict the facial emotion of the captured image.

We used Random Forest classifier for songs classification and trained the classifier using Spotify Music data to identify the moods [23] dataset. It is used by the above application to classify songs from the playlist that user selected.

4 Methodologies And Libraries Used

Requests - In online application, Requests is an HTTP library that simplifies the process of dealing with http re-quest/response requests and responses. Python is being used to create the library. In this example, I've used the requests library to submit queries to the Spotify WEB API and have received a json response containing the information I required.

Google Cloud Storage [20] - Google Cloud Storage is a service that allows you to store your data on the Google Cloud. An object is a piece of data that cannot be changed and is composed of a file in any format. Buckets are containers that are used to hold items. Every bucket is connected with a certain project. In this case, We utilised Google Cloud Platform Buckets to store the user's photos, the emotion prediction model, the song classification model, the haarcas-cade frontalface default.xml file, the anticipated emotion of the user, and the categorised songs information..

Google Cloud Pub/Sub [19] - Using Google Cloud Pub/Sub, services may interact asynchronously with latencies on the range of 100 milliseconds or less. Publishers and subscribers are terms used to describe systems of event producers and consumers that are created using this technology. It is up to the publisher to decide how and when to deliver events to the Pub/Sub service. After then, Pub/Sub distributes events to any services that need to respond to them. In this case, We utilised the Pub/Sub function to broadcast the playlist ID of the playlist that the user had selected.

Google Cloud Functions [19] - Google Cloud Functions is a serverless execution environment that allows you to construct and connect cloud services without having to worry about servers. It is possible to construct single-purpose functions that are coupled to events broadcast by your cloud infrastructure and services using Cloud Functions, which is a basic programming language. When an event that is being monitored occurs, these functions will be called as a result. In this case, I utilised the Cloud Function to trigger when a user's image is uploaded to a cloud bucket in order to anticipate their emotional state of mind. In addition, when a playlist ID is published to the Pub/Sub, a trigger is generated to categorise the songs in the playlist.

TensorFlow and Keras - TensorFlow is a numerical computing framework that is Python-friendly and that makes machine learning quicker and easier. It is available as a free download. Deep learning API Keras is written in Python and runs on the machine learning platform TensorFlow. Keras is an open source project developed by Google. In this example, I utilised TensorFlow and Keras to create a Convolutional Neural Network (CNN) Model for emotion recognition using the TensorFlow library.

OpenCV - OpenCV is a huge open-source library for computer vision, machine learning, and image processing. It can process images and videos to identify objects, faces, or even the handwriting of a human. Here, I used OpenCV to read and manipulate images.

Numpy - In computer vision, machine learning, and image processing, OpenCV is a massive open-source library that is free to use. It is capable of processing photos and movies in order to recognise items, faces, and even the handwriting of a living being. OpenCV was utilised to read and alter the pictures in this case.

Pandas - Pandas is a Python package that provides fast, flexible, and expressive data structures that are meant to make dealing with "relational" or "labelled" data both straightforward and intuitive. Pandas is a data structure that is fast, flexible, and expressive. It aspires to serve as the essential high-level building block for performing realistic, real-world data analysis in Python at the highest degree of abstraction. The Pandas module was used to read and preprocess the Song Classification dataset in this case.

Sklearn - Scikit-learn (also known as Sklearn) is the most usable and stable machine learning toolkit available for Python. Using Python, it provides a consistent interface to a variety of fast machine learning and statistical modelling methods such as classification, regression, clustering, and dimensionality reduction, among others. For the categorization of songs, I utilised the Random Forest Classifier and MinMaxScaler modules from the Sklearn library.

a) **Emotion Detection Model**

Emotion detection model is a Convolutional Neural Network(CNN) model build using Keras Sequential Model and it is trained using the FER2013 Dataset(Facial Expression Recognition 2013 Dataset) to predict happy or sad mood images.

b) **Song Classification Model**

Song Classification model is a Random Forest Classifier from Sklearn, it is trained using Supervised Learning by the Spotify Music data to identify the moods dataset taken from Kaggle by preprocessing the data using MinMaxScaler from Sklearn to predict happy or sad mood songs.

c) **Emotion Detection**

The main objective of this session is to capture images so here we are using the common device i.e., webcam. In the application, I used webcam.js to access the webcam and to capture the image of the user, then the

captured image is uploaded to the Google Cloud Storage and detect the facial emotion of the user's image upload to the cloud storage bucket. Firstly, from the image face of the user is detected using Haarcascade Algorithm designed by OpenCV is to detect the frontal face using the xml file haarcascade_frontalface_default.xml. Then the detected face is sent to Emotion detection model to predict the person's mood.

d) Song Classification

The main objective of this session is to classify the songs of the playlist selected by the user. Firstly, we extract the song features of the songs present in the playlist using the Spotify WEB API and then preprocess the extracted data using MinMaxScaler and the data is sent to the Song Classification model to classify them either to happy or to sad mood songs.

e) Create Playlist of Recommended Songs

The main objective of this session is to create a playlist for the songs recommended for the user by the system. Firstly, from the classified songs a list of tracks is selected based on the user's emotion and then, using Spotify API a playlist is created in the user's Spotify Library with the above selected list of tracks.

5. Performance

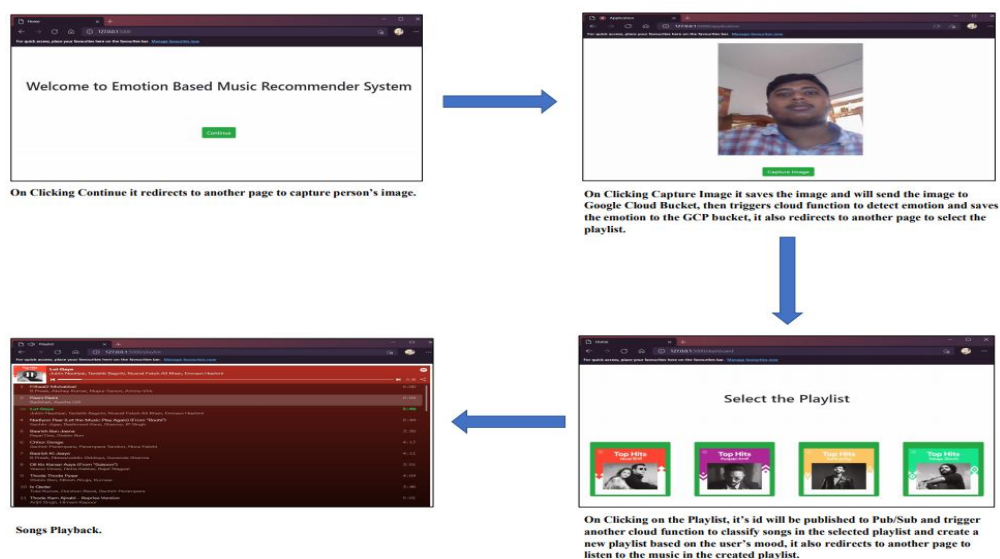


Fig. 20. Overall system Performance

5 Conclusion

In this project, The Emotion Based Music Recommender System is utilized to keep away from manual work of selecting song for their current mood and improves the listening experience of the end-user. This project is built to get the better music recommendations to the end-user by allow the music system to interact with the end-user. This system will make the work of the end-user easy by capturing the image using their webcam, predicting their emotion, and suggests a customized playlist using end-user's preference and their current emotion/mood. This music-system will decrease the work and time of end-user by generating a playlist of songs according to the end-user's facial emotion. We can use different approaches to get the emotion of the person more accurately (like EEG signals) and also, we can take more emotions into consideration to predict and classify. Instead of using Spotify API in future we can develop our own platform and databases and recommend accurate songs to the user.

References

- [1] Gangopadhyay, I., Chatterjee, A., & Das, I. (2019). Face detection and expression recognition using Haar cascade classifier and Fisherface algorithm. In *Recent Trends in Signal and Image Processing* (pp. 1-11). Springer, Singapore.
- [2] J Kuo, F. F., Chiang, M. F., Shan, M. K., & Lee, S. Y. (2005, November). Emotion-based music recommendation by association discovery from film music. In *Proceedings of the 13th annual ACM international conference on Multimedia* (pp. 507-510).
- [3] Ferwerda, B., & Schedl, M. (2014, July). Enhancing Music Recommender Systems with Personality Information and Emotional States: A Proposal. In *Umap workshops*.
- [4] Andjelkovic, I., Parra, D., & O'Donovan, J. (2016, July). Moodplay: Interactive mood-based music discovery and recommendation. In *Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization* (pp. 275-279).
- [5] James, H. I., Arnold, J. J. A., Ruban, J. M. M., Tamilarasan, M., & Saranya, R. (2019). EMOTION BASED MUSIC RECOMMENDATION SYSTEM. *EMOTION*, 6(03).
- [6] Ka, ShanthaShalini, Jaichandran Ra, Leelavathy SA, Raviraghul Ra, Ranjitha Ja, and Saravanakumar Na. "Facial Emotion Based Music Recommendation System using computer vision and machine learning techniques." *Turkish Journal of Computer and Mathematics Education* Vol 12, no. 1 (2021): 912-917.
- [7] Bansal, D., Bhatt, P., Dusane, M., Saluja, A., & Patel, K. (2020). EMOTION BASED MUSIC PLAYING DEVICE.
- [8] James, H. I., Arnold, J. J. A., Ruban, J. M. M., Tamilarasan, M., & Saranya, R. (2019). EMOTION BASED MUSIC RECOMMENDATION SYSTEM. *EMOTION*, 6(03).
- [9] Nimmagadda, S. C. (2017). Emotion Based Music Player. Department of Computing Sciences Texas A&M University-Corpus Christi Corpus Christi, Texas.
- [10] Kabani, H., Khan, S., Khan, O., & Tadv, S. (2015). Emotion based music player. *International Journal of Engineering Research and General Science*, 3(1), 2091-2730.
- [11] Gilda, S., Zafar, H., Soni, C., & Waghurdekar, K. (2017, March). Smart music player integrating facial emotion recognition and music mood recommendation. In *2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET)* (pp. 154-158). IEEE.
- [12] Velusamy, S., Kannan, H., Anand, B., & Sharma, A. (2014). U.S. Patent No. 8,879,854. Washington, DC: U.S. Patent and Trademark Office.
- [13] Andreasson, M. M. F. (2012). U.S. Patent No. 8,094,891. Washington, DC: U.S. Patent and Trademark Office.
- [14] Song, K. T., & Cervantes, C. (2017). U.S. Patent No. 9,570,091. Washington, DC: U.S. Patent and Trademark Office.
- [15] Khan, Y., Huybrechts, C., Kim, J., & Butcher, T. C. (2018). U.S. Patent No. 10,068,588. Washington, DC: U.S. Patent and Trademark Office.
- [16] Pereg, O., & Wasserblat, M. (2008). U.S. Patent Application No. 11/568,048.
- [17] Spotipy 2.0 documentation: <https://spotipy.readthedocs.io/en/2.18.0/>
- [18] Spotify Developers: <https://developer.spotify.com/>
- [19] Google Cloud Docs: <https://cloud.google.com/docs>
- [20] Google Cloud Storage: <https://googleapis.dev/python/storage/latest/index.html#>
- [21] Flask Docs: <https://flask.palletsprojects.com/en/2.0.x/>
- [22] FER 2013 dataset: <https://www.kaggle.com/msambare/fer2013>
- [23] Spotify Music data to identify the moods dataset: <https://www.kaggle.com/musicblogger/spotify-music-data-to-identify-the-moods>
- [24] Webcam.js: <https://cdnjs.com/libraries/webcamjs>

Implementation and Comparative Analysis of Various Energy Efficient Clustering Schemes in AODV

S. Dhanabal¹, K. Vengatesan², *P. William³, Udayabhanu N P G Raju⁴, S. Sangeetha⁵

¹Associate Professor, Department of Computer Science and Engineering,
Kongunadu College of Engineering and Technology, Tamil Nadu

²Professor, Department of Computer Science and Engineering,
Sanjivani College of Engineering, Kopargaon

³Research Scholar, Department of Computer Science and Engineering,
School of Engineering and Information Technology, MATS University, Raipur

⁴Assistant Professor, Department of Computer Science and Engineering,
Aditya College of Engineering, Andhra Pradesh

⁵Assistant Professor, Department of Computer Science and Engineering,
Kongunadu College of Engineering and Technology, Tamil Nadu
william160891@gmail.com

Abstract: The wireless sensor network allows for the surveillance of physical environments as well as the implementation of timely actions based on sensor data. The major objective of this sort of network is to extend the life of the network by utilizing energy-efficient connection technologies. Cooperation between sensor nodes and agents, on the other hand, must occur as quickly as feasible. This study offers a clustering methodology for building an automated framework for a WSN network's organization. We explore MANET security problems and suggest a few potential research areas. Many key and trust management systems have been devised to avoid external assaults from outsiders, and various secure MANET routing protocols have been proposed to prevent internal attacks from within the MANET system. A novel intrusion detection framework built particularly for MANET has been studied in terms of intrusion detection. Both preventive and detection approaches will be utilized to address the security problems in MANET. A hierarchical, salable energy optimization-based network model is used in the proposed clustering protocol. Because sensor node battery resources are limited and the physics detection cycle is brief, one of the most significant wireless sensor network techniques is to create energy-efficient network life routing protocols (WSN). This study presents a technique to heterogeneous clustering that is energy-efficient.

Keywords: WLAN, WSN, MANET, Energy Efficient, Throughput, End-2-End Delay

1. Introduction

Today, Wireless Sensor Networks (WSN) are of great interest not only to the scientific community but also to the business world. The development of this type of network poses new problems in software and hardware engineering, but at the same time brings scenarios that were previously thought of as science fiction closer to reality.

Sensor networks have become a reality as microelectronic-mechanical systems technology, wireless communications and digital electronics converge [1]. This has led to an increase in the number, scale and complexity of existing WSNs [2]. Wireless sensor networks are composed of sensor nodes and a sink node. In this network architecture, the sensor nodes are passive devices that only collect data locally and the sink in a global way, that is, it collects information from the whole network and does not interact in the environment to modify it according to the information obtained. In this case, human personnel are required to review the data collected in the sink and execute actions in response to such information. The need of WSN to guarantee a fast response to sensor nodes in an automatic way has allowed the development of Wireless Sensor and Actor Networks (WASNs), which represent an important extension of WSNs [3].

WSNs present a new paradigm based on the collaborative effort of a large number of sensors deployed near or within the phenomenon to be observed, which allows the provision of diverse services for numerous applications. This new network architecture can be applied to a variety of commercial, scientific and military applications such as environmental monitoring, industrial plant sensing and maintenance, military

surveillance, medical sensing, among many other new applications. In addition, the technologies developed can also be applied to aerospace industries [4].

A wireless sensor network is typically made up of a collection of low-cost, energy-efficient, multifunctional sensor nodes deployed in an area of interest. These nodes are small in size, but they contain sensors, integrated microprocessors, and transceivers, which allow them to detect, process data, and communicate. They communicate over short distances via a wireless medium and collaborate to accomplish a specific task, such as environmental monitoring or industrial process control, among other things.

1.1 WSN applications

The sensors can be used to monitor a variety of variables such as temperature, humidity, pressure, light, sound, and water quality, among others. Wireless sensor networks are less expensive and take less time to set up than wired networks, and they can be used in inhospitable environments such as battlefields, outer space, and oceans. WSNs were initially developed for military applications, but due to the low cost of sensors and the wireless communication system, a wide range of areas were explored, with significant advances made. Environmental monitoring, military applications, health, industrial process control, security and supervision, and, finally, home automation are among the most important. [5]

1.2 Parameters of a WSN network

It is important to understand the features and needs of the application for which you want to create the network when developing a WSN, because certain elements will be considered that will directly impact your influence [6]. It is crucial to remember that each application has unique requirements, therefore it is not required to take into account all of the elements throughout the design process.

1.3 Topologies in WSN

In a WSN network there can be from tens to thousands of nodes deployed in the area of interest, with a high density, being many inaccessible and unattended, prone to frequent failures, which entails a complicated maintenance of the network's topology area. To study the maintenance of this topology, three phases can be established

- 1. Pre-Deployment and Deployment:** Nodes can be launched or placed individually on site. They can be thrown from a plane, catapaulted, placed one by one by a person or a robot, etc.
- 2. Post-deployment:** After the deployment of the nodes, changes may occur in the topology también due to the position, accessibility, available energy posición and / or failures of them, as well as dinámica and according to the dynamics of the tasks.
- 3. Redeployment:** New nodes can be added to the network to replace failed nodes, or due to changes in task dynamics.

1.4 Energy consumption in WSN

The wireless sensor nodes have limited, self-contained power sources, and it is often impossible to recharge the batteries. As a result, the lifespan of a node is inextricably linked to that of its battery. When a node fails to function, it causes significant changes in the network's topology topology. As a result, energy conservation and management in the network are critical, and many research projects have been developed focusing on protocols and sensor network algorithms that optimize the use of energy administration. Among the various approaches, work has been done with sensor groups or clusters, studying sensor mobility and routing [7], and even changing the position of base stations. The communication protocols, which have been developed to ensure reliable communication and efficient use of energy energy within the network, are among the important aspects that have been developed in WSN networks. Among these protocols are those for access to the medium, links, and even routing protocols. [8].

1.5 Routing Algorithms

In the case of the TCP / IP (Transmission Control Protocol / Internet Protocol) architecture, this task is entrusted to the network level, in which the so-called routing algorithms are implemented, responsible for determining the path followed by each packet until reaching to the target.

There are several routing algorithms, those most used are Link State, Distance Vector and Source Routing [8]:

1. Link State Algorithm: a cost is assigned to each link or connection and each node manages a complete map of the network topology. Periodically, each node disseminates the cost of the links to which it is connected, and the remaining nodes update the network map and the routing table, applying an algorithm that takes into account the path at the lowest cost.
2. Vector-Distance Algorithm: the node already knows the cost of the links to which it is connected. Each node communicates to its neighbor which other nodes it can reach and at what cost. Thus, each node recalculates its own routing table following the information it has received, and using an algorithm that takes into account the path at the lowest cost.
3. Source Routing: Routing decisions are made from the source, and packets follow an established path.

2. Literature Review

A mobile ad hoc network is a wireless networking device that may be customized and arranged dynamically to form a mobile, multi-hop network. The reliance on fixed-network infrastructure is removed from mobile ad hoc networks by treating a mobile node as an intermediate transfer, which extends mobile node ranges well beyond the base transceivers. Mobile ad hoc networks extend the range of moveable nodes beyond their current level. However, if the nodes want to connect with other nodes that are not within their range, the packets must be routed from source to destination using routing algorithms. We selected the AODV routing protocol because it works well in a complicated network environment like a MANET. What happens in this scenario will result in specific security concerns. In the event of such an occurrence. Wormhole attacks are one type of successful denial of service attack. Because routing is the foundation of all mobile ad hoc communication, routing failures will bring the entire system to a halt.

LEACH (Low Energy Adaptive Clustering Hierarchy) [9] is an acronym for Low Energy Adaptive Clustering Hierarchy. This protocol chooses some nodes at random, such as CH, and rotates this function to distribute the energy load evenly among the network nodes. Each CH compresses the data from its group's nodes and sends the resulting packet to the base station at env a. Data collection is done on a regular and central basis, and it is useful when the sensor network needs to be constantly monitored. After a set amount of time, the role of CH is randomly rotated, resulting in uniform energy dissipation in the network (disipacio n). There is a variant of this protocol (LEACH-C [9]) that uses a centralized algorithm for cluster formation and can produce better results.

PEGASIS (Power-Efficient Gathering in Sensor Information Systems) [10]: In this protocol, nodes need only communicate with their nearest node in the chain, until the last node communicates directly to the base station. The collected information is transferred node by node in the chain, merged and transmitted to the sink by the final node. The protocol works on a round basis, alternating the node that has direct communication with the base station, so that power consumption is distributed evenly among all the nodes. On the other hand, the bandwidth required for communication is reduced due to the local coordination between nearby nodes.

TEEN (Threshold-Sensitive Energy Efficient Protocols) and APTEEN (Adaptive Periodic Threshold-Sensitive Energy Efficient Protocols) [11,12]: These protocols were created for applications that require a quick response to sudden changes in the parameters to be measured. TEEN nodes constantly monitor their surroundings, but data transmission is sporadic. Each CH sends a hard threshold (HT) to its group, which is the threshold value of the measured attribute; however, env an also sends a soft threshold (H S), which is a minor variation in the value of the attribute that causes the node to activate its transmitter and transmit. Only when the measured attribute is in the range of interest (above HT) and the difference from the previous transmitted measurement is greater than HS will the nodes transmit. The threshold values in APTEEN are modified on a regular basis based on the needs of the user and the type of application.

MECN (Minimum Energy Communication Network) [13]:. This protocol adapts dynamically to node failures or the deployment of new sensors, although it assumes that all the nodes can transmit to any other

node, which is not always possible. This is why an extension of this protocol was developed, called SMECN (Small Minimum Energy Communication Network) [13], which considers the possible existence of obstacles between nodes.

SOP (Self Organizing Protocol) [14]: Used in networks with heterogeneous sensors of varying capacities and functionalities, as well as mobile and stationary nodes. The router nodes are fixed; they self-organize in a network to implement the algorithm, whereas the special nodes only keep track of the active routers and those nearby. The algo-rhythm is divided into four stages: discovery, organization, maintenance, and self-organization. This algorithm has low table maintenance costs and a balanced routing hierarch. However, it can cause an overload, particularly in the organization phase of the algorithm, and if there are many outages in the network, the likelihood of using the self-organization phase increases, increasing operational costs.

Sensor array routing [15]: The goal of these protocols is to collectively monitor a target's behavior in a particular environment. A sensor suite is made up of network nodes that meet specific grouping criteria for a collaborative processing activity. [29] suggested three methods based on this concept: DAM (Distributed Aggregate Management), EBAM (Energy-Based Activity Monitoring), and EMLAM (Expectation-Maximization Like Activity Monitoring).

VGA (Virtual Grid Architecture) 16]: Within the network, data aggregation and processing are utilized to extend network life. Topologism is based on the assumption that the nodes have a fixed topology. To construct a linear virtual topology with a CH for each zone, we generally utilize square clusters. Aggregation takes place on two levels: local and global. The set of CH is referred as the Local Aggregator (LA), and it is used to conduct local aggregation, whilst a subset of these is referred to as the Master Aggregator (MA), and it is used to do global aggregation.

HPAR (Hierarchical Power-Aware Routing) [17] splits the network into clusters of geographically adjacent sensors, each of which is considered as a separate entity. Each zone is permitted to select how to route a message hierarchically via the other zones in order to maximize the battery life of the requirement system nodes. The messages are routed through the max min route, which is the maximum energy reserve route maxima of all remaining energy routes.

TTDD (Two-Tier Data Dissemination) [18]: This technique offers data delivery to multiple mobile base stations. The logic behind this notion is because utilizing high residual energy nodes can be more expensive than using the quickest way with the least amount of remaining energy consumption. Each information source builds a mesh ahead of time that is used to distribute data to mobile sinks, assuming that the sensors are fixed and their position is known. For data retransmission, multi-hop techniques are employed.

EECR (Energy Efficient Clustering Routing) [19]: The method suggests that the sink and member nodes work together to build clusters and choose the CH. The sensors are organized into clusters and can function in one of two modes: sensor mode or CH mode. This lets the algorithm to be more sophisticated, allowing for a more balanced connection between modes and energy savings.

Location-based routing The nodes are identified from their locations [20]. The distance between neighboring nodes can be estimated on the basis of the levels of the entrance and their relative coordinates through the exchange of information between neighbors [21]. The location of the nodes can also be obtained from a low-powered global positioning system (GPS) contained in the nodes. In many of these protocols, when there is no activity, the nodes are kept asleep, in order to save energy.

3 Routing Protocol

3.1 Traditional Routing Protocol in WAN Network

This section will review the protocols traditional developed for solve the problem of routing in networks wireless ad hoc mobile. These protocols are divided into two groups (Royer and Toh, 1999): proactive or table-driven and reactive or on demand. A third group is also added which is a combination of the previous ones, known as hybrids [22,23].

The proactive protocols are those that store information corresponding to the entire network topology and the routes to each of the possible destinations [24]. Therefore, they periodically emit packets to discover new devices on the network and routes to them, so that they can be used in the future. To save this information

and keep it updated, they use routing tables. The main advantage of these protocols is that if a device wants to send a packet to another device, it can obtain the routing information quickly and easily. Among its disadvantages is that to store this information, which is proportional to the size of the network, it requires a lot of memory in each of the devices.

Reactive protocols are distinguished by the fact that they perform route discovery only when a device wishes to communicate with a destination device, utilizing the route request message flooding process to reduce the number of control messages in the network [25]. When a route is discovered, it is kept until the destination is inaccessible or the route is no longer needed, which reduces the size of the routing tables. A disadvantage of reactive protocols is that if the devices do not have a baby an available path to communicate with a target device, this must be discovered, adding a significant time delay in the shipping of the first package, as well as increasing the possibility of net saturation due to the flood process. The reactive protocols AODV (Ad-hoc On Demand Distance Vector) and DSR (Dynamic Source Routing) will be examined again. The last type of protocol to examine are hybrid protocols, which are preferred in networks of a certain size and with a certain number of devices. The majority of these protocols use the zone concept, in which routing within the zone is performed proactively.

3.2 Optimized Link State Routing (OLSR)

The OLSR protocol proposed by [26] is a proactive routing protocol based on the state of the links, designed to work in a completely distributed way and that does not require an or - none of your packages. Being a proactive protocol, OLSR maintains routes for all destinations on the network, so it periodically sends control messages that allow it to learn the network topology. To reduce the amount of control messages, sent by broadcast or flood, he introduces the concept of Multipoint Relays (MPR). These are a set of devices selected from the same neighborhood within a hop of each device, capable of covering devices that are at a distance of a hop as shown in Figure 1.

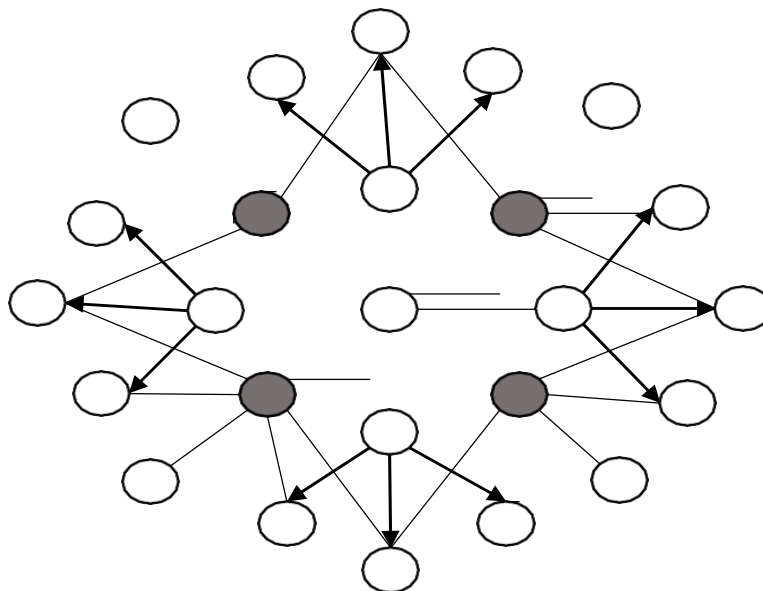


Fig. 1. Multipoint Relay in OLSR

Also, during the transmission of control messages, the protocol generates traffic and extra before the thereof a link or a new addition. The OLSR protocol manages three kinds of tables: neighbor tables, topology tables, and routing tables. The neighbor table is built from HELLO messages that are broadcast and contain information about neighbors at one hop, the state of the links, and information about their neighborhood at two hops. While the topology table is fed by Topology Control messages, which contain the addresses of all MPR-marked devices. Finally, the routing table is built using the information from the previous two tables, with each entry containing the destination's address, the address of the next hop, and the estimated distance to the destination.

3.3 Destination Sequenced Distance Vector (DSDV)

The DSDV protocol is a hop-to hop distance vector protocol proposed by [27], which allows multi-hop routing between network devices. Each device in the network maintains a routing table listing all possible destinations and the jump required to reach them. Each of these entries is linked to a sequence number originated by the destination device which indicates how old the route is, the higher the sequence number the more up-to-date the route information use of this sequence number makes DSDV a cycle-free protocol. DSDV must be periodically transmitting information corresponding to the network topology periodically and in case of a change in the network topology, immediately notify each device to keep its tables updated.

When a device discovers through these control packages that a route on its table is no longer available, the device creates a new package with a higher sequence number and the jump number marks, immediately after it notifies its neighbors of the change in topology. DSDV requires many control packages to keep route information up to date.

To reduce the amount of information contained in the packages, two types of packages are defined below as:

Full Dump: This package carries all the information available about the routing in the network.

Incremental Dump: This package only contains information that has changed since the last full dump received. Due to the constantly changing topology of an ad hoc mobile wireless network, the DSDV protocol requires all devices to periodically send all routing information they have learned periodically or to do so immediately when a change in information occurs. The main disadvantage of this protocol is that the information must be received by every device in the network, so it will take a time for the information to converge, producing a low in high density networks.

3.4 Ad-hoc Demand Distance Vector

The Ad-hoc Demand Distance Vector (AODV) protocol proposed by (Perkins and Royer, 1999) is based on the DSDV protocol; however, unlike the latter, AODV is purely reactive, i.e. a device does not need to discover or maintain a route with another device until the two need to communicate, in other words, route acquisition is purely on demand.

Only send discovery packages as required. Distinguish between the upkeep of local connectivity (neighborhood detection) and the overall upkeep of the topology. Distribute information about changes in local connection to neighboring devices that may require it. A fundamental feature of the AODV protocol is that, before providing routing information, the destination devices generate a destination sequence number (a concept borrowed from the DSDV protocol), which serves as a tool for determining when a specific route has been updated while avoiding the formation of loops. As a result, if a device must pick between many pathways to a destination, it will select the one with the higher sequence number, which corresponds to the most current routing information. The major advantage of AODV is that it does not generate additional traffic for communication; nevertheless, it has the restriction that it takes time to establish a connection and that the first conversation to build a route is more difficult than with other protocols.

3.4.1 Route discovery in AODV

When a device wishes to send a packet to another device, it first searches its routing database for a route to the destination, if one exists in the data packet; otherwise, it begins the route discovery process. In this scenario, the gadget will broadcast an RREQ message. The various parameters including different source and destination address are all included in the RREQ message. The gadget will wait for an RREP (Reset Route Message) message to arrive. If an RREP message is not received within a certain amount of time, the device will wait for it. The site will retry a certain number of times. If it fails again, the gadget concludes that there is no way to reach that location. When a device gets an RREQ message for a destination, it may do one of the following actions: Because the device that got the RREQ message is unaware of the mistake, it broadcasts the message to its neighbors. It also saves the route back to the destination device [28].

The destination device is the device that receives the message or has route information to destination t. In either scenario, it sends an RREP message back to the source device through the same way that the initial RREQ message was received. When the source receives the RREP message again, a route to the t-destination

is formed and available for usage. All devices keep a routing table containing information about each destination of interest, which is generated from received RREQ messages and stays active if utilized by any active adjacent device.

The following information is contained in each entry in this table:

Address of the destination device.

- Direction of the next jump.
- Number of jumps to reach the destination Sequence number of the destination.
- Active neighbors for that route.
- Expiry time for the route.
- Route maintenance in AODV

The devices' mobility has no effect on the active routing path to a good goal. If the moving device is the source device of an active route, the route discovery process can be resumed to find a new path to the destination. When the destination device or one of the intermediate devices is relocated, Destination receives a special RREP message. Network devices that are part of an active route transmit HELLO messages (a specific RREP message) to their neighbors on a regular basis. Failure to re-enter these messages will be construed as a loss of device connection. It broadcasts an RERR message (error message), and any device that receives it will cancel the routes that go via the device that has become unreachable. If a source device receives an RERR message, it may resume the route discovery process if the route is still needed. To ascertain this, the device must check to see if the route has recently been utilized. If it decides to reconstruct the route to the target device, it sends an RREQ message with a sequence number greater than the previously known one to guarantee that a new viable route is built.

Table 1. Comparative Analysis of Existing Protocol

S.No	Pause Time	No. of Nodes	40 (Energy of Node)		
		NOC	OLSR	AODV	DSDV
1.	1	5	16.28	57.54	24.98
2.		10	138.09	52.69	35.09
3.		15	48.50	114.89	77.69
4.	2	5	14.71	45.13	141.18
5.		10	79.45	95.04	501.30
6.		15	70.19	39.63	213.91
7.	3	5	9.07	78.10	27.49
8.		10	20.62	20.37	370.94
9.		15	89.29	19.23	253.89

4 Proposed Work

The suggested technique comprises of two packet forwarding mechanisms. When possible, the voracious technique, which consists of forwarding a data packet to a destination by using the geographical position of nearby devices to determine the next hop, employs it. Face-based routing entails sending a data packet to a destination device via a group of neighboring faces. It is employed in areas where the voracious technique does not work. To discover its neighbors, each device sends a mediating broadcast packet including the device's unique identification, its geographic coordinates (x,y), and the identifier of the cell in which it is presently situated. A jitter (i.e. delay) of 1 to 100 milliseconds is introduced to avoid synchronization with other broadcasts from nearby devices. When one of these packets is received, the device adds or updates a record in its connection database. The following illustrates the operation of the voracious technique and face-based routing using the plane partition.

Attack or Vulnerabilities in AODV

The vulnerabilities of the nodes within a WSN that allow the successful execution of The aforementioned attacks are closely related to the nature of WSNs, since as indicated in [44], nodes must have low power consumption characteristics energy and self-management, which leads to reduced computing capacities and therefore traditional security solutions cannot be applied effectively in WSN devices. Taking into account the above, in[Four. Five] the features what make to the WSN vulnerable to various attacks, the more important they are enunciated to continuation.

Device cost: The cost of the equipment is relatively low. Therefore, applying specialized security mechanisms can make this technology more expensive, thus affecting its popularity.

Susceptibility to theft: It is common to deploy nodes outdoors, without adequate protection the devices are exposed to being alienated from the network. In this way, an attacker can steal one of the nodes to extract information related to the network configuration, such as encryption keys, authentication credentials or modify the application software. This makes it easier to execute attacks since access to the data transported in the WSN is compromised, increasing the possibility of injecting malicious packets into the nodes.

Communication point-to-point or ad-hoc: In the absence of central nodes to channel network traffic, attacks can be executed from any direction and affect any of the nodes. In this way, an attacker can quickly go from hijacking data to interfering with it, causing loss of confidentiality and integrity of the information.

Balance between functionality and security: It is due to the conflict of interest of maintaining the characteristics of the nodes in terms of low resource consumption and increasing security levels. This translates into the need to develop solutions that maintain a balance between security and functionality, which is complex and difficult to apply.

Half from transmission shared: Dice what the WSN they use channels wireless shared on combination with the limiting on size, under consumption energetic and under cost from the nodes.

Limited encryption mechanisms: For many of the applications in WSN, the use of asymmetric encryption is costly due to resource limitations at the nodes. Therefore, the use of symmetric encryption protocols becomes one of the most widely used options. However, the management and storage of encryption keys is limited by the reduced storage capacity of the nodes, which results in keys of shorter length and easier to decrypt.

Given the preceding features, the WSN I am familiar with is vulnerable to a number of flaws that, if exploited, can seriously jeopardize the availability, integrity, and confidentiality of end-user and application information and data. An approach to these vulnerabilities that I am aware of focuses on the potential weaknesses and vulnerabilities to which smart homes are vulnerable. Those that are more common in the WSN are presented in this case.

Vulnerabilities what they can to be exploited to obtain successful sybil, sinkhole and wormhole attacks on real WSNs. Additionally, those vulnerabilities that are involved in the development of the attack and detection phases of this work are selected.

Table 2. Vulnerabilities exploitable by sybil, sinkhole and wormhole attacks

Vulnerability	Enumeration of attacks			
	Sybil	Sinkhole	Wormhole	Reason
A	X	X	X	Theft of encryption keys
B	X	X	X	Data from setting
C	X	-	-	Difficulty in order to detection
D	-	X	-	Injection from packages
E	X	-	-	Difficulty in order to detection
F	-	-	X	Facil decoded
G	X	X	X	Espionage / interception
H	X	-	-	False identity / impersonation
I	-	X	-	Execution Two
J	-	-	X	Loss of integrity

Where: X = Applies, - = Does not apply

It can be seen, it is possible to map the relationship between sybil, sinkhole and wormhole attacks and

vulnerabilities in WSN. In this way, the most common vulnerabilities that can be exploited by potential attackers are exposed. In these cases, each vulnerability allows the execution of one or more malicious activities, for example, vulnerability A. facilitates the extraction of encryption keys from the physically intervened nodes, with this information It is possible to decrypt the captured traffic or encrypt malicious packets to be sent within the network. On the other hand, vulnerabilities such as G. are typical of the WSN and although adequate controls are applied, it must be assumed as a latent risk that requires permanent monitoring since it is impossible to determine who or who are capturing WSN radio waves for malicious use. An exploration of the possible control measures to reduce the risk of exploitation of these vulnerabilities by sybil, sinkhole and wormhole attacks is presented.

Algorithm 1: Choice of the Next Jump using the Voracious Method

```

1: procedure RE-SENDING (package)
2:   minim Local  $\leftarrow$  true
3:   Current distance  $\leftarrow$  Distance (Current node, cell Destination)
4:   for all neighbor in V do: D Where V is the set of neighbors
5:     distance Actual  $\leftarrow$  Distance(neighbor, cell Destination)
6:   if distance Actual < distance Actual then:
7:     distance Actual  $\leftarrow$  distance V
8:   nextJump  $\leftarrow$  neighbor
9:   minimoLocal  $\leftarrow$  false
10:  if minimum Local == true then:
11:    Face Route(package)
12:  else
13:    Resend Package(next Shoot, package)

```

4.1 Environment of Attack Analysis with worm hole & without worm hole

The simulation was carried out on the NS 2 simulator, which is an event-driven simulation tool for wireless packet mode communication. It is a discrete event simulator that is object oriented and is used to investigate the dynamic dynamics of communication networks. It provides a full framework for building network protocols, generating and visualizing scenarios, and assessing their performance under specified conditions. We worked with 50 network nodes; the simulation lasted 90 seconds:

Firstly open the command user interface (CUI) or terminal on screen and type command and press ENTER Then we will enter the Desktop where the coding (.tcl) file is saved named richaworm.tcl.

Further when richaworm.tcl file runs, following figure occurs After running this file richaworm.nam file is generated which is basically an animator file and helps in showing the animation.

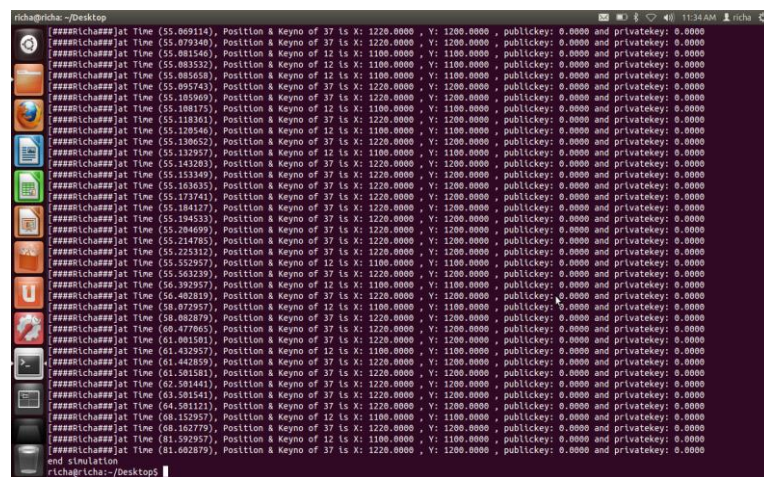


Fig. 2. After running step (ii) command

After that we have shown the file which protects worm hole and the file is named as richaisolate.tcl and to run this file .

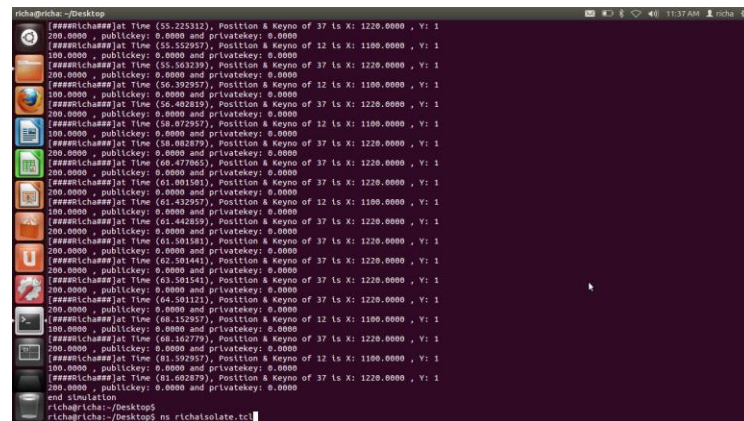


Fig.3. Command to run the .tcl file for introducing security algorithm to worm hole

Lastly, after running the richaisolate.tcl file, same happens as in step (iii). richaisolate.nam file is generated which is basically an animator file.

Now, we will discuss about the coding work done at frontend i.e. in .tcl file. In richaworm.tcl file following terms are used:

1. Channel - wireless channel is used
2. Propagation- two way ground
3. Internet information- physical, wireless
4. MAC address- MAC 802.11
5. Queue used- Priority queue
6. Link layer type- Data link layer
7. Antenna model- Omni antenna
8. Maximum packets used- 50
9. Number of mobile nodes- 50
10. Routing protocol- AOMDV
11. X axis dimension of topology- 2500
12. Y axis dimension of topology- 2500
13. Time of simulation start- 0 sec
14. Time of simulation end- 90 sec
15. Initial Energy value- 50

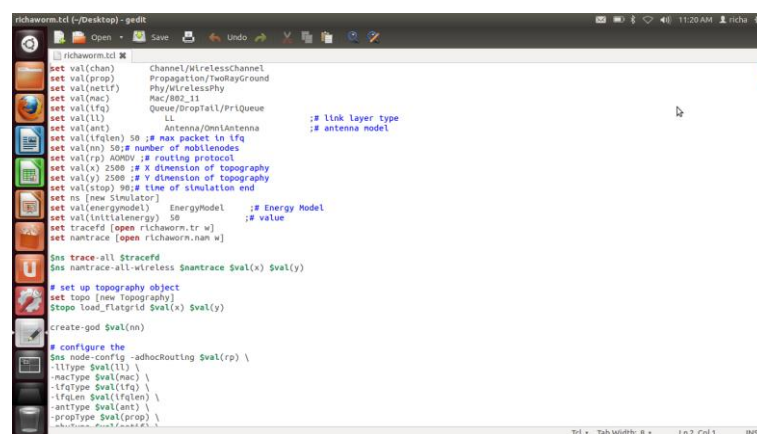


Fig. 4. Worm.tcl file coded which introduces worm hole in data communication

4.2 Parameters used for Comparison

Below given table (3) shows the comparison made between throughput with Worm hole attack and throughput after isolating Worm hole time slot from 0 - 91 seconds. Further a graph shows graphical representation of the same.

Table 3. Tabular representation of Throughput with worm hole and after isolating worm hole attack

TIME	WORM-THROUGHPUT	Proposed Work-THROUGHPUT
1	3072	3072
3	6141	6144
5	9207	9216
7	12270	12288
9	15330	15360
10	18387	18432
12	21441	21504
14	24492	24576
16	27540	27648
18	30585	30720
19	33627	33792
21	36666	36864
23	39702	39936
25	42735	43008
27	45765	46080
28	48792	49152
30	51816	52224
32	54837	55296
34	57855	58368
36	60870	61440
37	63882	64512
39	66891	67584
41	69897	70656
43	72900	73728
44	75900	76800
46	78897	79872
48	81891	82944
50	84882	86016

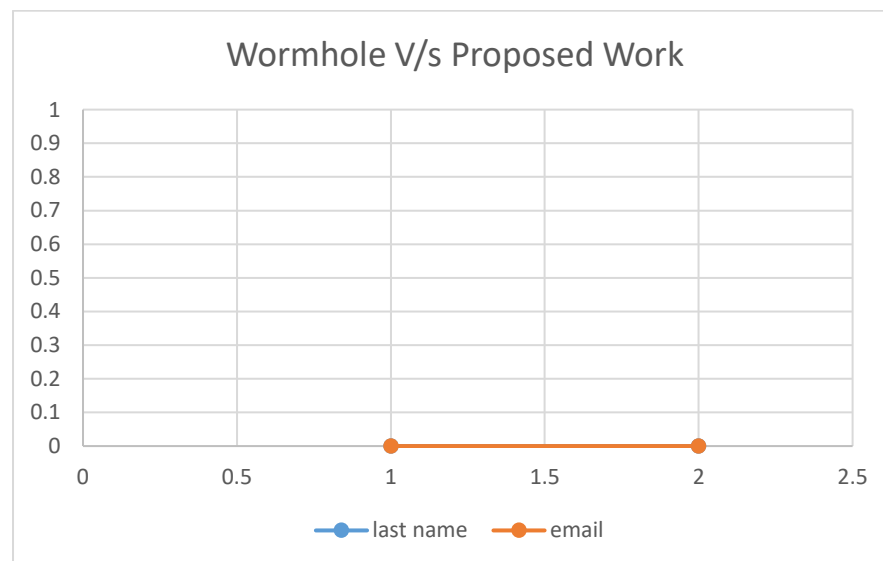


Fig. 5. Comparison of Throughput with wormhole and without wormhole

Table 4. Tabular representation of Throughput with SYBIL attack and after isolating Sybil attack

TIME	SYBIL-THROUGHPUT	PROPOSED-THROUGHPUT
1	87870	89088
3	90855	92160
5	93837	95232
7	96816	98304
9	99792	101376
10	102765	104448
12	105735	107520
14	108702	110592
16	111666	113664
18	114627	116736
19	117585	119808
21	120540	122880
23	123492	125952
25	126441	129024
27	129387	132096
28	132330	135168
30	135270	138240
32	138207	141312
34	141141	144384
36	144072	147456
37	147000	150528

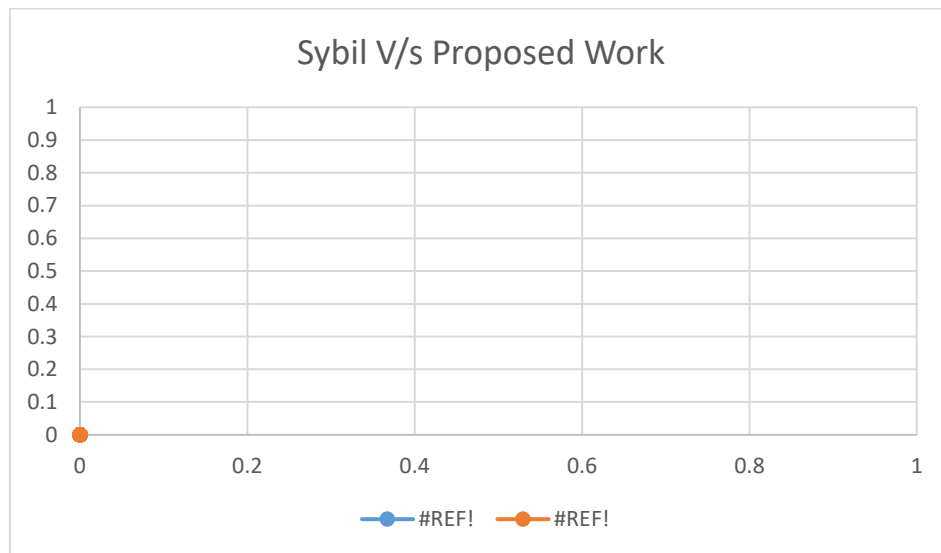


Fig. 6. Comparison of Throughput with Sybil and without Sybil

Table 5. Tabular representation of Throughput with GRAY HOLE attack and after isolating Gray hole attack

TIME	GRAY HOLE-THROUGHPUT	Proposed Work-THROUGHPUT
1	12270	12288
2	15330	15360

3	18387	18432
4	21441	21504
5	24492	24576
6	27540	27648
7	30585	30720
8	33627	33792
9	36666	36864
10	39702	39936
11	42735	43008
12	45765	46080
13	48792	49152
14	51816	52224
15	54837	55296
16	57855	58368
17	60870	61440
18	63882	64512

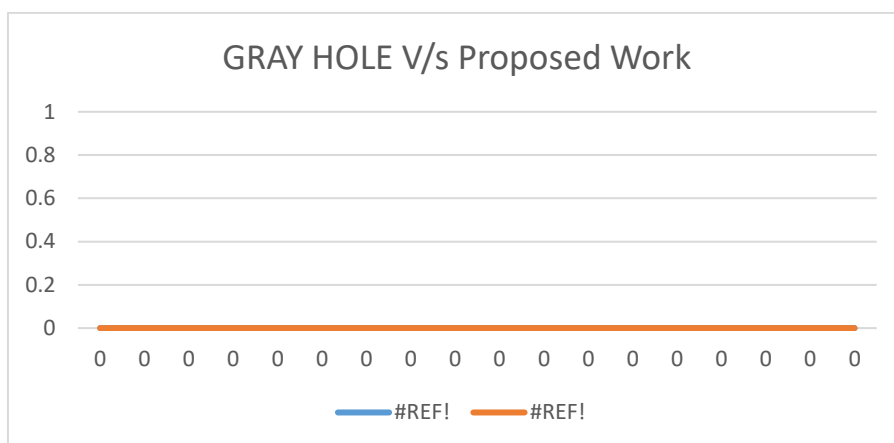


Fig. 7. Comparison of Throughput with Gray hole and without Gray hole

From above comparison it is clear that average rate of successful message delivery is more when worm hole is isolated rather than when worm hole is present over communication channel.

Utilization of energy can be in the form of heat or electricity. Above given table (5) shows the comparison made between Energy consumption during communication with Worm hole attack and Energy consumption after isolating Sybil attack during time slot from 0 - 91 seconds. Further a graph shows (Figure 4.9) the graphical representation of the same.

5. Conclusion

Technological evolution and especially in the field of communications networks has allowed more and more devices to connect to the network either by guided or wireless means, allowing the deployment of networks to places where a few years ago it was unthinkable that network services could be offered; In this context, Ad hoc mobile networks have become essential for the implementation of services that require characteristics of this type of network such as their autonomy, their ability to establish dynamic topologies as more devices access the network, the lack of need for a previously established physical infrastructure; Despite the benefits

of these networks, it should be borne in mind that they have limitations, mainly in terms of energy and processing capacity.

References

- [1] P. Rani, Kavita, S. Verma and G. N. Nguyen, "Mitigation of Black Hole and Gray Hole Attack Using Swarm Inspired Algorithm With Artificial Neural Network", in *IEEE Access*, vol. 8, pp. 121755-121764, 2020. doi: 10.1109/ACCESS.2020.3004692.
- [2] E. O. Ochola, L. F. Mejaele, M. M. Eloff and J. A. van der Poll, "Manet Reactive Routing Protocols Node Mobility Variation Effect in Analysing the Impact of Black Hole Attack", in *SAIEE Africa Research Journal*, vol. 108, no. 2, pp. 80-92, June 2017. doi: 10.23919/SAIEE.2017.8531629.
- [3] R. J. Cai, X. J. Li and P. H. J. Chong, "An Evolutionary Self-Cooperative Trust Scheme Against Routing Disruptions in MANETs", in *IEEE Transactions on Mobile Computing*, vol. 18, no. 1, pp. 42-55, 1 Jan. 2019. doi: 10.1109/TMC.2018.2828814.
- [4] Ankit Kumar & Madhavi Sinha (2019). "Design and development of new framework for detection and mitigation of wormhole and black hole attacks in VANET", *Journal of Statistics and Management Systems*, 22:4, 753-761, DOI: 10.1080/09720510.2019.1609555.
- [5] T. Li, J. Ma, Q. Pei, H. Song, Y. Shen and C. Sun, "DAPV: Diagnosing Anomalies in MANETs Routing With Provenance and Verification", in *IEEE Access*, vol. 7, pp. 35302-35316, 2019. doi: 10.1109/ACCESS.2019.2903150.
- [6] S. Djahel, F. Nait-abdessalam and Z. Zhang, "Mitigating Packet Dropping Problem in Mobile Ad Hoc Networks: Proposals and Challenges", in *IEEE Communications Surveys & Tutorials*, vol. 13, no. 4, pp. 658-672, Fourth Quarter 2011. doi: 10.1109/SURV.2011.072210.00026.
- [7] X. Li, Z. Jia, P. Zhang, R. Zhang and H. Wang, "Trust-based on-demand multipath routing in mobile ad hoc networks", in *IET Information Security*, vol. 4, no. 4, pp. 212-232, December 2010. doi: 10.1049/iet-ifs.2009.0140.
- [8] Pankaj Dadheech, Ankit Kumar, Chothmal Choudhary, Mahender Kumar Beniwal, Sanwta Ram Dogiwal & Basant Agarwal (2019). "An enhanced 4-way technique using cookies for robust authentication process in wireless network", *Journal of Statistics and Management Systems*, 22:4, 773-782, DOI: 10.1080/09720510.2019.1609557.
- [9] D. O. Akande and M. F. MohdSalleh, "A Network Lifetime Extension-Aware Cooperative MAC Protocol for MANETs With Optimized Power Control", in *IEEE Access*, vol. 7, pp. 18546-18557, 2019. doi: 10.1109/ACCESS.2019.2895342.
- [10] Y. Song, H. Luo, S. Pi, C. Gui and B. Sun, "Graph Kernel Based Clustering Algorithm in MANETs", in *IEEE Access*, vol. 8, pp. 107650-107660, 2020. doi: 10.1109/ACCESS.2020.3001137.
- [11] M. El-Semary and H. Diab, "BP-AODV: Blackhole Protected AODV Routing Protocol for MANETs Based on Chaotic Map", in *IEEE Access*, vol. 7, pp. 95197-95211, 2019. doi: 10.1109/ACCESS.2019.2928804.
- [12] Y. Chen, C. Hu, E. H. Wu, S. Chuang and G. Chen, "A Delay-Sensitive Multicast Protocol for Network Capacity Enhancement in Multirate MANETs", in *IEEE Systems Journal*, vol. 12, no. 1, pp. 926-937, March 2018. doi: 10.1109/JSYST.2017.2677952.
- [13] Z. Chen, W. Zhou, S. Wu and L. Cheng, "An Adaptive on-Demand Multipath Routing Protocol With QoS Support for High-Speed MANET", in *IEEE Access*, vol. 8, pp. 44760-44773, 2020. doi: 10.1109/ACCESS.2020.2978582.
- [14] Ankit Kumar, Pankaj Dadheech, Dinesh Goyal, Pawan Kumar Patidar, S. R. Dogiwal and Neha Janu, "A Novel Scheme for Prevention and Detection of Black Hole & Gray Hole Attack in VANET Network", *Recent Patents on Engineering* (2020) 14: 1. <https://doi.org/10.2174/1872212114999200512120211>.
- [15] S. Hao, H. Zhang and M. Song, "A Stable and Energy-Efficient Routing Algorithm Based on Learning Automata Theory for MANET", in *Journal of Communications and Information Networks*, vol. 3, no. 2, pp. 43-57, June 2018. doi: 10.1007/s41650-018-0012-7.
- [16] B. Ojetunde, N. Shibata and J. Gao, "Secure Payment System Utilizing MANET for Disaster Areas", in *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 12, pp. 2651-2663, Dec. 2019. doi: 10.1109/TSMC.2017.2752203.
- [17] N. Jiang, P. Xu, Y. Yao, T. Bui and Q. Chen, "Exploiting Radio Irregularity for Location Verification in Sparse MANETs", in *IEEE Communications Letters*, vol. 22, no. 6, pp. 1284-1287, June 2018. doi: 10.1109/LCOMM.2018.2828406.
- [18] M. Ahmad, A. Hameed, A. A. Ikram and I. Wahid, "State-of-the-Art Clustering Schemes in Mobile Ad Hoc Networks: Objectives, Challenges, and Future Directions", in *IEEE Access*, vol. 7, pp. 17067-17081, 2019. doi: 10.1109/ACCESS.2018.2885120.
- [19] T. Rahman, I. Ullah, A. U. Rehman and R. A. Naqvi, "Notice of Violation of IEEE Publication Principles: Clustering Schemes in MANETs: Performance Evaluation, Open Challenges, and Proposed Solutions", in *IEEE Access*, vol. 8, pp. 25135-25158, 2020. doi: 10.1109/ACCESS.2020.2970481.

- [20] Kumar A., Dadheech P., Beniwal M.K., Agarwal B., Patidar P.K. (2020). "A Fuzzy Logic-Based Control System for Detection and Mitigation of Blackhole Attack in Vehicular Ad Hoc Network", In: Chaudhary A., Choudhary C., Gupta M., Lal C., Badal T. (eds) *Microservices in Big Data Analytics*. Springer, Singapore. https://doi.org/10.1007/978-981-15-0128-9_15.
- [21] M. Ponguwala and S. Rao, "E2-SR: a novel energy-efficient secure routing scheme to protect MANET-IoT", in *IET Communications*, vol. 13, no. 19, pp. 3207-3216, 3 12 2019. doi: 10.1049/iet-com.2019.0039.
- [22] S. Doss et al., "APD-JFAD: Accurate Prevention and Detection of Jelly Fish Attack in MANET", in *IEEE Access*, vol. 6, pp. 56954-56965, 2018. doi: 10.1109/ACCESS.2018.2868544.

Plant image Classification: State of Art

Sukanta Ghosh¹, Amar Singh²

^{1,2}Lovely Professional University, Punjab, India
sukantaghoshmca@hotmail.com

Abstract: Plant image-data classification is a perplexing procedure which might be influenced by numerous factors. The technology of image segmentation and its classification is difficult but important problem in computer vision and generally used in medical image processing, face recognition pedestrian detection, etc. in this paper will discuss on image classification techniques include CNN, SVM, Random Forest etc. The main scope of image classification is to extract difference features of the image that are used for analysing, interpretation and understanding of images. Image classification splits a digital image-data into numerous partitions to analyse them for classification purpose. It is additionally used to recognize various partitions in the plant image-data. Successful utilization of numerous highlights of information and the choice of an appropriate classification strategy are particularly critical for classification precision improvement. This paper analyses these algorithms of image classification, also analysis and compares the advantages and disadvantages of different algorithms. Furthermore, some significant issues influencing order execution are talked about. Eventually, we make a divination of the formation orientation of image classification with the combination of these algorithms. Clearly it is seen that CNN performs better than all other classification algorithm.

Keywords: Image classification, CNN, SVM, Random Forest, Thresholding, Multivariate data, Clustering.

1 Introduction

The human capacity to break down and order images quickly and precisely is something that everyone finds profoundly helpful in regular. Thorpe and his partners found that people can order complex regular images rapidly [1]. So as to comprehend unpredictable images, the initial step is to perceive the articles and afterward perceive the class of the images [2].

Before, numerous classifiers have been created by different researchers. Various techniques are naive Bayes classifier, support vector machines, k-nearest neighbours, Gaussian mixture model, decision tree and radial basis function (RBF) classifiers [3][4]. Anyway, image identification is trying for a few reasons [3][4]. The first and most evident explanation is that there are around 10,000 to 30,000 diverse image classifications. The subsequent explanation is the perspective variety where numerous images can appear to be unique from various edges. The third explanation is brightening in which marks similar images look-a-like changed images. The fourth one explanation is foundation mess in the classifier can't recognize the image from its experience. Different difficulties incorporate scale, twisting, impediment, and intra-class variety.

In most of the cases, plant image-data classification will be done by using the mathematical classification techniques or neural network-based techniques. Image classification will be done according to predetermined procedures which shown in Figure:

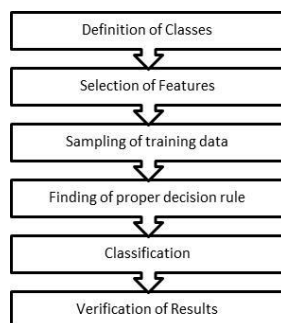


Fig. 1. Plant Image Classification Procedure

The focuses of the paper are on assessing a practical review of various optimized classification algorithms. Common classification techniques, like K-means, K-Nearest-Neighbour, Support-Vector-Machines, Random

Forest techniques, CNN and other neural network-based techniques will be discussed. We are going to use PlantCLEF dataset and weed dataset for the comparison of provided algorithm.

Plant leaf classification is a very challenging task. Most of the plant species have unique leaves pattern, even, plant in same species have unique features which makes the process very difficult. Leaves comes in different shapes, sizes, color and pattern make the classification more challenging. Sometime plants in different species have the similar vascular structure which further let to fall in accuracy. According to PlantCLEF 2018, plant image classification is still a challenging issue because of the emergence of new species. CNN provides a good accuracy figure for many datasets available online but still lacking its accuracy in many public datasets. Optimization of existing algorithm is another challenge which most of the researches are facing in context to plant image classification.

2 Plant Image Classification-Techniques

Table 1:Plant Image Classification

Category	Characteristics	Classifier Example
Supervised classification Technique	Goal is to take an input data point, class with the highest score is then the chosen (or assigned) class, focus on arriving at a reasonable loss function, sufficient image-data is available and same is used for training.	Decision Tree (DT), Max-likelihood classifier, min-distance classifier, ANN
Unsupervised classification Technique	Grouping calculations are castoff to parcel the picture information into various bunches dependent on some property. No development depictions of the classes are given.	ISODATA, k-means algorithm
Parametric classification Technique	Given information is in type of Gaussian circulation. The boundaries (for example mean vector and covariance lattice) are regularly created from preparing tests. At the point when scene is unpredictable, classifiers regularly yield 'boisterous' output.	Max-likelihood, discriminant analysis (DA)
Non-parametric classification Technique	Classifiers don't utilize factual boundaries to figure class partition and are particularly reasonable for fuse of plant image data to a grouping technique.	DT classifier, ANN, SVM, ES
Per-pixel classification Technique	Conventional classifiers ordinarily build up a mark by consolidating the spectra of all training-set pixels from a given component.	Max-likelihood, min-distance, ANN, DT, SVM
Object-oriented classification Technique	Image division combines pixels into items and arrangement is led dependent on the articles, rather than an individual pixel. No GIS vector information are utilized.	e-Cognition
Per-field classification Technique	GIS assumes a significant job in field arrangement, coordinating the information in a grouping.	GIS-based classification
Hard classification Technique	Settling on a complete choice of the land spread section that every 3-pixel is allotted to a solitary section.	Max-likelihood, min-distance, ANN, DT, and SVM
Soft (fuzzy) classification Technique	Accommodating every 3-pixel a proportion of the level of comparability for each section.	Fuzzy classifiers, pixel classifier, spectral analysis

Spectral classification Technique	Unadulterated ghostly data is utilized in picture order.	Max-likelihood, ANN, min-distance
Contextual classification Technique	The neighbouring 3-pixel data is utilized in picture characterization.	Conditional modes, P-2-P contextual classifier, frequency-based classifier

The learning algorithms are comprehensively arranged into supervised and unsupervised learning strategies. The differentiation is drawn from how the researcher orders information. In supervised learning, the classes are foreordained. These classes can be thought about as a limited set, recently showed up at by a human. In practical, certain classes of the data are labelled data.

a. K-Means Clustering

K-mean bunching is misused to amass the pixel powers in an image into K number of packs. This offers the division of an image into K segments similar forces and it gives more motorization than manual limit of a picture. [5]

Expect the elements of an image are $m \times n$; convert to a vector with $(m \times n)$ lines and 1 section. With the photos, a 3-dimensional component vector per pixel is utilized. K-implies places pixels into K number of groups bolstered likenesses of power esteems. At long last got a gathering of different sectioned picture, where each fragment is relatively homogeneous as far as pixel forces. K-implies bunching is utilized for a simple procedure for picture characterization.

Principle Component Analysis (PCA) based K-Means Clustering. It is a measurable method generally utilized for measurement decrease. Normally the clustering techniques are used for various purposes. Clustering calculations utilized for unaided arrangement of far-off detecting information change as indicated by the proficiency with which clustering takes place [6].

Fuzzy-C-Means Clustering (FCM). This algorithm allows information to have a place with multiple groups. This technique is every now and again utilized in change identification, pattern-recognition and classification. It is for the most part planned to limit the target work [7].

b. Support Vector Machines (SVM)

Vladimir Vapnik introduced the concept of SVM in year 1995. A SVM is an idea of related supervised machine learning techniques that dissect image data, perceive examples, and afterward utilized for taxonomy and recession analysis. The classification issue can be limited to thought of the two-class issue without losing of general statement. The objective is to isolate the two image classes by a capacity that is prompted from accessible preparing training image data sets [8]. Support vector machine is a two-dimensional depiction of the ideal surface developed from the directly distinguishable case [9].

SVM have a place with a group of summed up linear classification. A maximal separating hyper plane will be developed. On each side of the hyper plane two equivalent hyper planes are created using different image data. This separating hyper plane will expand the separation between both the equivalent hyper planes. It has been supposed that the speculation error is dependent on the separation of hyper plane [10].

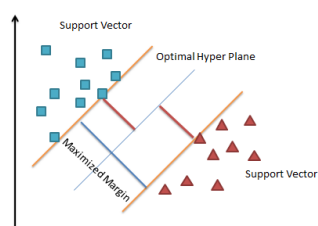


Fig. 2. Optimal Hyper plane using the SVM algorithm

SVM calculation finds the focuses neighbouring to the qualification line from both the classes. The focuses are known as help vectors. Presently, we ascertain the separation between the qualification line and the help vectors. The separation is known as the edge. The objective is to exploit the edge. The hyper plane for which the margin is max is known as the optimal hyper plane. Thus, SVM makes a choice in such a way that the separation between the two classes should be as wide as possible [11].

c. Decision Tree

Decision trees are one of the image pattern reorganizations techniques in machine learning domain. It depends on the "partition and vanquishes" methodology. It splits the image data into small partitions, where every partition has a position with one class. The partitions are succeeding by number of tests. It is made out of a root, further this root is having terminals called "takes off". The root node and inside terminals are ended with non-terminal hubs are connected into Decision stages. Every inside terminal in the Decision tree tests the estimation of some information, and the branches from the root are marked with the imaginable outcomes of the test. The terminal node expresses to convincing alliance. The alliance procedure is realized by a prearrangement of principles that establishes the way to be followed, beginning of the root node and closure at one terminal node, which expresses to the mark for the objection being ordered. At each non-terminal node, a Decision must be taken. [12] [13]

d. Random Forest (RF)

Random Forest is a collective tree-based learning algorithm. Random forest uses the concept of choosing a number of decision tree randomly from selected training image data set. It uses the collection of all decision trees and concludes with the final class of test image object. Basically, there are two different types of random forest algorithms, RF for a classification problem and RF for a regression problem [14]. RF is a type of supervised machine learning algorithm. The term forest here is a collection of relatable DT and is trained with the help of bagging training method. Bagging methods combines different types of learning models to increase the overall efficiency in results. Random forest is used to constructs multiple DT and combine them to get a more precise and constant extrapolation. Random forest is both applicable on both classification and regression problems [15].

A classification approach based on Random Forest and Linear Discriminant Analysis to classify the different categories of plants was generated. The author has discussed various challenges which are faced in plant image classification. The experiment conducted in this research article shows the improvement in the results. Experiments results shows that LDA classification have achieved an accuracy of 92.65% whereas random forest have achieved 88.82% [16][17]. The main limitation of this technique lies in the large number of trees, which makes the system slow and ineffective [18].

e. Convolutional Neural Networks (CNN)

The Convolutional neural networks are standardized forms of multilayer perceptron (MLP). They were created dependent on the working of the neurons of the human visual cortex. The main aim behind using CNN is that the user gives array of numbers and system will generate the probability of the plant image class. The working is similar to the working of human brain. When we have a look at any plant image then we memorize the pattern of the plant or leaf of plant to identify the plant. Similarly, a machine can recognize an object will looking into the granular level features of the object [19][20][21].

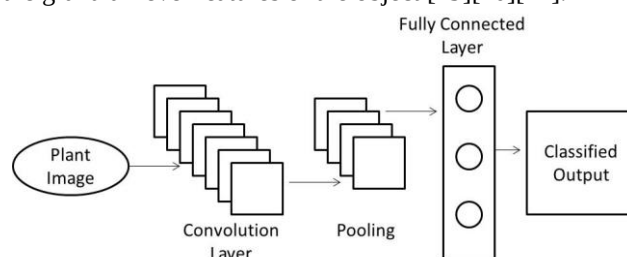


Fig. 8. Convolution neural network

A CNN consists of convolution layers, activation function (ReLU) which is followed by every convolution layer, a pooling layer especially max pooling layer and a fully connected layer. A CNN always starts with an input image. Feature map is created using different filters. ReLU function is applied to increase the non-linearity. Pooling layer is applied to each feature map. Further pooled images are flattened into one long vector. Then the vector is inserted into a fully connected artificial neural network. The features are processed through this network. The newly created model has been trained using forward and backward propagation. This is repeated till we get our optimized model for classification [22][23].

ReLU is a activation function used in CNN working. It helps the convolution layer to increase the linearity in the network. Mainly the images are basically highly non-linear in nature. ReLU also trains the network promptly without any major loss to accuracy. Other activation functions namely tanh, sigmoid etc. are also used in CNN [24][25]. Deep CNN have shown promising results in single and multi-labelled image classification. CNN has shown outstanding performance when it comes to the computer vision and image processing. Selection of hyper-parameters highly influences the performance of CNN. A minute change in hyper-parameters can impact the performance of CNN. Moreover, performance of CNN is highly dependent on the hardware of system also. A good GPU will enhance the performance [26].

3 Algorithms Performance and Discussions

Table 2: Image classification method

Article Title	Classification Techniques	Dataset Used	Accuracy
Automatic Plant Image Identification using Transfer Learning via Convolutional Neural Networks [21]	CNN	PlantCLEF	45.67%
Deep Learning and SVM Classification for Plant Recognition in Content-Based Large Scale Image Retrieval [22]	SVM	PlantCLEF	36%
Deep Learning and SVM Classification for Plant Recognition in Content-Based Large Scale Image Retrieval [22]	CNN	PlantCLEF	59%
Random Forest Boosted CNN: An Empirical Technique for Plant Classification [23]	CNN + Random Forest	PlantCLEF	67%
Random Forest Boosted CNN: An Empirical Technique for Plant Classification [23]	CNN + Adaboost	PlantCLEF	59.6%
Random Forest Boosted CNN: An Empirical Technique for Plant Classification [23]	SVM	PlantCLEF	58%
Non-local DenseNet for PlantCLEF 2019 Contest [24]	ResNet-18	PlantCLEF	51%
Non-local DenseNet for PlantCLEF 2019 Contest [24]	ResNet-152	PlantCLEF	78%
Non-local DenseNet for PlantCLEF 2019 Contest [24]	DenseNet-121	PlantCLEF	81%
Non-local DenseNet for PlantCLEF 2019 Contest [24]	Non-local DenseNet-121	PlantCLEF	86%

4. Conclusion

In this article we have tried to give an overview of the most popular algorithm used for plant image classification. A detailed review has been done on plant image classification techniques. CNN surpasses SVM, random forest and other techniques to win this battle. Performance of plant image classification is highly dependent on first stage where feature extraction is done. The performance of these plant image classification techniques is depended on the dataset used. PlantCLEF is a highly recognized dataset for plant images classification. Every year a lot of new species are added into this dataset which makes it a real ground to test the newly developed techniques. When more than two approaches are blended together there is a significant jump in the accuracy. Such more blended version of techniques could be used in future for better accuracy and performance for plant image classification.

References

- [1] S Ghosh and A Singh, "The scope of Artificial Intelligence in Mankind: A Detailed Review," in *Proc. International Conference on Recent Advances in Fundamental and Applied Sciences (RAFAS)*, Punjab, India: Journal of Physics: Conference Series, vol. 1531, pp. 012045, 2020.
- [2] Jain, A., Duin, R., Mao, J. (2000), Statistical Pattern Recognition: A Review. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22, 4-37
- [3] Thanh N. T., Ron W., Lutgarde M.C. B. (2006), KNN-kernel density-based clustering for high-dimensional multivariate data. *Computational Statistics & Data Analysis*, Volume 51, Issue 2, Pages 513-525
- [4] Rui X., Donald W. (2005), Survey of Clustering Algorithms. *IEEE Transactions on Neural Networks*, Vol. 16, No. 3
- [5] Evgeniou, T., Pontil, M. (2001), Support Vector Machines: Theory and Applications. 2049. 249-257
- [6] Yongli Z. (2012), Support Vector Machine Classification Algorithm and Its Application. *Information Computing and Applications*, pp 179-186
- [7] Boser, B. E., Guyon, I., Vapnik, V. (1992), A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, pages. 144 -152
- [8] Chan, J., Paelinckx, D. (2008), Evaluation of Random Forest and Adaboost tree-based ensemble classification and spectral band selection for ecotope mapping using airborne hyperspectral imagery. *Remote Sensing of Environment*. 112. 2999-3011
- [9] Naidu, G. G., Srinivas, Y. (2017), Survey of Methods for Effective Image Classification and Identification. *International Journal of Applied Engineering Research*, ISSN 0973-4562 Volume 12, Number 18 (2017) pp. 7132-7136
- [10] Surekha S. (2017), A Survey of Various Tree Based Classification Techniques. *International Journal of Advanced Research in Computer Science and Software Engineering*, Volume 7, Issue 3, March
- [11] Xu, B., Yunming, Y., Nie, L. (2012), An improved random forest classifier for image classification. *2012 IEEE International Conference on Information and Automation, ICIA 2012*. 795-800
- [12] E. Elhariri, N. El-Bendary and A. E. Hassanien, "Plant classification system based on leaf features," 2014 9th International Conference on Computer Engineering & Systems (ICCES), Cairo, 2014, pp. 271-276
- [13] Danny Roobaert, "DirectSVM: A fast and simple support vector machine perceptron", *Proceedings of IEEE International Workshop on Neural Networks for Signal Processing*, December 2000.
- [14] Tóth, B. P., Osváth, M., Papp, D., Szűcs G. (2016), Deep Learning and SVM Classification for Plant Recognition in Content-Based Large Scale Image Retrieval, *CLEF 2016 Proceedings, Springer*
- [15] Toma, A., Ștefan, L. D., Ionescu, B. (2017), Automatic Plant Image Identification using Transfer Learning via Convolutional Neural Networks. *CLEF 2017 Proceedings, Springer*
- [16] Kulkarni V., & Sinha, P, (2013). Random forest classifiers: A survey and future research directions. *International Journal of Advanced Computing*.
- [17] S. S. Keerthi, S. K. Shevade, C. Bhattacharyya and K. R. K. Murthy, "A fast iterative nearest point algorithm for support vector machine classifier design", *IEEE Transactions on Neural Networks*, vol. 11, no. 1, pp. 124-136, 2000.
- [18] Khan, A., Sohail, A., Zahoor, U., Qureshi, A. S. (2020). A Survey of the Recent Architectures of Deep Convolutional Neural Networks. *Artificial Intelligence Review*, doi: 10.1007/s10462-020-09825-6
- [19] Pearlina, A., S., Kumar S. S., Harini, S. (2019). A Study on Plant Recognition Using Conventional Image Processing and Deep Learning Approaches. *Journal of Intelligent and Fuzzy Systems*
- [20] Haug S., Michaels A., Biber P., Ostermann J. (2014), Plant classification system for crop /weed discrimination without segmentation. *IEEE Conference on Applications of Computer Vision*

- [21] T. Friess, N. Cristianini and C. Campbell (1998), The kernel adatron algorithm: a fast and simple learning procedure for support vector machine. *Proceedings of 5th International Conference on Machine Learning*
- [22] S. Ghosh, A. Singh and S. Arya, “Plant Disease Detection Using Machine Learning Approaches: A Survey,” in *Machine Learning and Data Analytics for Predicting, Managing, and Monitoring Disease*, 1st ed., vol. 1. Chocolate Ave, Hershey, USA: IGI Global, pp. 9, 2021.
- [23] Thanh, D. N., Qu’énott, G, Goeuriot, L. (2019), Non-local DenseNet for PlantCLEF 2019 Contest. *CLEF 2016 Proceedings, Springer*
- [24] Thorpe, S., Fize, D., Marlot, C. (1996), Speed of processing in the human visual system. *Nature*, 381, 1996, pp. 520–522
- [25] Banerjee S., Pamula R. (2020). Random Forest Boosted CNN: An Empirical Technique for Plant Classification. *Proceedings of the Global AI Congress 2019. Advances in Intelligent Systems and Computing*, vol 1112.
- [26] S. Ghosh and A. Singh, “Image Classification Using Deep Neural Networks: Emotion Detection Using Facial Images,” in *Machine Learning and Data Analytics for Predicting, Managing, and Monitoring Disease*, 1st ed., vol. 1. Chocolate Ave, Hershey, USA: IGI Global, pp. 11, 2021.

Study on Mushroom Image Classification Using Different Machine Learning Methods

¹Janvi Shetty, ²Anushree Rahate, ³Suvarna Pansambal

¹²³Department of Computer Engineering
Atharva College of Engineering
Mumbai, India
janvi.shetty25@gmail.com
anushree.r.rahate@gmail.com
suvarnashirke@atharvacoe.ac.in

Abstract: Mushrooms are one of the most familiar and delicious sustenance which are cholesterol-free and at the same time, rich in vitamins and minerals as well. Due to the tastiness and various health benefits of mushrooms, it often appears in people's daily meals even though nearly 45000 known mushrooms breeds throughout the world are poisonous and an even smaller known percentage are lethal for consumption. Identifying if or not any given mushroom is a poisonous mushroom through human eyes is quite difficult if the person in context is not an expert. Thus, building an automatic image identification of the mushrooms could be of great importance. Several methods can be deployed to recognize mushrooms, including regular experience, running chemical tests and classification of fungal groups. These methods are time-consuming and hence, cannot produce quick and accurate results. An application would be beneficial if it is real-time. By doing so, it could give its users enough information to classify on their own with the help of a detailed description of any mushroom whose image has been submitted to the software. We aim to find a reliable way for identification of mushroom edibility with a larger data-set of images that performs more efficiently than existing works. In this project, we shall study and implement the various machine learning techniques such as Support Vector Machine, Convolution Neural Network, Decision Tree techniques and find the best one to implement in our software. This project is a software to identify the type of mushroom given and if it is edible or not, using the image of the mushroom.

Keywords—Machine learning, Image analysis, Support Vector Classification, Image Data-set, Support Vectors Machine, Mushroom Genus, Poisonous, Image Processing.

I. Introduction

In most kitchens, Mushroom, a fleshy, earthy fungi is a delicacy. It has different nutritional values, and is also one of the healthy variants of good cholesterol in food. There are different genus like Amanita, Hygrocybe, Inocybe, Lactarius and many others that can be distinguished based on their physical characteristics.

Around 70 percent of natural poisoning cases are caused by ingestion of toxic mushrooms. As alarming as this number sounds only around fifty species of mushrooms are actually poisonous as compared to the rest thousand others out there. Even among these 50 species only ten are actually capable of causing deaths. There may be a very negligible percentage of harmful mushroom kinds in the world, various others can cause severe and unsightly symptoms. The lack of knowledge and ability to identify is the main reason for mushroom related fatal cases these days. Ingestion of poisonous mushrooms can result in symptoms ranging from slight intestinal pain to imminent death.

Separating edible mushrooms from lethal needs a lot of experience and detailed knowledge. All kinds of mushroom have varied features that help an expert to distinguish whether they are dangerous or not. The various features of features are the primary factor for identifying the species of the mushrooms. These features include the cap shape, the colour of the cap, if the mushroom has bruising or not, the spore colour, the stalk height, etc. Mushroom experts have already stated which species are harmful or not based on these features. Unfortunately there are too many features and exceptions in this classification for an average person to be able to spot a lethal mushroom from the rest.

The software we have developed is something that will help people identify what the genus of the mushroom they have and if it is edible or not. For our project, we shall study and implement the various machine learning techniques such as Convolution Neural Network, Support Vector Machine, Residual Neural Network and find the best one to implement in our software. The purpose of this project is to build an automated toxicity detection method with the help of their perceptive features of the mushrooms.

2 Related Works

According to a paper by J. G. van de Vooren et al. [1], the usefulness of image analysis in variety deduction of mushrooms was demonstrated. Here, in order to increase the number of morphological characteristics that can be measured, image analysis techniques were used. Shadow removal, which is a minimum of grey level preprocessing technique was used to fulfil the aim of segmenting the image into background and three objects [one cap (pileus) and two gills (lamellas)]. Their resultant model was able to distinguish 45 different breeds of mushrooms based on their genres. Red-filter was used for the segmentation and it could identify 80 percent of the mushroom data it was tested upon.

Paper by Preechasuk, Jitdumrong, et al. [2] worked in the benefit of the populace that was at risk of directly encountering harmful mushrooms. The process begins by resizing the image to 96x96 and using CNN on it. The preprocessing included rotating, zooming, shifting width wise, height wise shifting, rotating by adjusted parameters, scaling (zooming-in), shifting by width, height wise shifting, adjusting the brightness, flipping the image vertically and horizontally.

A paper by Rahmat, R. F., et al. [3] concluded that KNearest Neighbor can identify the poisonous mushrooms with an accuracy rate of 90 percent on a self-made dataset of mushroom images consisting of 40 images in all. Canny edge detection and threshold methods were used for segmentation.

A mushroom recognition application made by Lidasan, et al. [4] made use of GrabCut algorithm to separate the background from the foreground and then running the resultant image through canny algorithm for first converting into grey scale and then edge detection which resulted in a 92 percent accuracy. The drawback of using the Grabcut algorithm is that the result of the algorithm depends on the quality of the image that is fed to the application. In cases where the background and foreground image had the same color or showed close to negligible differentiation among them, the result of the image segmentation would be incorrectly done. Running Grabcut through a large dataset like ours made it even more susceptible to such errors and hence feeding such compromised images to the training set resulted in further reduction in accuracy.

Paper [5] was a basic research paper on use of Support vector machines for edible mushroom identification, where SVM is used on a csv dataset with 22 attributes of mushroom and the resultant model gave an output with an accuracy score of 100 percent.

3 Research Methodology

One of the main objectives of this project was finding an algorithm that gives us the most accurate results in the least amount of time. Efficiency was our main goal here. This project included testing out three of three image classification techniques. Convolution Neural Network, Residual Neural Network and Support Vector Machine were the methods implemented. The basic preprocessing was exactly the same in all the techniques. It is explained with the functioning of the currently functional project. It was found that for the data-set used, Support Vector Machine was the algorithm that proved to give the best results. The functioning of CNN and RNN will be explained further.

A. Convolution Neural Network

A Convolution Neural Network is an algorithm in the Deep Learning methodology which takes an image, assigns the value (weights) to various components in the image and then differentiate between the various images with the help of set weights[10]. This method was used to classify the images into the given classes. The above figure explains the basic functionality of the Convolution Neural Network. In the implementation [8], we have made a total of three layers. The Activation Function used is 'relu'. And the Max

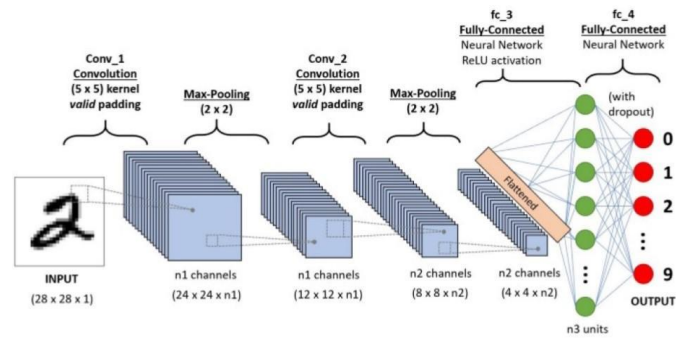


Fig. 1. Diagrammatic representation of CNN [10].

pooling used (2,2) And the activation function used in the last layer is Sigmoid. Adam optimizer is used and the model is fit for training. During fitting there is a validation split of 25 percent with three epochs.

B. Residual Neural Network

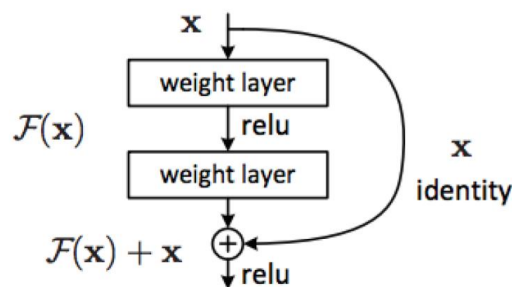


Fig. 2. Diagrammatic representation of RNN[11].

The figure above demonstrates the ‘Skip Connection’, which is basically mapping of identity. In identity mapping, there are no external parameters to be considered. Here, the input of one layer is the output of its previous layer [11]. During implementation the training, testing and validation data are pre-split into different folders. The model is built using ResNet50. For the internal layers activation function used is ‘relu’ and for the prediction layer the activation function used is ‘softmax’. Adam optimizer is used in this model as well.

C. Scikit-learn

This model was built on the scikit-learn module in python, it is a machine learning model to preprocess data and train a model. The images were first resized and transformed using ski. The images were then turned to gray-scale and Hog was calculated. The HOG of an image is calculated by dividing any image into defined blocks. Here these images were divided into 80 by 80 pixel blocks. For every individual block, a number of given directions are used to calculate the magnitude of the gradient [12]. The images were then normalised to train the model on. The SGDClassifier was used for image classification.

4 Methodology

A. Pre-processing

As the machine can only process numerical data for making predictions, the data had to be converted in the ideal format. First the images are converted into gray-scale images. This is then resized to 50×50 pixel images.

Then the Numpy module was used to convert the images in the dataset to a numerical dataset and it was stored as an object. To be able to access this numerical dataset at any time, it was stored as a pickle file using the Pickle module.



Fig. 3. Sample of resized 50×50 pixel image

B. Working of the Model

1. The working of this project is based on image classification using Machine Learning. The various classes of the images are pre-defined. The categories were still well defined in this dataset.
2. The next step was shuffling the data. Shuffling leads to a better model as it will help us when training the model. Unshuffled data will lead to training the model to identify selective classes instead of all the classes.
3. After shuffling the data is trained and tested to verify and check the accuracy of the model trained.
4. This model is saved and can be further used to make predictions that classifies it into a specific class. This project is also capable of giving the user some more information on the classified mushroom and will also redirect the user to a website for a further, more detailed, description website.

5 Experimental Results

In the case of CNN, it was deduced that it works the best for binary classification and thus it gave the absolute accuracy of 9 percent. For RNN, even as a multi class classifier, it gave the average accuracy of 57 percent with the model being tweaked. 57 percent is not as bad as the results CNN gave but there was still room for improvement. The result accuracy here was fairly low for Scikit-learn and when optimization was tried, it only gave more errors. From the table above it can be concluded that the Support Vector Machine algorithm gave the best results for the dataset [13, 14] that was used in this project. Thus, SVM was implemented to build the best possible model with highest accuracy and it also proved to be a quick algorithm in terms of predicting.

Table I

Train And Test Accuracy Of Each Mode After Tweaking Values For Optimum Accuracy

Algorithm	Train Accuracy	Test Accuracy
Convolution Neural network	0.67	0.09
Residual Neural Network	0.8	0.58
Scikit-learn	0.53	0.17
Support Vector Machine	0.92	0.82

Table 2

Train And Test Accuracy Of Each Mode After Tweaking Values For Optimum Accuracy

Category	No. of True Value	No. of False values	Percentage Accuracy
Agaricus	26	4	86.66
Amanita	80	19	80.80
Boletus	103	31	76.86
Cortinarius	82	21	76.86
Entoloma	33	7	82.5
Hygrocybe	32	12	72.72
Inocybe	56	16	77.77

Lactarius	180	31	85.30
Russula	105	34	75.53
Suillus	27	2	93.10
Total Average	692	177	82.43

6 Results

On running the 'model.predict(xtest)' code multiple times, we found that the accuracy score ranged from 80-84 percent. The result obtained showed an increase of around 10 percent than the past suggested model [2]. Our model still shows a lesser accuracy score as compared to the research done in [3, 4], but considering the large number of mushroom images that the model was fed in order to train and test, the score obtained for the model is highly satisfactory.

A. Sample Accuracy

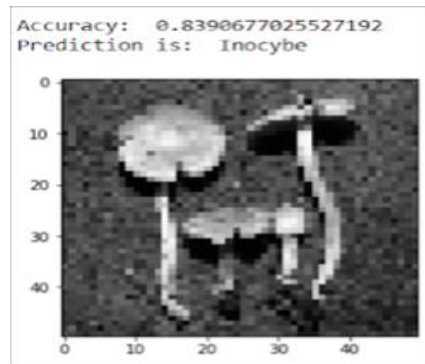


Fig. 4. Finding Accuracy and display of gray-scale image in set image size

B. False Positives



Fig. 5. Input image of Hygrocybe

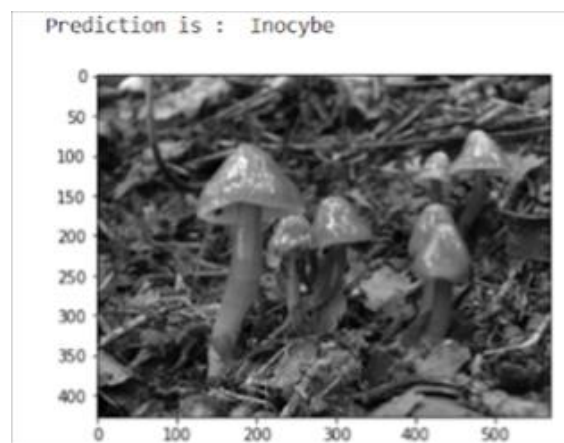


Fig. 6. Output image of Hygrocybe falsely identified as Inocybe



Fig. 7. Reference image of Inocybe

C. True and False Values per category

Sr no.	prediction made by system	truth value	true category
0	Lactarius	True	Lactarius
1	Russula	True	Russula
2	Cortinarius	True	Cortinarius
3	Lactarius	True	Lactarius
4	Inocybe	True	Inocybe
5	Amanita	True	Amanita
6	Cortinarius	True	Cortinarius
7	Lactarius	True	Lactarius
8	Lactarius	True	Lactarius
9	Boletus	True	Boletus
10	Russula	False	Hygrocybe
11	Amanita	True	Amanita
12	Russula	False	Cortinarius
13	Cortinarius	True	Cortinarius
14	Inocybe	True	Inocybe
15	Lactarius	True	Lactarius
16	Hygrocybe	True	Hygrocybe
17	Hygrocybe	True	Hygrocybe
18	Russula	True	Russula
19	Lactarius	False	Amanita

Fig. 8. Observations for the first twenty outputs

For Fig. 5, Fig. 6 and Fig. 7, as the system converts the input image into gray-scale while reading the image and then carries out the training and testing process, the false positives obtained are mostly because of similarities in the cap shapes of mushrooms of different genres. Hence, in these cases the system carries out the prediction of the image without taking into consideration attributes such as clustered/singular or colour of the cap shape. In order to increase reliability of the said prediction, an additional feature- 'to know more' button is placed in the final front end, where the user could go to an expert website to compare their findings to all the attributes associated with that mushroom and make an informed decision of its edibility.

7 Conclusion

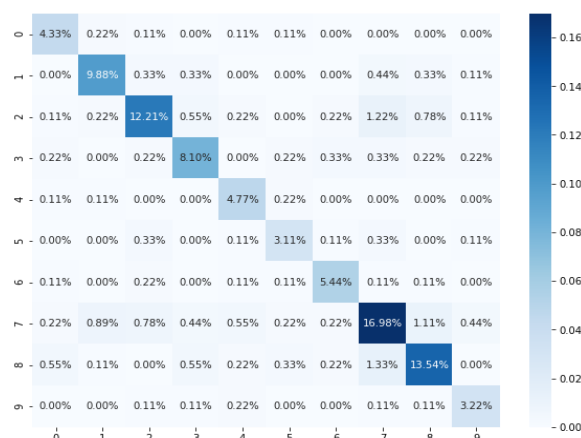


Fig. 9. Overall percentage of true and false predictions for each category of mushroom

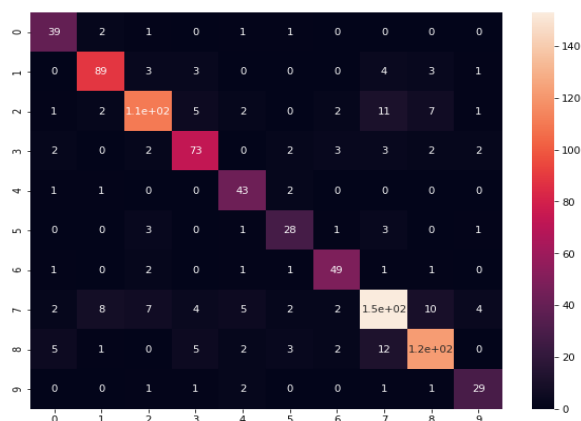


Fig. 10. True and false predictions for each category of mushroom

Due to the absence of a system that can successfully identify any mushroom at any time with results of high accuracy proofs, the methods employed till date have only been manual classification and this is inevitably subjected to high rates of human error. With the use of this online application, targeted audiences, even with minimal skills, could be greatly benefited with the services it promises to provide.

The scope of this system, not being limited to just identification of edibility of the mushroom, could be expanded into providing aid to the medicinal practitioners and paramedics.

8 Future Scope

Provided the right datasets, this application could also be used as an application for foraging fruits, berries, roots, flowers, etc. making it a true foragers guide to making use of all that nature has to offer in a camping trip. Feature of adding to the image dataset of mushrooms on every new mushroom image input received by the system can be done.

ACKNOWLEDGMENT

We would like to thank and dedicate this research to our college, Atharva College of Engineering for giving us the appropriate platform to present our project on the topic “Book of Shrooms (Image Analysis of Mushroom Types Classification)” and like to give our sincere thanks to our respected Principal, Dr. Shrikant Kallurkar for igniting the drive within us for this research and giving us the opportunities, time, and aid to carry out and present research on the topic during testing times. We are sincerely grateful for having Dr. Suvarna Pansambal as our guide and as the Head of the Computer Engineering Department who further pushed us to better our best and drove us to pursue a better idea than that was originally presented to them.

Moreover, the completion of this research would have been impossible without the cooperation, suggestions and help of our friends and family.

References

- [1] Van De Vooren, J. G., G. Polder, and G. W. A. M. Van der Heijden. "Identification of mushroom cultivars using image analysis." *Transactions of the ASAE* 35.1 (1992): 347-350
- [2] Preechasuk, Jitdumrong, et al. "Image Analysis of Mushroom Types Classification by Convolution Neural Networks." *Proceedings of the 2019 2nd Artificial Intelligence and Cloud Computing Conference*. 2019.
- [3] Rahmat, R. F., et al. "Fungus image identification using K-Nearest Neighbor." *IOP Conference Series: Materials Science and Engineering*. Vol. 420. No. 1. IOP Publishing, 2018.
- [4] Lidasan, Johaira U., and Martina P. Tagacay. "Mushroom recognition using neural network." *International Journal of Computer Science Issues (IJCSI)* 15.5 (2018): 52-57.
- [5] Wibowo, Agung, et al. "Classification algorithm for edible mushroom identification." *2018 International Conference on Information and Communications Technology (ICOIACT)*. IEEE, 2018.
- [6] "Image Classification Through Support Vector Machine (SVM) — Machine Learning" YouTube, uploaded by Md. Iqbal Hossain, 14 April 2020, www.youtube.com/watch?v=0rjlvIOQlbc
- [7] Ahadli, Tarlan. "Linear SVM Classifier: Step-by-step Theoretical Explanation with Python Implementation" *Towards Data Science*, Medium, 23 March 2020, towardsdatascience.com/linear-svm-classifier-step-by-step-theoretical-explanation-with-python-implementation-69767437e0e6
- [8] "Convolutional Neural Networks - Deep Learning basics with Python, TensorFlow and Keras p.3" YouTube, uploaded by sentdex, 9 August 2018, www.youtube.com/watch?v=WvoLTXIjBYU&t=653s
- [9] "Training Residual Neural Network with your own dataset" Youtube, uploaded by Jay Bhatt, 14 September 2020, www.youtube.com/watch?v=1GbcP66yYX4&t=378s
- [10] Shah, Sumit. "A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way" *Towards Data Science*, Medium, 15 December 2018, towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53
- [11] Shorten, Connor. "Introduction to ResNets" *Towards Data Science*, Medium, 24 January 2019, towardsdatascience.com/introduction-to-resnets-c0a830a288a4
- [12] Steegstra, Patrick. "Tutorial: image classification with scikitlearn" *Kapernikov*, 10 April, 2018, kapernikov.com/tutorial-imageclassification-with-scikit-learn/
- [13] March 19-March 19, Mushrooms classification - Common genus's images, Kaggle, Mushroom World, www.kaggle.com/maysee/mushroomsclassification-common-genuss-images/metadata. Accessed 17 Aug 2020. Dataset
- [14] September 2020- September 2020, deepmushroom, Kaggle, www.kaggle.com/anilkrishna/deepmushroom/metadata. Accessed 17 Aug 2020. Dataset

Artificially Intelligent System for Tourism Industry

¹Ashima Arya, ²Rattan Deep Aneja, ³Mitu Sehgal, ⁴Amit Kumar bindal, ⁵Shakti Kumar

Department of IT Panipat Institute of Engineering & Technology, Samalkha
Research Scholar Department of CSE Maharishi Markandeshwar (Deemed to be University), Mullana
Department of IT Panipat Institute of Engineering & Technology, Samalkha
Department of CSE ,Maharishi Markandeshwar (Deemed to be University), Mullana
Department of ECE ,Panipat Institute of Engineering & Technology, Samalkha
ashima.it@piet.co.in,rattananeja@gmail.com,mitu.it@piet.co.in
bindalamit@gmail.com,shaktik@gmail.com

Abstract: The conversational agents or chatbots have become a part of our everyday life and are designed to make our life easier. Recent advancements in natural language processing and artificial intelligence have made them surge past the popularity of social media apps as they can understand and process more information and respond more intelligently. Chatbots have covered most of the consumer and business markets and help these sectors gain huge profits due to customer engagements. In this paper we intend to provide such engaging and interactive services in the tourism industry as well. This is intended to boost the number of tourists every year thus giving a boost to the tourism sector which will further help and improve economic growth too. The voice enabled chatbot has been implemented as a Node.js application backed by a retrieval-based model. The dataset containing information on Indian monuments in a QA format was fed to the system to build intents and answer user queries. The proposed approach attempts to deliver an end-to-end system for tourists, foreign as well as domestic, visiting various monuments. The research process also led to the creation of an unavailable dataset covering almost all monuments of India.

Keywords—chatbot, NLP, Neural Network, Dataset, api.ai

1. Introduction

India has proudly owned its widely diverse culture and rich heritage for centuries now. The plethora of monuments and tourist spots the country houses, evidently vouch for this fact. Glancing through the 2017 statistics, the number of Foreign Tourist Arrivals (FTA) in India was 8.8 million with the foreign exchange earnings amounting to 1,54,146 Crore. Concurrently the number of domestic tourist visits summed up to 1613.55 million. These figures alone portray a distinct picture of the dimensions the Indian tourism industry boats of. But this huge volume of tourist visits also demands for a great number of personnel to make the tours well-informed and fruitful. This crucial job has been performed through the years by whom we know as ‘tour guides’. Although a lot of effort is put in while effectively describing and translating the relevant information about any monument, there’s always the possibility of ambiguity due to miscommunication or language differences. This is where our chatbot plays its part. As an attempt to inculcate and harness the latest in technology, namely Artificial Intelligence, the chatbot tries to reduce the communication disparities that hamper the effective transfer of information.

Chatbots are essentially digital platforms that attempt to simulate human-like conversations. The platform under discussion here uses Natural Language Processing (NLP) and Understanding (NLU) to recognize and subsequently understand what the user says, followed by intention detection in order to generate a response using retrieval-based machine learning algorithms.

An essential prerequisite to train and develop any machine learning model is feeding it with relevant information related to the application domain or usage environment. This information is generally organised into ‘datasets’. The unavailability of the required dataset during this research also led to the creation of a unique QA dataset containing and covering essential information about most of the monuments of India.

This piece of work aims to build an end-to-end intelligent system for the tourism industry that facilitates a seamless experience to tourists right from obtaining their entry pass to a premise, to actually perceiving and

knowing about it. The process (discussed further in detail) involves purchasing the entry ticket, thus eradicating queues, automated gate entry and an operational web-interfaced, voice-enabled chatbot to provide information in response to user queries using its knowledge base developed through the provided data contained in the created dataset.

2. Background And Related Work

The history of the conceptualization of chatbots dates back to the year 1950 when Alan Turing introduced the problem called the “imitation game”, which was based on the question Turing raised saying “Can machines think?”[1]. This later became to be known as the “Turing test”. In 1964, the idea of the world's first ever chatbot began taking shape at MIT's Artificial Intelligence laboratory. Developed over a period of 2 years and released in 1966, ELIZA[2] was the earliest of chatbots that made 'human-computer interaction' possible through speech as well. ELIZA was essentially a retrieval- based chatbot functioning on a pattern matching approach. Left to right parsing was applied on each input sentence to ELIZA

and each word was searched into a 'database' of keywords. The aim was to respond appropriately for each user's query depending on the keyword identified and matched in database. The release of ELIZA gave a new direction to computer scientists and linguistic researchers [springer] to work for.

The next notable development done in 1995 in the field of chat bots was ALICE[3] (Artificial Linguistic Internet Computer Entity). ALICE was developed using Artificial Intelligence Markup Language that may be considered as an extension to XML. It employs heuristic pattern matching using conversation patterns stored in knowledge-base.

ALICE won the Loebner Prize thrice in the years 2000, 2001, 2004 for being the most human-like bot by clearing the “Turing test”. While ALICE aimed at creating generic responses for a wide range of topics, having an open- domain approach, "VPbot"[4] had a more targeted topic, 'closed-domain' approach. VPbot or Virtual Patient bot was designed to simulate patient conversations in order to train medical students. It used a relational database model to store its linguistic rules and patterns.

ViDi (Virtual Dietician) presented in [5] is a virtual physician for Diabetic patients. It has a special proposed designed to remember the conversational path as well using the vPath parameter. Later an improved design was introduced in [6] using a relational database and ViDi was also updated to be used as a web-based chatbot by using HTML, PHP, XHR and AJAX.

A test environment was created for a new technique of keyword matching using ViDi [7]. This technique was called One Match or All Match Categories (OMAMC). The keywords generated by other chatbots with the same sample were used to compare the generation of keywords by ViDi technique. This technique was found to improve keyword matching.

NLP

There are inherently three procedural steps to any natural language analysis approach:

- a) voice recognition and conversion to text
- b) text processing
- c) response and action taking

As the first step, user voice is taken as an analog signal input through any microphone which then converts it into a digital stream of information. This is in the form of digital pulses which are then converted to text through certain instructions based on linguistic rules.

This converted text serves as an input to the next step of analysis which involves processing the text to understand what the user intends to say. It can be implemented through various algorithms or models. This involves breaking up the text sentences into words or tokens and then identifying the keywords present in the sentence in order to determine what the user means by studying the keyword mappings.

Lastly, an appropriate rule-based response is generated or action is executed in response to the user query.

This can either be a text response, a voice response (post text-to-speech conversion), running another program in response or facilitating any real-time application through a webhook.

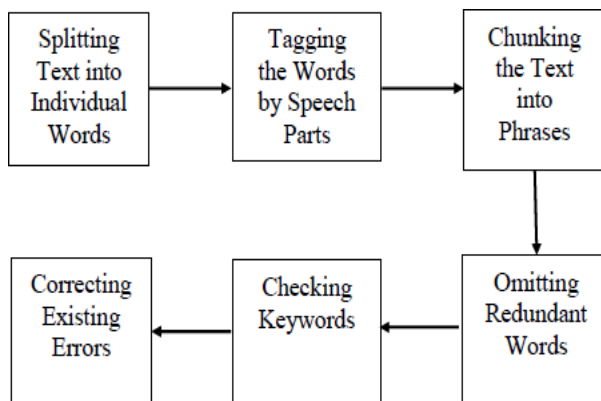


Figure 1: Natural language Analysis procedure

Recent Developments

The current stable and widely used chatbots in the market are Microsoft's Cortana, Amazon's Alexa, Google Voice and Apple's Siri. All these are open domain bots that generally rely on generating web-search based responses.

The most recent developments in natural language processing platforms have been done by giants like Google (Dialogflow), IBM (Watson), Amazon (Lex) and Facebook(Wit.ai). These can be used to build customised chatbots but require to be trained by creating intents or feeding examples of relevant data.

In the present scenario, there aren't any domain specific chatbots installed or available for Indian monuments per se. Also no relevant dataset till date is available that contains factual information regarding these monuments.

3. Chatbot Algorithms

Chatbots can be implemented by using various algorithms and text classification techniques. Since a chatbot is simply a program and does not understand human language so it uses text classification to generate appropriate response. The text classifier is a program that classifies or associates the text inputted into intents which then produce a response. Classifier is nothing but a way to categorize sentence or data into numerous categories called intents. The input “Who built Taj Mahal?” is associated with a response “Shah Jahan”. As soon as the program receives an input it starts to categorize it and will match the intent with a suitable response. The text classifiers are based on three approaches – pattern matchers, algorithms and neural networks. Any of the classifiers used produce the same end result that is a response. Specific “parts of speech” may be referenced by a response in a sentence. Depending on the state of conversation using conditional logic, different responses can be provided for intent.

Pattern Matchers

Pattern matching was used by early chatbots to match text and generate response corresponding to that text. It was basically a brute force approach since each and every pattern had to be defined for which there was a response. AIML (artificial intelligence markup language) is used as a standard structure for these patterns. AIML is an XML dialect for creating natural language software agents.

A simple pattern matching explanation:

```
<aiml version = "1.0.1" encoding = "UTF-8"?>
<category>
<pattern> WHO BUILT QUTUB MINAR </pattern>
<template> Started by Qutub-ud-din Aibak and completed by
his son-in-law Iltutmish.</template>
</category>
<category>
<pattern> WHO BUILT TAJ MAHAL </pattern>
<template>SHAH JAHAN.</template>
</category>
<category>
<pattern>DO                YOU      KNOW      WHO
                        *BUILT</pattern>

<template>
<srai>WHO BUILT <star/></srai>
</template>
</category>
</aiml>
```

The output given by machine is

Human: WHO BUILT QUTUB MINAR

Chatbot: Started by Qutub-ud-din Aibak and completed by his son-in-law Iltutmish.

The chatbot knows who built a particular chatbot because its name has an associated pattern. Hence it will respond to any query due to an authored pattern. This brute force mechanism is very difficult to put into practice since it needs thousands of patterns to be authored for chatbot to perform.

Algorithms

To make the classifier into a more manageable machine the work can be approached algorithmically that is we can have an equation for the same. The problem is reduced so as to get a simplified solution. "Multinomial Naive Bayes" is a classic text classification algorithm which gives the following equation:

```
class: weather
"is it nice outside?" "how is it outside?" "is the weather nice?"
class: greeting
"how are you?" "hello there" "how is it going?"
Classification of a few sample input sentences: input: "Hi
there"
term: "hi" (no matches)
term: "there" (class: greeting) classification: greeting (score=1)
input: "What's it like outside?"
term: "it" (class: weather (2), greeting) term: "outside (class:
weather (2) ) classification: weather (score=4)
```

For the classification of sentence "What's it like outside" into two Classes named weather and greeting, terms used in the input sentence are matched with two given classes. It was found that class weather scored 4 that is highest as compare to score of class greeting as 1 only. We skip the matching of every pattern in the input sentence and taken into consideration only the words that matched by the help of equation and generated the score on the basis of similar words only.

Figure 2: Naive Bayes Classifier

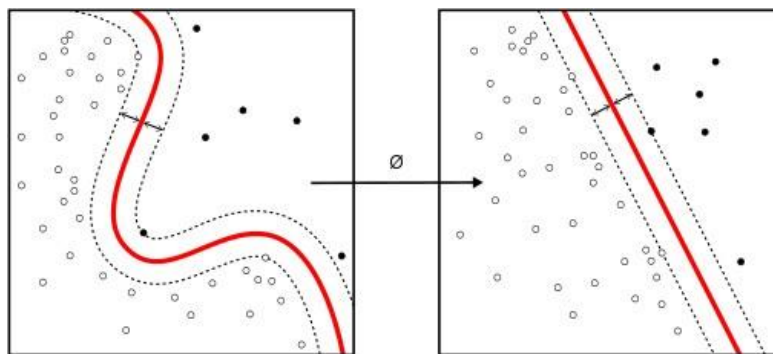


Figure 2: Naive Bayes Classifier

The class with the highest term matches is identified by the classification score produced. A score can tell which intent is more similar to the input sentence but never tells the possibility of a match. Thus, a threshold can not be applied to decide the acceptance of sentence on the basis of score given by the equation. A high score signifies only the more relativity keeping aside the strong classification. Or in other terms we can say that it provides an inherently weak classification. Also, the algorithm only counts what a sentence is like but it never count for what a sentence is not. Such type of approaches can not be considered to verdict about that a given sentence doesn't belongs to any specified class.

This classifier is "naive" because it assumes independence between "features", in this case: words. Each word is treated

A set of sentences that are belonging to a class will be provided along with a new input sentence. Then after a score will be generated for each class on the basis of occurrence of common words in each class. While matching, one of the important factors is matching the more meaningful words rather than less meaningful words. For example, match for the word "apple" will be more meaningful as compare to searching the word "it". A class with highest score gives the response that belongs to the input sentence.

A sample training set:

as having no connection with other words in the sentence being classified. Emphasis is given on the matching words but not on dataset training., it's "naive" but effective.

Neural Networks

Artificial neural networks (ANNs), invented in the 1940's provides the way of calculating output by giving some input (a classification) and applying adjusting weighted connections ("synapses") calculated from repetition of process through training data.

To improve the accuracy of neural networks, weights are adjusted as per the requirements of training datasets to create effective ANNs.

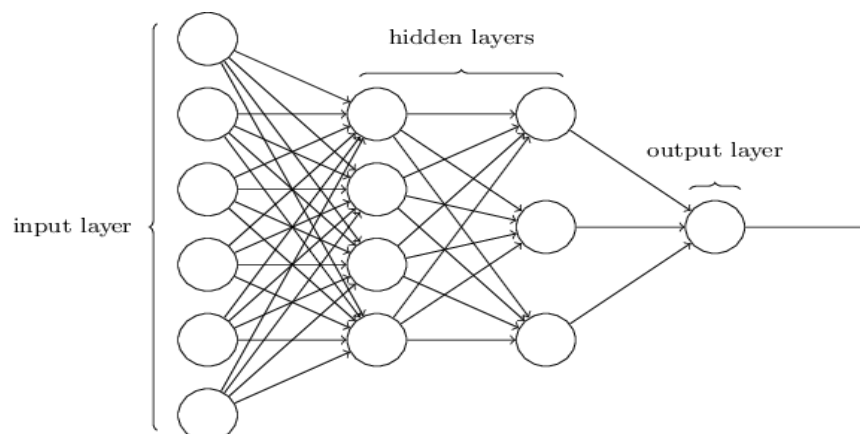


Figure 3: 2 hidden layers ANN with 6 inputs and one output: lines representing synapses and circles representing nodes

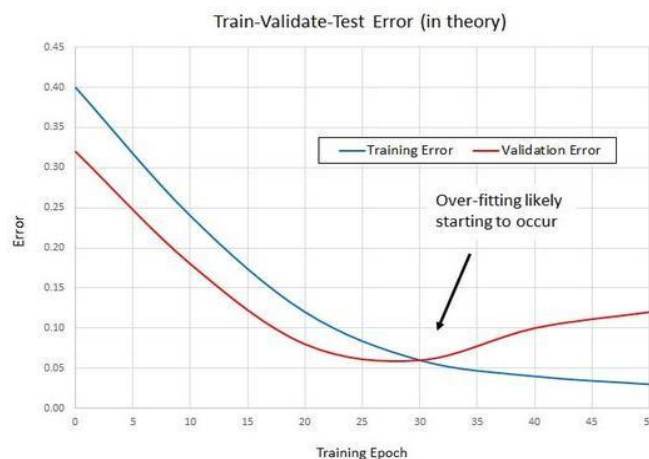


Figure 4: Train-Validate Test Error

While comparing a trained neural network with algorithm, an ANN is having less lines of code with respect to a comparable algorithm but ANN requires very large matrix to store the “weights”. In comparatively undersized sample, where only 140 unique words are provided in the training sentences and 25 classes are taken, a matrix of 140 x 25 will be required to store the weights. To find the product of such matrix and repeating the iterations 1,00,000 times so that error rate can be minimized, we need to have a very high speed processor.

As there are multiple pattern matching algorithms so there will be many more neural networks as well and they will be having varying capacity and complexity but the basic purpose of using ANNs will be classification only in this domain.

Several machine learning algorithms have been used to implement efficient chatbot systems. The Decision Tree (DT)

is one of the supervised learning method that is non-parametric and used for regression and classification. DTs are organized as trees that have hierarchical structures and are used to predict a value (basically taking a decision) on the basis of simple decision rules those are made up from data features. Capability of working with categorical as well as numerical data is one of the major advantages of DTs. Generally, DTs can be explained easily and facilitating human interpretation but DTs can lead to complex tree structure and thus cause problem of over fitting.

Yet another supervised learning approach is The Naive Bayes (NB) classifier that implements Bayes theorem along with the ‘naive’ assumption. Naive assumption says that there are independent features in context to a given class. Many of the applications in real world perform pretty well with Naïve Bayes classifier. Some of the applications are spam filtering, intent recognition and document classification. NB algorithms can be divided into two categories: Multinomial Naive Bayes (MNB) and Gaussian Naive Bayes (GNB). The distribution of continuous features is done in Gaussian for GNB algorithm, whereas, a multinomial distribution is assumed in the MNB algorithm for text classification.

The K-Nearest Neighbors (KNN) method is also very popular method that stores training data instances but never tries to construct a general internal model. This method uses the technique of voting and counts the majority for applying the nearest neighbors around the query point. A query point is decided and assigned as a class that will be most representative for all the nearest neighbors of that particular point.

In continuation, another classifier Deep Recurrent Neural Networks (RNNs) uses the sequential information for classification purpose. Traditional neural network assumes that all the input-output pairs are independent of each other. However, for many of the classification problems this assumption doesn’t seems be appropriate. RNNs are called “recurrent neural networks” because they perform the similar procedure for each element of a sequence, with the output taken as input in next step. Another good feature of RNNs is that they are having a persistent ‘memory’ that may be used to capture and store the critical information on temporally ordered events.[9]

The processing of query and providing the responses by a chatbot can be done using LSTM cells and deep learning[10]. Their study also showed information retrieval systems are outperformed by deep learning, in terms of quantitative evaluation metrics as well as qualitative human judgment.

The analysis of content of conversation using neural natural language processing techniques is done [11]. Another classification is based on GRU sentence analysis and the

Cosine similarity measure is used for estimating the similarity of the sentences [12,13].

4. Implementation

The proposed approach of an end-to-end system for tourists has been implemented technologies like Node.js, and Java along with various API(s) and libraries integrated into them. The Natural Language Processing and intention detection is performed using Google's Dialogflow platform. The following components were used/built to implement the proposed system:

The Chatbot

The chatbot here is essentially a Node.js web application. The complete workflow of the application is depicted in Figure.5 "Web Speech API" is used for speech-to-text and text-to- speech conversions. Its "SpeechRecognition" interface is used to listen to what the user says using the microphone. This is then fetched as a text string, which is immediately sent to the server using WebSocket. The server is setup and connected to Dialogflow using its Node.js SDK and hence the received string is fed as a query to the Dialogflow agent. This is followed intent detection and mapping to generate an appropriate text response which is transmitted back to the browser using the WebSocket. Now, the "SpeechSynthesis" interface of the "Web Speech API" is used to speak the response back to the user.

The server framework used to run Node.js is Express. The WebSocket is implemented via "Socket.IO". A WebSocket is used here for real time communication between the browser and the server. The Dialogflow agent is trained by creating intents based on the questions in the dataset (discussed below).



Fig 5: Chat Application Workflow

The Dataset

A dataset is a collection of related information organised into a defined format. The dataset required for this tourism domain chatbot was not available at the time of implementation. Hence, a Question-Answering(QA) dataset was created with 900 records containing important information about almost all monuments and famous tourist spots of India.

5. Results

The required dataset was created and populated with adequate entries. The Dialogflow agent was subsequently trained, as a result of which the chatbot can be seen responding to user queries for two different scenarios in Figure 6 and Figure 7 through the output log on the terminal.

```

Message: hey
Bot reply: Good day! What can I do for you today?
Message: tell me something about Gateway of India
Bot reply: A statue of Chhatrapati Shivaji was inaugurated on 26 January 1961, o
pposite the gateway as a symbol of Maratha pride and glory
Message: who designed Gateway of India
Bot reply: It was designed by the British architect, George Wittet.
Message: tell me something more about it
Bot reply: It has an Indo-Saracenic Revival architecture
Message: fun fact about Gateway of India
Bot reply: First Battalion of the Somerset Light Infantry passed through the Gat
eway on their way out in a ceremony on February 28, 1948, signalling the end of
British rule
  
```

Figure 6: Conversation scenario 1 (user querying about gateway of India)

```

Message: hello
Bot reply: Hello! How I can help you?
Message: what is the height of Qutub Minar
Bot reply: Its around 73 metres< 240 ft>.
Message: when was it constructed
Bot reply: The construction of its first storey started around 1192.
Message: who built it
Bot reply: Started by Qutub-ud-din Aibak and completed by his son-in-law Iltutmi
sh.
Message: thank you
Bot reply: Anytime. That's what I'm here for.
  
```

Figure 7: Conversation scenario 2 (user querying about Qutub Minar)

6. Conclusion And Future Scope

Although chatbots have gained a lot more recognition in the recent past through their presence in modern smartphones and personal gadgets in the form of Siri, Cortana, Alexa and many more, but we saw that they've been into development and use since long. Over the years, there have been improvements in terms of Natural Language Processing and hence intention detection, with the aim to simulate human conversation accurately. The proposed approach attempted to deliver an end-to-end system for tourists, foreign as well as domestic, visiting various monuments. In this attempt, Natural Language Processing was harnessed to give a digital tour guidance experience to the end user. The research process also led to the creation of an unavailable dataset covering almost all monuments of India. Most contemporary chatbot algorithms in use are retrieval-based and attempts to develop and implement efficient generative-based models are being made and looked forward to. Such algorithms would significantly reduce the need to extensively pre-define intents. This would also limit the frequent use of "Fallback intents" and hence generate much more reliable results. Apart from that, the chatbot built here can be extended to mobile platforms so users can carry their personal guide along to any place they visit. For that to be efficient, the dataset will need to be extensive. Geolocation tagging can also be used to make auto-detection possible and make the chatbot smarter.

References

- [1] J. Cahn. "CHATBOT: Architecture, design, and development." University of Pennsylvania School of Engineering and Applied Science Department of Computer and Information Science, Apr. 26, 2017.
- [2] J. Weizenbaum. "ELIZA—a computer program for the study of natural language communication between man and machine" Commun. ACM, 1966.
- [3] Richard S Wallace "The anatomy of ALICE. Parsing the Turing Test" 2009, 181–210.
- [4] Webber, G.M.: "Data Representation and Algorithms For Biomedical Informatics Applications", PhD thesis, Harvard University, 2005
- [5] A. S. Lokman, and J. M. Zain "An architectural design of Virtual Dietitian (ViDi) for diabetic patients"; pp. 408-411, 2009.
- [6] A. S. Lokman, and J. M. Zain, "Extension and prerequisite "An algorithm to enable relations between responses in chatbot technology" Journal of Computer Science, vol. 6, no. 10, pp. 1212, 2010.
- [7] A. S. Lokman, and J. M. Zain, "One-Match and All-Match Categories for Keywords Matching in Chatbot" American Journal of Applied Sciences, vol. 7, no. 10, pp. 1406, 2010..
- [8] Rokach, L., Maimom, O., "Data Mining with Decision Trees: Theory and Applications" World Scientific, Singapore, 2014.
- [9] Kamphaug, Å., Granmo, OC., Goodwin, M., & Zadorozhny, V.I., "Towards Open Domain Chatbots—A GRU Architecture for Data Driven Conversations" In: Diplaris S., Satsiou A., Følstad A., Vafopoulos M., Vilarinho T. (eds) Internet Science. INSCI 2017. Lecture Notes in Computer Science, Springer, vol 10750, November 22, 2017.
- [10] .Xu, A., Liu, Z., Guo, Y., Sinha, V., Akkiraju, R., "A new chatbot for customer service on social media", In: Proceedings of the CHI Conference on Human Factors in Computing Systems, 2017.
- [11] Lee, D., Oh, K.J., Choi, H.J., "The chatbot feels you - a counseling service using emotional response generation", In: 2017 IEEE International Conference on Big Data and Smart Computing, BigComp 2017, 2017.
- [12] Chung, J., Gulcehre, C., Cho, K., Bengio, Y., "Empirical evaluation of gated recurrent neural networks on sequence modeling", arXiv preprint arXiv:1412.3555, 2014.
- [13] Greff, K., Srivastava, R.K., Koutn'ik, J., Steunebrink, B.R., Schmidhuber, J., "LSTM: a search space Odyssey", IEEE Trans. Neural Netw. Learn. Syst. 28, 2222– 2232, 2017.

Wireless sensor network communication protocol: A Survey

¹Swati Juneja, ²Dr Karamjit Kaur, ³Dr Hardeep Singh

swati_juneja347@yahoo.com

kkaur@ggn.amity.edu

hardeep@thapar.edu

Abstract-A survey done on wireless sensor network (WSN) concluded that energy efficient routing based protocols are one of the basic energy saving methods for increasing network lifetime. It also prevents the conservation of energy. In today's wireless sensor networks, these two considerations are among the most significant issues. Wireless sensor network is a network that consists of a single base station, a sink node, and a number of sensor nodes that operate on very little battery power. The energy of these small nodes can be efficiently saved by designing various routing protocols. The different types of energy conservation/saving methods in a WSN are routing, node scheduling, clustering, etc. In this paper, we first define the various design challenges of routing protocol and then we will explain the existing routing protocols. As a result, this paper conducts a study of communication protocols.

1. Introduction

The wireless sensor network is basically a network which consists of a limited battery-powered, low cost, less duration, and small sensor nodes which are restricted in memory capacity and energy. These nodes are provided with the interfaces with which they can easily communicate with each other wirelessly in order to form a network. These sensor nodes are actually deployed randomly in a quite large geographical area. The primary objective of a WSN is to keep track of the target area for an extended period of time. The nodes are deployed in unapproachable areas where the replacement of sensor battery is usually not possible. The conservation of sensor capacity, as well as the optimization and extension of the network lifetime, is a major concern in WSN applications.

Each sensor node consists of basic components such as the sensor component, processing component, communication component and power supply component. These components perform together to provide a sensorial operation such as in tracking or monitoring of events in multiple areas, monitoring in the medical field, environmental monitoring, etc. as shown in figure (1). To figure out how much energy each node consumes, we must first figure out how much energy each part consumes [1]. A sensor and an analogue to digital converter make up the sensing portion (ADC). The unit is then used to sense the information data and collect all the valuable information. The processing component is employed to process the data, interprets the command to the ADC and then to send control signals. Power supply component consists of a power generator which is needed to supply power to the sensing unit, processing unit and communication unit. The communication unit is used to receive instructions and it, then, transmits the data from the processing unit to the user.

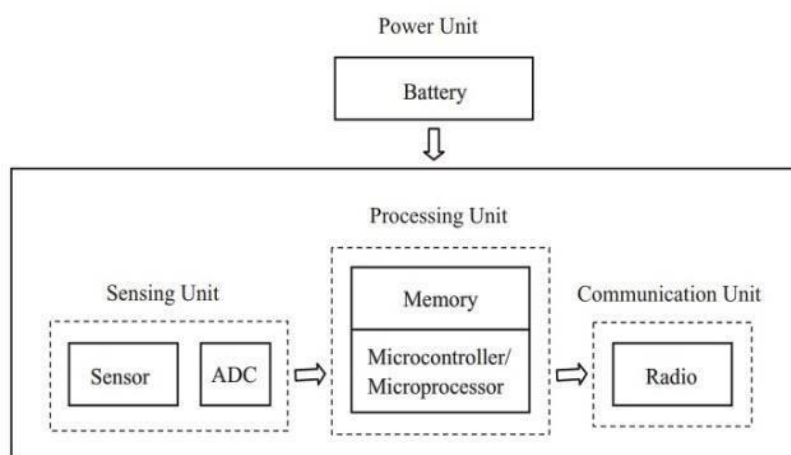


Fig 1: Wireless sensor node block diagram

2. Related Work

Along with prior knowledge, efforts are being made to plan and improve different routing schemes have been proposed. These protocols are configured and built based on the application requirements as well as the network architecture. However, when designing routing protocols for WSNs, certain considerations must be made. The energy usage of the sensors, which directly affects the network lifetime, is the most significant or crucial parameter for the architecture of any WSN.

The reviewer conducts a survey on architecture problems and strategies for WSNs in [2] (2002). At all layers of the network stack, the architectural restrictions of the sensor nodes, as well as the suggested protocols, were reconsidered. However, routing protocols are also not well classified.

In WSN (2004), a study on routing algorithms is presented in [3]. It divides routing protocols into three categories based on network structure: flat, hierarchical, and position-based routing protocols. Based on the protocol functioning, they are also subcategorized as multipath-based, query-based, negotiation-based, and QoS-based routing techniques. It contains a total of 27 protocols and discusses the problems and implementation concerns that must be addressed in WSNs. It also specifies some of the sensor node's drawbacks, such as restricted energy supply, processing capacity, and wireless sensor node bandwidth.

In this article, we extend Al-classification Karaki's to aid in the explanation of each protocol's activities and the improvement of energy efficiency concerns.

In [4], the author addresses several WSN routing protocols and divides them into three categories: data-centric, hierarchical, and location-based (2005) approaches. It just covered routing algorithms for sensor network and did not explain into insight on how to save energy.

In [5], the authors do a comprehensive review of emerging state-of-the-art protocols (2007). In wireless ad hoc networks, the authors characterize algorithms in two areas: minimal energy broadcast/multicast (MEB/MEM) and maximum lifetime broadcast/multicast (MLB/MLM) (AHN). As a result, the two key energy constituents that must be addressed are lowering the overall transmission power usage of all multicast session nodes and increasing the operating time until the first node's battery is exhausted. In addition, each node in the network has an omnidirectional antenna that is accountable for conveying and gathering data signals.

In (2008), the author discusses a top-down approach to a number of submissions and reviews on different facets of WSNs [6]. It divides the routing problem and conservation of energy into three different parts. However, the paper did not include a clear explanation of the protocol relation.

The author in 2009 [7] presents a report on energy efficiency by focusing on the physical part of a standard member nodes. The sensing device is categorized into four parts: one or two sensors for data collection, a microcontroller and memory for data acquisition, a radio component for wireless data transmission, and a power source.

The author also spoke about how to reduce power consumption in WSNs by taking into account the design and power breakdown. The methodology of energy saving schemes is then defined, along with its various features.

Furthermore, they show that the network's energy consumption is significantly higher than the energy consumption ascribed to samples or statistical analysis, owing to the employment of various forms of energy conservation. As a result, it is concluded that the sampling process needs more time than the contact phase, implying that the sensor's energy consumption can be very high.

The routing protocols are listed again in [8], along with their implementation problems (2009). Furthermore, different routing mechanisms are discussed in the context of their various features and processes for extending network life. The authors, on the other hand, lack a clear evaluation of multiple methods in question.

In [9], the paper emphasized the difficulties and problems in developing energy-efficient Medium Access Control (MAC) protocols for WSNs (2009). It also discussed the advantages and disadvantages of a few MAC

protocols for WSNs.

The authors of [10] conducted a review of a few existing methods, which is close to the exhaustive study carried by the authors of [12]. Just a handful of the protocols in the survey were energy sensitive, and they were all dependent on a specific sink node. In [13], the authors concentrated their research on clustering techniques and addressed them (2007).

Extensive research on biologically determined routing protocols has been conducted in [14] and [11]. The emphasis is on the concepts and implementations of different algorithms, as well as their modeling conditions, which are dependent on energy saving schemes. In [18], the authors looked at various traditional routing protocols as well as biological protocols in sensor network. The protocols are classified depending on a number of factors, including the information's complexity, the infrastructure, route selection, and the network's efficiency. Frequently selected from methodological and observational analyses, as well as certain flaws, were outlined in the classification of routing protocols, which served as a benchmark for future assessment.

In 2014, a survey [19] proposed strategies for allowing exchange amongst implementation criteria and energy consumption when designing WSNs. In [15], the author summarised prior studies and listed routing algorithms for sink node.

The authors explored and categorised energy-aware routing algorithms depending on different features of the sensor devices in their paper [16] (2016). In [17] the author presented a recent research of increasing network lifespan and lifespan strategies, identifying emerging developments in WSNs that included implementations, architecture limits, and lifetime model parameters.

3. Routing Protocols

Routing techniques are needed for transmitting information signals and information among sensors and base nodes in order to communicate. Because of many reasons, the routing process in sensor systems is extremely important and difficult. A routing protocol is a type of internet protocol that determines which outgoing path an arriving message can take.

It is, in several other terms, the method for evaluating an information transfer route between a transmitter node and a target node [3,4,5]. A sink device or a base station is the name given to the target node (BS). The following are the standard characteristics and restrictions of WSN routing:

- A sensor network is made up of several different sensing devices. As a consequence, employing a universal identifying address to reach each node may be impractical. Independent technologies can be needed by a WSN. As a result, the WSN's architecture should be primarily requirement-dependent.
- Where several wireless sensing devices feel the identical environmental conditions, a routing algorithm can minimise duplication of the sensed information.
- Because of their minimal price and battery-powered technology, sensor networks are more susceptible to error or loss.

In WSNs, a few of the primary goals of the routing algorithm is to save resources and reduce power usage. Some other network's routing algorithm, on the other hand, are developed to maximize a very good Quality of Service (QoS) across information transmission. The sensor devices have a number of drawbacks, including a low power source, a reduced memory capacity, and limited computing ability.

3(A) Design Issues of Routing Protocol

Although offering a wide range of uses, sensor networks have a number of flaws. Specifically, the mobile links linking the sensor nodes have low energy, power, and throughput. As a result, the main purpose of a sensing network is to produce knowledge from sensor nodes using sensing device while also extending the WSN's lifespan.

The restricted and limited amount of power availability of sensor nodes is crucial for the creation of energy-efficient routing algorithms. Routing in WSN is a very hard task as it is incredibly different from an ad-hoc network and mobile network. The following is a collection of the most important influences on routing protocol architecture [20]:

1. Node distribution is an utilization-based process that necessitates close monitoring of all resources because it impacts the routing algorithm performance. It may be either deterministic or random.
2. Utilization of energy without sacrificing precision: every node needs some amount of energy to perform the task. Thus, many algorithms are defined so that energy consumed by the node is conserved, but accuracy should be maintained.
3. Fault tolerances: Nodes fail because of less power, damages or environmental interference. When creating a protocol, it should be kept in mind that if a node breaks or a hole forms, it should not impact the network's efficiency.
4. Data Reporting Model: Protocols are profoundly concerned about energy consumption and route stability. It is entirely based on utilization.
5. Quality of Service: Information should be supplied over a set period of time. Even though, energy consumption, which is immediately correlated to network longevity, is sometimes considered more important than data quality.

3(B) Classifications Of Routing Protocols

Path establishment, Network Structure, and Protocol Operation are the three major types of routing protocols (figure 2). To define the protocols, they are divided into four groups.

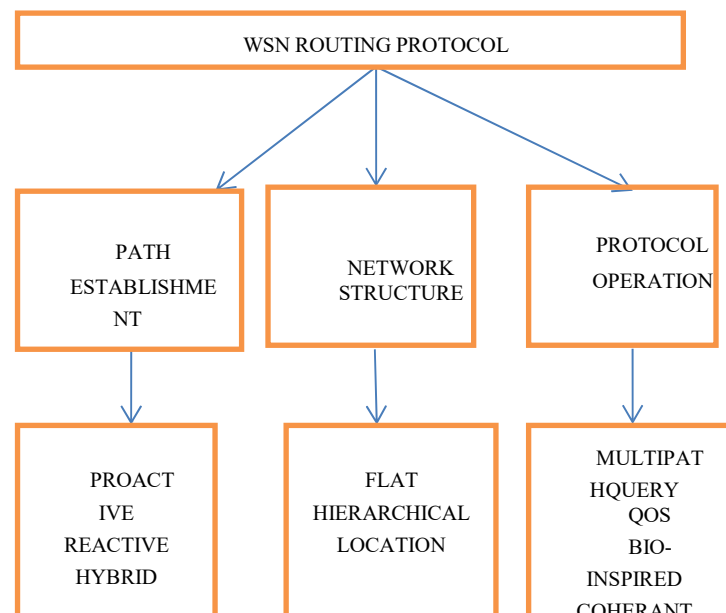


Figure 2: Routing Protocols

I. Path Establishment

These protocols differ in a number of ways. Despite belonging to all the same network infrastructure, they do not have the same features.

1. **Proactive or Table-Driven Routing:** Proactive(or table-driven) routing algorithms operate similarly to wired networks. It is based on the idea of sharing routing information across different nodes on a regular basis, with each node creating its own forwarding table. This table is important for determining a path to a specific location. Each and every sensing device must have one or more tables to hold all of the routing information. They still maintain every periodic network improvements and refresh their own table if any changes in the main network topology occur. As a result, if a path to a specific target or even a packet has to be forwarded, there is also no unnecessary delay due to route exploration.
2. **On-Demand Routing Protocols that are Reactive or Source-Initiated:** On-demand routing triggered by the source is an illustration of table-driven routing. When compared to proactive (table-driven) routing methods, reactive (on-demand) protocols only begin a route discovery process when it is required. A global search protocol method is required whenever a path from sources to target is required. It does not need frequent network updates. However, since the required routes are not visible and must be searched, it leads to delays. Any routes are indeed present in the cache retained by the sensor devices in some instances. As a result, there is no extra delay and no paths must be explored.
3. **Hybrid Routing Protocols:** They incorporate the benefits of both pro-active and re-active routing algorithms: they use pro-active routing locally and re-active routing inter-locally. It is founded on two major premises: 1) In WSNs, the majority of transmission occurs between nodes which are already near with one another, and 2) If a topology change takes place in the concentration of a particular node, it is to be expected. Whenever a route or a sensing device on the other portion of the network goes missing, this only affects the nearby neighbourhoods; devices on the other end are unaffected.

II. Network Structure

The node consistency is used to classify the network structure. Any networks have equally dispersed nodes which are of the same size. They distinguish amongst different nodes in certain networks. Flat, hierarchical, or location-based network structures are all possible.

All sensor devices in a flat network have the same degree of capabilities and tasks. They transmit messages to their neighbouring nodes regardless of the network architecture.

- i. **Flat Protocols:** The network's nodes are exactly similar and comparable. It has a number of benefits, including a low cost of infrastructure upgrades within network devices.
 - ii. **Hierarchical Protocols:** Sensor nodes are arranged in clusters in this category of protocols, with the head of the cluster usually being the node with the strongest residual energy. The cluster head is in charge of gathering data from the cluster's sensor devices, transmitting data amongst clusters, and passing information to the sink node. Clustering decreases energy demand and increases the network's lifespan. Clusters can be used to offset energy consumption since they have a high distribution ratio and scalability. The nodes closest to the base station or cluster head would exhaust all of their sources of energy first, followed by the other nodes.
- Location-based Protocols:** They use benefit of positioning information to redirect collected data signal and information to specific areas rather than the entire WSN. This family of protocols will discover a path from an origin to destination while conserving the energy of sensing devices.

III. Protocol Operation

1. **Query-Dependent Protocols:** In these protocols, the intended sensing device sends an information query over the network. When a node possesses this information, it sends the information that matches the query to the node that asked the inquiry in the first place [21]. Real vocabulary or high-level query vocabulary are commonly used to characterize these questions.
- **Coherent and Non-Coherent-Dependent Routing Protocols:** Manipulation of information is expected at the simple node level in sensing networks. Inside the network system, the sensing devices work together to analyse the information. It is suggested to use a routing method to start the data acquisition module. There are two types of processes in this system:

- Coherent Data Processing-Dependent Routing: This is an energy-saving process in which the sensing device only does the bare minimal of processing. The tasks of time stamping and duplication reduction are completed with minimal manipulation. The information is then transmitted to the collectors after minimal processing.
- Non-Coherent Data Processing Dependent protocol [89]: The information is analysed internally by the sensor devices in this category before it is sent to the next device for analysis. Sorting is done by aggregators, which are sensing device. In non-coherent navigation, there are three stages of information gathering.
 - a. Detecting targets, gathering data, and pre-processing (b) Membership announcement; (c) Central-node voting.
 2. An incident is discovered during the objective monitoring step the data is gathered and pre-processed. The sensor device decides to engage in a collective process and reveals this to all of its neighbours during the membership announcement step. A central device is expected to do more refined data processing during the central node election stage.
 3. Multipath Dependent Protocols: For wireless sensor networks, multi-path routing is an important routing process. Multi-path routing has the benefit of doing load balancing while still being more robust to routing faults. In contrast to single-path routing protocols, efficiency analyses of several multi-path routing protocols reveal that they take advantage of lower routing latency, less end latency, and congestion alleviation.
 4. QoS-Based Protocols: There must be a compromise between power demand and network data efficiency. When a sink generates information requests from the sensed node, a particular level of consistency must be retained.
 5. Bio-inspired: Biological systems vary from scheduled systems as the nature inspired networks have a normal cooperation motive, while scheduled networks which need certain efforts to rejuvenate sensing devices for cooperation. Randomness, multiple interactions, constructive feedback, and constructive criticism are the four core concepts that social insects use to self-organize. They mimic the social pattern of biological beings such as ants, bees, termites, spider monkeys, and bats to overcome critical challenges caused by the lack of centralised planning.

4 Conclusion

WSNs have grown significantly, and they now play a key role in the efficient collection and distribution of signal data. Energy saving is a utmost job for the sensor networks, as that have minimal battery life. Because of the difficulty and dependence of multiple operations on Sensor networks, energy-saving routing strategies and routing-based efficient standards are needed to ensure network access and information routing by using less resources.

Firstly, we discussed the various routing protocols and their design issues. We classify them into three categories, path establishment, network structure, and protocol operation. Furthermore, we looked at various survey articles on routing algorithms and protocols and discussed how they differed from our analysis. Since WSNs are mostly rechargeable batteries, the amount of energy usage of such various sensors must be controlled correctly and effectively.

References

- [1] Nikolaos A. Pantazis, Dimitrios D. Vergados, "A Survey on Power Control Issues in Wireless Sensor Networks," *IEEE Communications Surveys*, 2007, Vol. 9, Issue 4, pp. 86-107.
- [2] I.F. Akyildiz, W. Su, Y. Sankarasubramaniam, E. Cayirci, "Wireless Sensor Networks: A Survey," *Computer Networks*, 2002, Vol. 38, Issue 4, pp. 399-422.
- [3] Al-Karaki, A. Kamal, "Routing Techniques in Wireless Sensor networks: A Survey," *Security and Networks*, 2004, Vol. 11, Issue 6, pp. 6-28.
- [4] K. Akkaya, M. Younis, "A Survey on Routing Protocols for Wireless Sensor Networks

Role of IoT Based Self Automated Agriculture System

¹Dr. Sangeeta*, ²Ms. Preeti Dahiya*, ³Dr. Annu Dahiya*, ⁴Ms. Sakshi Wadhwa* and ⁵Ms. Ananya*

^{1,2,3}Assistant Professor, Department of Management Studies,

^{4,5}UG Student, Department of Business Studies

^{1,2,3,4,5} Panipat Institute of Engineering & Technology, Panipat, Samalkha, India.

¹sangeetaturk@gmail.com, ²preeti.sangwan@gmail.com, ³dahiya.annu05@gmail.com,

⁴swadhwa056@gmail.com, ⁵ananyagoyal017@gmail.com,

Abstract: In the past decades, agriculture has seen many transformation in technology. As the world is adopting internet of things (IOT), agriculture industry should also opt the IoT for crop management and improving productivity. Self automated IOT based agriculture system can be able to reduce the cost in production. The article tries to study the role of self automated agriculture system to improve the efficiency of agriculture sector. Agriculture is the backbone of developing country like India for the growth of the economy. So the paper aims to highlight the role of smart techniques in agriculture by combining with physical devices. The paper focus on the crop management, waste management, warehouse management with the help of remote smart devices. The paper suggested that farmers should use the smart devices for better production. Large amount of crops are required to reduce the hunger index in world. If smart and self automated systems are used in farming, the scenerio can be changed.

Keywords: Agriculture, Automation, IoT, Self Automation

1 Introduction

For the majority of the population in India, agriculture is the livelihood. Total foodgrain production within the country is estimated to be a record 291.95 million tonnes in 2019-20. according to the estimate of Indian Council for Agricultural Research (ICAR), demand for foodgrain would increase to 345 million tonnes by 2030. Although contribution of agriculture is reduced in the GDP of country but still agriculture has a good place in the growth of country [1]. Agriculture is the oldest human industry, but it's certainly no stranger to technological change. Agriculture is required to become technology oriented for better production. Smart farming and precision agriculture involve the mixing of advanced technologies into existing farming practices in order to increase the production efficiency and therefore the quality of agricultural products [2].

IoT is a combination of data, web associated items, is an integral component of the future Internet. IoT focuses on the automation of processes to reduce human intervention. IoT in agriculture focus is on automating the aspects of agricultural methods to make it more efficient and effective [3]. Farm automation, often related to "smart farming", is technology that creates farms more efficient and automates the crop or livestock production cycle. An increasing number of companies are performing on robotics innovation to develop drones, autonomous tractors, robotic harvesters, automatic watering, and seeding robots. Although these technologies are fairly new, the industry has seen an increasing number of traditional agriculture companies adopt farm automation into their processes [4]. The paper has four sections. First section covers the introduction of IoT and status of agriculture in India. In second section paper tries to explore the literature available regarding the self automated agriculture system. Section third covers the applications of IoT in agriculture and fourth section provides the suggestions for the implementation of IoT based self automated system.

2 Literature Review

Internet of things has changed the human lives and brings revolution in the all kind of industries. Agriculture industry is adopting the smart farming techniques for reshaping the future of agriculture. IoT based smart farming is not limited to the cultivation only but it has extended role in the water utilisation, crop management, preservation of quality crops also. IOT based smart farming helps in monitoring the farms more effectively as compared to conventional methods of farming [5]. Many kind of techniques are used in smart farming like artificial intelligence, big data and compute vision etc.

This is evident that due to global warming, there is dire need of conserving water at large scale. In farming, water is utilized at the fullest. Hence moisture sensors and temperature sensors are invented to preserve water. Wireless sensor networks made it simple to monitor the greenhouse parameters [6].

IoT has enabled the robotics based sensing system in order to develop the agriculture system. Self automated system is providing access to the crop data analysis and crop monitoring to the small farmers too so that they can also enhance the production. IoT have provided many solutions to the farm management and crop management system [7].

Researchers have focusd on various kinds of IOT based techniques to monitor the information regarding crop management. In agriculture sensors, robots are used to achieve the desired smart agriculture [8].Autonomus tractors and agriculture drones have transformed the agriculture sector into a new business model. These self automated influence the process and products of agriculture very effectively. Along with profitability and productivity, IoT also enables the farmers towards sustainble practices and farm management. Smart farming tools are very helpful in the enhancement of sustainable farming practices [9].

3 Research methodology

The paper is based on exploratory research design. The current study is primarily based on the secondary data engendered from the report of several ministries and organizatios. The article discusses the benefits of self automated agriculture system for improving the productivity.

4 Objective

Objective of the paper is to study the role of self automated agriculture system to improve the efficiency of agriculture sector.

5 Discussion

IoT based techniques used in farming

In the last decades, farming has seen many transformational changes. Technological changes takes place in agriculture which are helpful in controlling the practices of farming like livestock and crop management with the help of agriculture gadgets. IoT has improved the efficiency of the farming.Sourav Sarkar (2018) has highlighted the IoT based techniques with the help of following diagram



Fig.1 IoT based techniques

1. **Autonomous Tractors:** These kinds of tractors are controlled by remote devices. Autonomous tractors are pre-programmed according to the requirements of producers. Autonomous tractors are based on driverless technology. This will lead towards increased efficiency across operations and at last increase the production. Many firms have crafted tractor's autonomous kits so that farmers can access the automation more easily [10].
2. **Harvest Robots:** Robots are used in the agriculture for harvesting. Harvesting requires gentle treatment for fruits and vegetables. Harvest robots remain gentle on the fruits and vegetables and many technologies are innovated in order to evade staining and damage. Many Robotics manipulators work together on the improvement of quality of fruits and vegetables [11].

3. Robotics in Seeding and Weeding: In agriculture, robots are innovated for accelerating the seeding process. Seeding proves very effective on specific crops. Robotics used for seeding purpose reduces the cost of labour and speed up the whole process. Weeding is removal of unnecessary plants from the main crop. For labours it is a very complicated and time consuming task. Moreover, pesticides are also being used by the farmers in order to eliminate unwanted plants. Hence weeding robotics enables the farmers to make less use of labours and lessen the amount of pesticides [12].

4. Drones in Agriculture: Drones are innovated to keep an eye on the situations from a remote area. In agriculture, drones are used to keep a check on the farming conditions remotely. Drones are also appropriate for other kind of agriculture activities like application of fertilizers and pesticides. Agriculture drones are operated from a base station and analyse the issues very quickly and in a cost effective manner [13].

Benefits of the IoT based techniques

IoT based techniques are being used in agriculture for the progress of farming. Self automated agriculture systems are developed for crop management. In order to effectively utilise these self automated system, at first data related to smart farming is collected with the help of sensors. Sensors check the quality of soil, condition of weather and efficiency of equipments etc. IoT based techniques have better control over the agriculture processes and, as a result, lower production risks [14]. Farmers can foresee the output of their production so that they can make preparations for the circulation of production. If a farmer recognize exactly what proportion crops are getting to harvest, IoT techniques can be more useful. IoT based techniques are also helpful in the waste management because IoT devices have more control as compared to non-IoT devices.

The IoT magazine (June, 2019) has explained the benefits by using the following diagram-

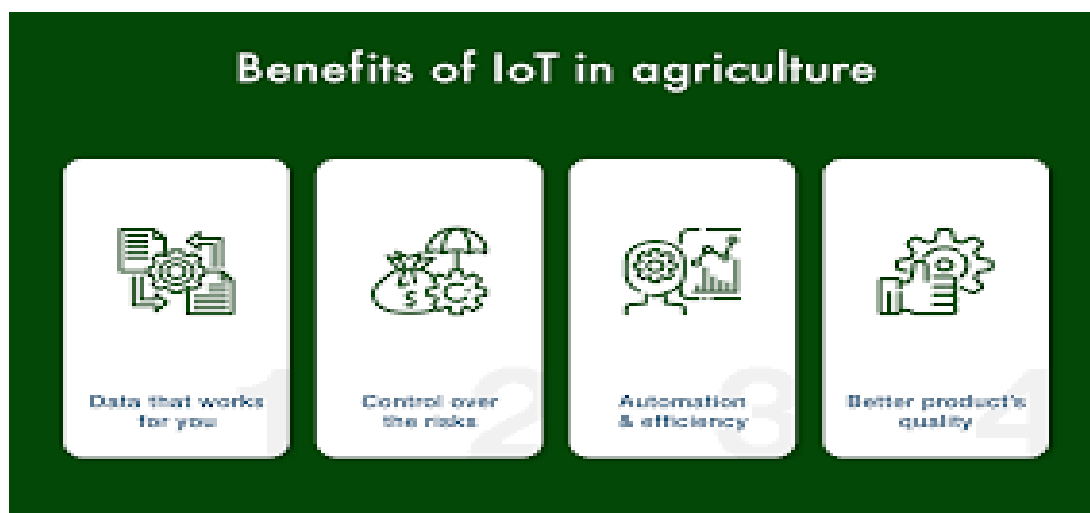


Fig.2 Benefits of the IoT based techniques

IoT devices are helpful in the management of various processes followed by production cycle. Moreover, self automated agriculture system has reduced the burden of routine tasks performed by the farmers. Agriculture industry always faces the shortage of labour which is taken care by the IoT based system [15].

6 Conclusion

The paper suggested that farmers should use the devices for better production. Large amount of crops are required to reduce the hunger index in world. Many kind of automation are being used in farming, still there are challenges which are not overcome yet. If smart and self automated systems are used in farming, the scenario can be changed to a great extent. High costs of IOT techniques are a big obstruction for the farmers of developing countries. Furthermore, high maintenance cost and breakdowns in machinery create issues in front of farmers.

References

- [1] Ahmed, N., De, D., & Hussain, I. (2018). Internet of Things (IoT) for smart precision agriculture and farming in rural areas. *IEEE Internet of Things Journal*, 5(6), 4890-4899.
- [2] Young (1989). *The Technical Writers Handbook*. MillValley, CA: University Science, 1989.
- [3] Athira.V , Arvind, Haripriya.H, Rani.R and Aravind.S (2017), Automated irrigation with advanced seed germination and pest control, in *IEEE in ICT for Agriculture and Rural Development (TIAR)*.
- [4] Joaquín Gutiérrez, Juan Francisco, Villa-Medina, Alejandra Nieto-Garibay, and Miguel Ángel Porta-Gándara "Automated Irrigation System Using a Wireless Sensor Network and GPRS Module",*IEEE*.
- [5] K. Swarna Krishnan, K. Jerusha, Poonam Tanwar, Shefali Singhal (2020). Self-Automated Agriculture System using IoT , *International Journal of Recent Technology and Engineering (IJRTE)* , ISSN: 2277-3878, Volume-8 Issue-6, March 2020
- [6] Farooq, M. S., Riaz, S., Abid, A., Abid, K., & Naeem, M. A. (2019). A Survey on the Role of IoT in Agriculture for the Implementation of Smart Farming. *IEEE Access*, 7, 156237-156271.
- [7] Y. Kim, R. Evans and W. Iversen, (2008). Remote Sensing and Control of an Irrigation System Using a Distributed Wireless Sensor Network, *IEEE Transactions on Instrumentation and Measurement*, pp. 1379–1387.
- [8] Kite-Powell, J. (2016).Take a look at how technology makes smart and sustainable farming, available at: www.forbes.com/sites/jenniferhicks/2016/12/31/take-a-look-at-how-technology-makes-smart-and-sustainable-farming/#262949cb3deb (accessed 16 August, 2021).
- [9] Walter, A., Finger, R., Huber, R. and Buchmann, N. (2017). Opinion: smart farming is key to developing sustainable agriculture. *Proceedings of the National Academy of Sciences*, Vol. 114 No. 24, pp. 6148-6150.
- [10] Q. Wang, A. Terzis and A. Szalay, A Novel (2010). Soil Measuring Wireless Sensor Network, *IEEE Transactions on Instrumentation and Measurement*, pp. 412–415, 2010
- [11] S.Vinothini (2020). A survey on the role of iot in agriculture for the implementation of smart. *Journal of University of Shanghai for Science and Technology*. ISSN: 1007-6735, Volume 22, Issue 11.
- [12] Shruthi B S, K B Manasa, Lakshmi R (2019). Survey on Challenges and Future Scope of IOT in Healthcare and Agriculture", *international journal on computer science and mobile computing* aliyavari.com/paers/2016_Sensors_SmartFarming/SmartFarm.pdf.
- [13] S. R. Nandurkar, V. R. Thool and R. C. Thool (2014). Design and Development of Precision Agriculture System Using Wireless Sensor Network, *IEEE International Conference on ACES*
- [14] Nóbrega, L., Gonçalves, P., Pedreiras, P., & Pereira, J. (2019). An IoT-based solution for intelligent farming. *Sensors*, 19(3), 603.
- [15] Prem Prakash Jayaraman et al (2016), Internet of things platform for smart farming: Experiences and lessons learnt", *Sensors* 2016, 16, 1884; doi:10.3390/s16111884.

Suggested Readings

- [1] S. Sagar, G. Kumar, L. Xavier, S. Sivakumar and R. Durai (2017). Smart irrigation system with flood avoidance technique, in *Third ICONSTEM*.
- [2] S. Saraf and D. Gawali (2017). IoT based smart irrigation monitoring and controlling system. in *IEEE International Conference on RTEICT*.
- [3] Tamilyanan.L and Venkatesan. R, (2017). A sustainable agricultural system using IoT," in *ICCSP*.
- [4] Zhao. W, Lin. S, Han. J, Xu. R and Hou. L (2017). Design and Implementation of Smart Irrigation System Based on LoRa, *IEEE GC Workshops*.
- [5] Sarkar(2018). Smart agriculture and irrigation monitoring system using IoT. https://www.researchgate.net/publication/327235058_Smart_Agriculture_and_Irrigation_Monitoring_System_using_IOT/figures?lo=1

OTT Streaming: Real Time Engagement

¹Dr. Annu Dahiya*, ²Mr. Vikas Nain*, ³Dr. Sangeeta*, and ⁴Ms. Riya Arora*

^{1,2,3}, Assistant Professor, Department of Management Studies,

⁴UG Student, Department of Business Studies

^{1,2,3,4} Panipat Institute of Engineering & Technology, Panipat, Samalkha, India.

¹dahiya.annu05@gmail.com, ²vikasnain1516@gmail.com, ³sangeetaturk@gmail.com,

⁴riyaarora82122@gmail.com,

Abstract: A shift from traditional media or entertainment industry to online streaming is witnessed. It bypasses all the traditional cable and satellite TV. These services give freedom to subscriber to enjoy content on their convenience. The study focusses on the various factors that are responsible for evolution and growth of these services. Relevant literature is studied to achieve the objectives of the study. Technological advancement, High speed internet facility (4G), drop in price of internet charges, access of digital devices to large population and idle time during lockdown period due to covid-19 are the factor that stimulates the growth of online streaming services.

Keywords: OTT, Technological advancement, Internet Facility, Covid-19

1 Introduction

The term “Over-the-top” refers to the services which provided freedom to the subscriber to use according to his convenience. It distributes content with the infrastructure that is not controlled by content provider. It bypasses all the traditional cable and satellite TV. These services are popular in developed countries but after rapid growth in technological advancement in telecommunication services also gain popularity in developing countries like India. Initially number of OTT apps provide content free of cost. Now OTT service provider offer premium services for that subscriber need to pay charge. In premium services various other services are provided in addition to video on demand.

Various factors like Technological advancement, High speed internet facility (4G), drop in price of internet charges, access of digital devices to large population and idle time during lockdown period due to covid-19, willingness to pay for OTT services stimulate the growth of OTT platforms. Global and Indian both brand shows strong position in Indian market. according to the annual report of M&E the Boston Consulting Group (BCG) along with the Confederation of Indian Industry (CII) during the period of covid-19 pandemic paid OTT services grow by 60% and expectation that at least half of new consumer will continue to subscribe these services in future.

2 Objectives of the Study

The study tries to explore the current trends and various factors that influence the evolution and growth of online streaming industry in India. So the main objective of the study is to hit upon the identification of different factors that encourage the growth of OTT industry.

3 Research Methodology

The framework mainly considers the factors that have significant impact on OTT industry. The focus of this article is on the service quality, ease of use, and subscription price. The relevant literature from respective journals and reports was studied which are most important source of study. The paper covers wide collection of academic literature on online streaming services which specifies different approaches and emphases of research in diverse disciplines.

4 Literature Review

Lee, Choia, Cho & Lee (2016) the availability of internet has great impact on various industries especially on procedure and structure of the industry. This paper focus on the impact of online streaming on music industry. Earlier with the use of internet sale of music industry goes down as all download songs from various websites and even share it with others but with the stabilization of online music platforms this industry gets reviving. To assess the impact of digital product on physical product a study has been conducted on adoption of online music streaming services in Korea. It is found that online music streaming services has positive impact on the sale of music industry. In addition to this it is also found that male artists work has more sale than female and logic behind this is female fan of male artist are more willing to pay for individual services. Online streaming services affects sales of both number of songs in one album as well as number of artists in one group positively.[1].

Matrix (2014)The technological upgradation in various fields also affect the entertainment industry especially television industry. It leads to change in program production decision, distribution deals and promotional strategies. This brings change in media habit of consumers also. Now consumer prefer over the top streaming service in comparison to old cable subscription. Consumer wants video on demand and do not want to wait for a predetermined television schedule. A large number of viewers using Netflix and other OTT services for watching TV shows and web series. This study examines the Netflix impact on young people's relationships, identities, their values, as well as their media use, habits, and literacies. A criticism is faced by this trend, it hampers the "water cooler talk" that create bonding among people through shared and mass-mediated experiences. A lot of quality kids' TV content also available on demand through these services. [2].

Kumari (2020)The trend of OTT services in the world makes a great change in various industries and entertainment industry is also one of them. India is not exception to this, it witnessed to various OTT services like hotstar, Amazon prime, Netflix. These services reorganise the Indian entertainment industry. OTT players get success in becoming the part of mainstream media and have influence on the consumption behaviour of the viewers. The relationship of internet with entertainment industry leads to change in mode of content distribution as well as nature of content creation. The objective of this paper is to study the trend of OTT services in India and factors responsible for the success of these services. Primary data is collected through online questionnaire and convenient sampling technique was used for this study. It is found that most of the respondents agreed that their movie and TV watching habits are change with the emergence of OTT services. They use these services at late night. Availability of good quality smartphones and affordable internet packs are the important contributing factors in success and growth of these services in India. [3].

Ota, Ray & Chandra (2020)A media service that is offered directly to viewers through internet is known as over the top media platform. In technical terms it bypasses the traditional cable, broadcast and satellite television platforms. The well-known OTT player in India is Netflix and hotstar. OTT has the biggest market in US and India is expected to second largest OTT market with over 500 million users in 2020. The aim of this paper is to analyse adoption of OTT platform and factors influencing the same. A structured questionnaire has been used to gather the data and descriptive statistical techniques are used to analyse the data. Results tell that hotstar is the preferable OTT service in Jamshedpur.[4].

Patel, Khadia & Awasya (2020)The era of technology and internet helps a lot during the period of lockdown for completing official task, enhance learning as well as spending free time. Trend of spending free time on digital screen is increasing tremendously during lock-down and great increase in viewership of the OTT video streaming services evidence for the above statement. Due to Shut down of cinema halls, prohibition on releasing of movie on big screen and technological shift by internet create a new kind of viewing experience. This new experience is created by OTT platform. This study tries to examine the impact of increasing use of OTT platforms on future of collective watching experience in India. It develops a different attitude of Indian viewers towards Big screen after getting new experience of viewing. It may create a threat for Big screen industry and collective watching experience.[5].

Jain (2021)Technological changes and internet have changed the structure and function of various industries and same with the entertainment industry. The operations of entertainment industry witnessed drastic changes due to increasing use of internet among target customer of this industry.This scenario of technological upgradation and increasing the awareness and technical literacy among consumer emerge the new trends in the entertainment industry such as use of different OTT platform. Amazon Prime, Netflix, Hotstar, Zee5, Disney+ are some popular OTT platform in India specially among salaried and youth customer.[6].

Lee, Nagpal, Ruane & Lim (2018) Digitalization and continuous expansion of digital media stimulate competition between cable TV and online streaming services. Advertising need to be coupled with customized product according to the demand of the customer is required to attract the customer. Various studies found that cost, ease of use and social trends are the key factor that determine the adoption of cable TV and online streaming services. Effect of various factors on adoption of Cable TV and online streaming services has been studied. Primary data has been collected through a structured questionnaire at a large public university. To assess the impact of various factor multivariate regression model was used. Model for both options showed statistically significant. The regression model found statistically significant for cost and customer service factors in cable TV option while only social trend was significant in regression model of online streaming services. Relation of these factors also studied with various demographic factors. It is found that there is no significant role played by gender in choosing online streaming over cable TV but age has significant role in adoption of online streaming services over traditional cable TV.[7].

Jose (2020) Rapid technological growth coupled with improvement in internet penetration across India leads to significant change in media habits of the people. One more factor that play a significant role in this change is emergence of over the top media services. In today's busy schedule a service allows you to watch movies, TV serials and other content at your own convenience. Majority of people in India have cable TV but subscription of various OTT platform like Netflix, Amazon Prime and hotstar also increased significantly. Some of these services add lot of other benefits and charge premium cost in comparison to cable TV. Users ready to pay premium cost for services like less break and advertising, freedom of time, availability of international content. These services also recommend contents to the consumer on the basis of their past streaming. The aim of this study is to better understand this new online streaming industry and various factor that contribute in shifting of media habit from traditional cable TV to online streaming services in India.[8].

Gupta & Singharia (2021) Technological advancements in various fields like telecommunication, availability of technological sound devices and a long lockdown period due to covid-19 pandemic create a war between online streaming service provider for attracting and retaining customers. This study examines the impact of customer engagement and quality of service experience on subscription of OTT service and willingness to continue these services in future. In other words, it studied the relationship between satisfaction and shifting in media habits. Partial least square structure equation modelling was used to analyse the relationship among various factors.[9].

Haridas& Sivapriya (2020) The concept of watching movie and other entertaining content is re-defined by the online streaming services like Netflix, Amazon Prime etc. these services gain a huge popularity day by day. The graph of subscriber goes very high in short span of time. There is an intense competition in India between OTT service provider. This paper studied the preference among customer with reference to Netflix and Amazon Prime. Various factors are studied that determine the preference of OTT services such as sentiment relating to content, audio-visual quality of video, ease of use and price of subscription. Primary data is collected through various social media platforms like Instagram, Facebook, Twitter, blogs, discussion forums, etc. Reviews available on various social media are analysed with the help of online text mining software. It was extracted that ease of use is the factor that significantly contribute to the quality of e-services.[10].

Anbumalar, Antony, Aiswarya.S, Mithun P.M & kumar (2021) A pandemic which changed the life of various people across the world and mostly negatively hit there life but this pandemic proves positive stimulator for the online streaming service industry. In India global players like Netflix and Amazon Prime grab good market share while Indian players like hotstar and Jio cinema also gain popularity in market. This paper examines the role of covid-19 pandemic in emergence of new media (OTT) consumption trend and also studied impact of this trend on people of various age such as child, adult and elder. Current trends and future aspects of these online streaming services were also studied. Both primary and secondary data was used to examine the above concepts.[11].

Gopi, Faizal & Vipin (2021) The pandemic of covid-19 badly hampers to most of the industries worldwide and theatre movie releases is one of these. When this time period became long then OTT platform become the choice for movie release. OTT platform supplies TV serial, movie and other entertaining content through high-pace internet connection instead of cable or satellite providers. This increased the amount of viewership on this platform. Viewers finds lot of benefits but this platform also have some negative impact. This paper aims to examine the negative impact of OTT movie releases on viewers. OTT movie releases has great impact

on Indian Big screen movie experiences. In addition to this a great impact on promotional activities, especially on big-ticket movies. The finding of the study is helpful for film industry to get an insight about attitude of Indian viewers towards OTT movie releases.[12].

Bellamkonda, Gangwar, Sudhagani & Adepu (2021) In India number of OTT service subscribers is continuously increasing as pattern of few years back is analysed. Technological advancement enables OTT providers to provide high quality content to near screen through internet. Lockdown also plays a significant role in growth rate of OTT subscribers. OTT provider make good effort and investment in producing content in addition to their depository. So increasing use of OTT services is resultant of various factors like technological advancement, access of digital devices to large population, drop in data charges, improvement in internet speed and prohibition on physical movement during lockdown period. This paper also extracted the information regarding user profiles and preferences from an Indian perspective.[13].

Chaithra (2020) In this century evolution of streaming apps becomes the biggest pivots in the entertainment industry specially among students. OTT platform is used for entertainment and academic purpose by youth. OTT apps become the most preferable option for spending leisure time. These apps change the culture of watching movies and TV serials now online streaming for watching entertaining videos become culture. This culture raises both the number of OTT subscriber and OTT providers. A good competition is seen among various online streaming app providers and they try to understand the psychology of people regarding the willingness of subscribing these apps. OTT providers use different strategies to attract potential subscribers.[14].

Pranami & Jain (2021) The extensive use of 4G, availability of affordable internet packs, local-oriented content makes OTT platforms more popular in India and expected to grow more in future (emergence of 5G). A great increase in online streaming service subscriber is record in a single year during covid-19 pandemic. Online streaming service providers like Amazon prime, hotstar and Netflix captured substantial part of market while Atal Balaji and voot also gain popularity. A survey is conducted with a sample size of 200 respondents to assess the increase the use of OTT platform, increase in time and perceived reason for this increase. A various reasons are found for increase in usage of OTT platform during Covid-19 such as citing various reasons such as 'Excess time', 'Boredom', 'New content' etc. The daily usage duration increased by 3 hours due to this pandemic. This fact make OTT platform a lucrative industry and attract number of new players in the market. Demand of these services in India resultant in launch of OTT service named Hulu and HBO Max. [15].

Jhala & Patadiya (2021) Some events, happening bring changes in different aspect of life. Pandemic of Covid-19 is one of those, it brings lot of changes in personal and professional life of the people. In this paper author studied impact of this pandemic on media consumption habit of Indian consumer. Various reports support that demand of OTT platforms increased during this pandemic. OTT is the service which directly offered to consumer through internet instead of cable TV and satellite. OTT platform now a commodity and no more considered as luxury. Both Indian (Hotstar and Jio cinema) and global players (Netflix and Amazon Prime) showed strong presence and tremendous growth in Indian market. It surpasses the cable, broadcast, and satellite television platforms. The value of this industry expected to reach to USD5.12 billion by 2026.[16].

5 Findings of the Study

This paper studied about online streaming services and existing literature provides good insight about it. Firstly, A media service that is offered directly to viewers through internet is known as over the top media platform. It bypasses the traditional cable TV and satellite TV. It provides freedom to watch content on his convenient time and device. OTT app providers make lot of effort in producing good content instead of relying on old depository. Structure of distribution as well as producing content in entertainment industry has been changed by the evolution of OTT platform. It changed the media consumption pattern of the consumer. OTT platform is more popular among youth and salaried people.

A tremendous growth was recorded in last few years by online streaming services. This growth attracted new players in the market. The launch of OTT app Hulu and HBO Max is the result of attractiveness of this industry. In India global players like Netflix and Amazon Prime grab good market share while Indian players

like hotstar and Jio cinema also gain popularity in market. This industry become threat for the Indian traditional TV serial industry and Big screen cinema. Now online streaming becomes the trend.

It is extracted that mobile is the most preferred device to use online streaming services.

Technological advancement, High speed internet facility (4G), drop in price of internet charges, access of digital devices to large population and idle time during lockdown period due to covid-19 are the factor that stimulates the growth of online streaming services.

A various reasons are found for increase in usage of OTT platform during Covid-19 such as citing various reasons such as 'Excess time', 'Boredom', 'New content' etc. The daily usage duration increased by 3 hours due to this pandemic. Customer engagement, quality of service experience, audio-visual quality of video and ease of use are the factors that determine the willingness for subscription of OTT apps.

Criticism faced by this trend, it hampers the "water cooler talk" that create bonding among people through shared and mass-mediated experiences and badly affects the collective watching experience.

References

- [1] Lee, M., Choi, H., Cho, D., & Lee, H. (2016). Cannibalizing or complementing?. The impact of online streaming services on music record sales. *Procedia Computer Science*, 91, 662-671.
- [2] Matrix, S. (2014). The Netflix effect: Teens, binge watching, and on-demand digital media trends. *Jeunesse: Young People, Texts, Cultures*, 6(1), 119-138.
- [3] Kumari, T. (2020). A study on growth of over the top (OTT) video services in India. *International Journal of Latest Research in Humanities and Social Science (IJLRHSS)*, 3(9), 68-73.
- [4] Ota, R., Ray, S.S., & Chandra, A. (2020). An analysis of customer preference towards OTT Platform during a pandemic: A special reference to Jamshedpur Market. *Science, Technology and Development* 9(8), 192-202.
- [5] Patel, M. K., Awasya, G., & Khadia, R. (2020). A Study: OTT Viewership in "Lockdown" and Viewer's Dynamic Watching Experience. *International Journal on Transformations of Media, Journalism & Mass Communication* (Online ISSN: 2581-3439), 5(2).
- [6] Jain, M. K. (2021). The rise of ott platform: changing consumer preferences. *EPRA International Journal of Multidisciplinary Research*, 7(6), 257-261.
- [7] Lee, C. C., Nagpal, P., Ruane, S. G., & Lim, H. S. (2018). Factors affecting online streaming subscriptions. *Communications of the IIMA*, 16(1), 2.
- [8] Jose, R. J. (2020). Factors influencing the shift from traditional TV to OTT Platforms in India. *International Journal of Advanced Science and Technology*, 29 (7s), 4044-4051. Retrieved from <http://sersc.org/journals/index.php/IJAST/article/view/22888>.
- [9] Gupta, G., & Singharia, K. (2021). Consumption of OTT Media Streaming in COVID-19 Lockdown: Insights from PLS Analysis. *Vision*, 25(1), 36-46.
- [10] Haridas, H. & Sivapriya, D. (2020). Customer Perception towards Networked Streaming Service Providers with Reference to Amazon Prime and Netflix. *International Journal of Recent Technology and Engineering* 9(1) 513-517.
- [11] Anbumalar, S., Antony, B., Aiswarya, S., Mithun P.M & Kumar, A. (2021) Adoption of OTT Platform in India During Covid-19. *Compliance Engineering Journal* 12(4), 221-231.
- [12] Gopi, J., Faizal, M., & Vipin, H. (2021). An Analysis of the Impact of Covid-19 on OTT Movie Releases and Its Viewers in India. *Asian Basic and Applied Research Journal*, 1-8.
- [13] Bellamkonda, S. T., Gangwar, V. P., Sudhagoni, V. S., & Adepu, N. (2021). profiles and preferences of ott users in indian perspective. *euopean journal of molecular & Clinical Medicine*, 7(8), 5106-5142.
- [14] Chaithra & Reshma (2020). Proliferation of OTT apps in India: an empirical study of OTT apps and its impact on college students. *International Journal of Research and Analytical Reviews* 7(1), 427-435.
- [15] Parnami, S., & Jain, T. (2021). A Study on Increase in the Usage of OTT Streaming Services. *International Journal of Research in Engineering, Science and Management*, 4(8), 142-145.
- [16] Jhala, B. D., & Patadiya, V. B. (2021). a study on consumer behaviour towards ott platforms in india during covid era.

Integration of AI and IoT in Supply Chain Management of Indian SMEs

Siddharth Sahu¹ and Anchal Gupta²

¹ Amazon Development Center, Bangalore, India; siddharthsahu.asm@gmail.com

² Apeeyaj School of Management, Dwarka, New Delhi, India;
anchal.gupta1503@gmail.com

Abstract: SMEs are considered as the backbone of Indian manufacturing sector and one of the core functions of SME is supply chain, which creates a connection between consumers, suppliers and other intermediate organizations (including large enterprises). The 4th Industrial revolution brings a lot of new and advanced technologies, which the SMEs have found difficult to cope up with. This paper aims to identify the advantages of these emerging technologies, especially Artificial Intelligence (AI) and Internet of Things (IoT) in the supply chain domain of Indian SMEs. Furthermore, there has been an attempt to study the barriers which the SMEs are facing or might face in the future, while implementing these technologies. A list of 27 benefits and 15 barriers has been derived from studying various literatures, which are further analyzed using mathematical models like AHP and ISM. The logistics has been found to be the most promising area, followed by production and procurement. On the other hand, it has been observed that many SMEs are not even aware about Industry 4.0 and also they lack basic IT facility to accommodate these advancements. There is a need for the SMEs to tackle these challenges and take a step towards industry 4.0, where automation will be the key to manufacturing.

Keywords: Artificial Intelligence, Internet of Things, Supply Chain Management, Analytical Hierarchy Process (AHP), Interpretive Structural Modelling (ISM), Barriers

1. Introduction

Supply Chain Management (SCM) is a very crucial vertical in every organization which includes many aspects like supplier selection, transportation and logistics, retail, communicating information from one node of the chain to another, etc. Every node in the supply chain has its own importance and managing all of them together, can be a complex and difficult task. Since the increase of globalization, companies are getting inter-connected and rely on each other for various reasons like raw material, manufacturing equipment, various services etc. Indian SMEs and their Supply Chain play a very important role in fulfilling the requirements of large manufacturers like Toyota, Ford, Delphi, Fiat, Volvo etc., who have entered the Indian market for production and prefer to get the complete assembly of a part, rather than having a large number of individual parts (Thakkar et al., 2012). The quality of product and delivery process can be improved by focusing on SCM of these companies (SMEs), resulting in more reliable delivery promises, faster delivery, risk reduction etc. However, studies show that SMEs are unable to harness the true potential of SCM (Thakkar et al., 2011).

Information technology (IT) plays a vital role in enhancing the operational efficiency of SCM and improves infrastructural stability by integrating SC partners with inter-organizational functions, which will result in improving the competitiveness of the SMEs (Tripathy et al., 2016). Considering the industry 4.0 which is mainly data driven, Artificial Intelligence (AI) and Internet of Things (IoT) have emerged as a cutting edge technology, which can help SMEs to solve the problems of SCM, giving them a competitive advantage over others.

The AI is completely data driven, and the algorithms used are as good as the data available with the system. Data has always been available around us, what we were lacking was the technology to gather this information and analyze it to get some meaningful insight. These analysis can further be used to take intelligent decision (Ben-Daya et al., 2017). IoT can play a vital role in gathering and communicating information among different SC partners, and AI can be used to analyze and process this information to take intelligent decision.

Integrating AI and IoT with the SCM of SMEs will not be an easy task as there are many requirements for this system to work efficiently. Many of the SMEs don't even have the basic computer facility and even if they did, their IT capabilities are limited to basic administrative work (Thakkar et al., 2012).

SMEs need to coordinate with different supply chain partners, at different levels to gain the competitive advantage. To deal with uncertainty in the market for demand, SMEs require speed in decision making, which needs to be accurate and intelligent (Singh, 2011). Since, there is a need for better manufacturing ecosystem as more companies are entering India for production, there is a need to improve the SMEs of Indian manufacturing sector. This paper is an attempt to study this need of improving the SMEs from a supply chain point of view, and what role AI and IoT can play to strengthen the supply chain.

To make an approach towards the integration of AI and IoT in SCM of Indian SMEs, there is a need to understand the advantages of this integration, and also an awareness is required about the barriers and challenges for the implementation of AI and IoT. Hence the key objectives of this study are as follows:

- Identify the benefits of AI and IoT in SCM of Indian manufacturing SMEs
- Identify the challenges the Indian manufacturing SMEs may face, in implementing AI and IoT in their SC.
- Rank the benefits using AHP method and derive meaningful insights.
- Develop a model to derive relationships between different barriers using ISM approach.
- Categorise the barriers based on driver and dependence power using MICMAC analysis.

2. Literature Review

The study involves three main aspects, Supply Chain Management, Artificial Intelligence and Internet of Things. AI and IoT have been studied with respect to their application on SCM.

2.1. Supply Chain Management: Current scenario of SMEs and issues involved

SMEs are neither much concerned with the activities of SCM, nor able to fulfil the need of effective and efficient planning of their supply chain activity which could lead to better product quality, cost optimization, integration with other SC partners, and implementation of e-based solution (Vaaland and Heide, 2007; Thakkar et al., 2012). Thakkar et al. (2011) also discussed that a key factor in the survival of SMEs is to produce more in a less time and cost, while taking care of the quality, which can be attained by efficient SC practices. In an attempt to be a part of global business, flexibility in changing needs and deciding prices, remains a challenge for the SMEs. SMEs need to adapt more IT tools and practices for better responsiveness, fast and intelligent decision making and to cope with the external competition (Singh, 2011; Tripathy et al., 2016). Since the SMEs operating parameters are very different from that of the large enterprises, it is difficult to evaluate the effect of integration of IT services. Focusing of SC practice can lead to many benefits for the SMEs like - faster delivery, reliability in delivery promises, ability to tackle delivery disruptions, cost optimization, ability to handle risks and optimized inventory management (Thakkar et al., 2011).

Adapting SC practices is costly, which is one of the reasons that SMEs are reluctant to implement SCM (Thakkar et al., 2012). It is also said that SCM may also have a negative effect on the SME's performance (Vaaland and Heide, 2007). There are other reasons which make it difficult for an SME to implement SC practices - like they have to follow rules and regulations as per the larger customers, large companies believe that SMEs can be easily replaced and hence buyers hesitate to partner up with SMEs, SMEs lag behind in adequate tools, systems and method like electronic interface and IT tools and they also have limited resources. Apart from these challenges, Thakkar et al. (2012) explained that Indian organizations, on an average, carry a total of 33.41 days of raw material, 14.25 days of WIP and 16.09 days of finished goods, which indicates that SMEs have a poor infrastructure for logistics.

2.2. The need of Information Technology: Advanced computing and Analysis

To study the application of IoT and AI in SCM, first the understanding of these emerging technology was necessary. For this various literature was referred, which helped in understanding the architecture of IoT system and how these systems worked (Chun et al, 2019). The same approach was used to understand AI and the different components and techniques of AI. These fields of technology are very vast and studying every aspect in detail seemed irrelevant to the objective of this research. Hence, a brief study has been done in order to get the basic understanding of AI and IoT. Another purpose of studying the literature on IoT and AI was to find the area of application of these technologies, and how it benefits different domains of SC. Also, an attempt was made to discover different challenges or requirements which has to be fulfilled, in order to make the integration of these three areas of studies i.e. SCM, IoT and AI, successful.

Advanced tools like agent based systems (a subset of AI) can prove to be very useful for coping up the fast changing and dynamic environment, by improving coordination and managing complex SC networks (Chun et al., 2019; Chiu & Lin, 2004). SMEs are trying to extract information from every asset they have like inventory, transportations and logistics, warehouses etc. and the firms are becoming more data oriented (Min, 2010). Automatic collection and analysis of data to make better decision is required for the I4.0 (Calatayud et al., 2019). For the SME have a competitive advantage, smart technologies (smart production, smart transportation and logistics etc.) are very important (Witkowski, 2017). Better communication and collaboration between SC partners, better responsiveness, faster processes and better decision making is essential for the survival of SMEs, in a data-driven world (Thakkar et. al., 2012; Tripathy et. al., 2016; Singh, 2011; Gupta et al., 2020).

The common definition retrieved from former literature, Internet of Things (IoT) is a technology which collects, transmits, and analyze data. IoT enablers play a crucial role in collecting data from different sources which will improve the ability to retrieve data. According to a number of literature studied, a similar basic architecture of IoT application has been observed, consisting of four layers (Calatayud et al., 2019; Da Xu et al., 2014, Abdel-Basset et al., 2018; Verdouw et al., 2016). Studying different literature on the application of IoT in SCM, three main domains of SCM has been observed, which has the potential to gain some benefits from the application of IoT – Demand and order management, Production and Manufacturing, and Logistics and transportation. Apart from these fields, a separate field of Information and analysis has been identified which helps the three fields in improving their functioning.

As explained in Oxford English Dictionary (2008), Artificial Intelligence is the theory and development of computer systems able to perform tasks normally requiring human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages. Dash et al. (2019) defined AI as the ability of machines to address problems independently, for which they have not been programmed explicitly. Simply put, the main function of AI is to understand and mimic the human intelligence and behavior, to design and develop computer systems which can solve problems without human assistance. (Min, 2010).

The list of application for both AI and IoT, a combined lists for benefits has been derived on which further mathematical modelling will be applied is shown in Table 1.

Table 5. List of combined benefits of AI and IoT in SCM

Classification	Benefits	Description	Related Field	Reference
Procurement	Supplier selection	Evaluating supplier performance and selection	Rough set theory, expert learning, GA, Fuzzy logic, IoT	Min, 2010; Abdel-Basset et al., 2018
	Demand Forecasting	Increased efficiency in forecasting demand	ML, Expert system, ABS, IoT	Min, 2010; Calatayud et al. 2019; Dash et al., 2019;
	e-Procurement	Effectiveness in bidding and placing order online	ABS	Min, 2010
	Order Management	Real time communication between customer and logistics partner	IoT	Govindan and Hasanagic, 2018; Mangla et al., 2017
	Make or buy decision	Deciding whether to outsource job or to do it in-house	AI	Min, 2010
Production and Manufacturing	Production planning	Improved production planning, better shop floor control, minimizing waste etc.	ABS, IoT	Min, 2010; Dash et al., 2019; Mangla et al., 2017; Witkowski, 2017 ; Abdel-Basset et. al., 2018;
	Improved Quality	Error free production	IoT, AI	Dash et al., 2019;

	Automation	Robotic automation and smart manufacturing	IoT, AI	Dash et al., 2019; Cardoso & de Sousa, 2019; Radanliev et. al., 2019; Mittal et al., 2017
	Improved R&D	Helping in R&D and NPD	ABS	Min, 2010; Dash et al., 2019; Plastino & Purdy, 2018
	Asset Management	Resource optimization, maintenance and early detection of breakdown	Prediction Algorithms	Calatayud et al., 2019; Plastino & Purdy, 2018
Logistics and Transportation	Logistics planning	Distribution problems, delivery and pickup, TSP, and e-logistics	IoT, AI	Min, 2010; Cardoso & de Sousa, 2019; Calatayud et al., 2019
	Vehicle routing and maneuvering	Improved route planning, assisted driving, traffic analysis	ANN, GA, IoT	Min, 2010; Calatayud et al., 2019; Witkowski, 2017
	Inventory Management	Improved Material handling, inventory control, order picking etc.	ANN, GA, Expert systems, Fuzzy Logic, ABS, IoT	Min, 2010; Cardoso & de Sousa, 2019; Calatayud et al., 2019; Abdel-Basset et. al 2018
	Order Fulfilment	Improved Order fulfilment	ANN, Fuzzy logic	Min, 2010
	Location Allocation	Facility location optimization	AI	Min, 2010; Calatayud et al., 2019
	Resource Optimization	Optimization of logistics resources	IoT	Govindan and Hasanagic, 2018; Da et. al., 2014
	Container Load Planning	Loading the container in an efficient way	AI	Min, 2010; Calatayud et al., 2019
Communication and Information Exchange	Communication and visibility	Better communication and information exchange	ANN, ABS, IoT	Min, 2010; Plastino & Purdy, 2018; Witkowski, 2017; Calatayud et al., 2019
	SC Partnership	Improved communication between SC partners, assistance in negotiation	ML, ABS	Gupta et al., 2020; Min, 2010
	Early Warning and control system	Identifying potential problems and suggesting control measures	IoT, AI	Govindan and Hasanagic, 2018; Radanliev et al., 2019; Verdouw et al., 2016
	Real Time Monitoring	Reduced communication gaps, real time availability of data	IoT, AI	Min, 2010; Dash et al., 2019; Plastino & Purdy, 2018; Witkowski, 2017; Verdouw et al., 2016; Abdel-Basset et. al 2018; Calatayud et al., 2019
Other Benefits	Strategic Advantage	Firms with better technology will definitely have an upper hand	AI	Dash et al., 2019
	Big data analysis	Ability to analyze large complex data	AI	Cardoso & de Sousa, 2019; Calatayud et al., 2019; Plastino & Purdy, 2018

		from different sources		
	SC Planning	Better SC strategies and planning	ANN, Rough set theory, ABS	Min, 2010
	Resource utilization	Optimization in resource allocation and utilization	AI	Plastino & Purdy, 2018; Mittal et al., 2017; Gupta et al., 2018
	Cost Reduction	Reduction of cost in many fields	AI	Dash et al., 2019; Plastino & Purdy, 2018
	Data Analysis	Analyzing data obtained from IoT enablers to get valuable insights	IoT, AI	Govindan and Hasanagic, 2018; Calatayud et al., 2019
	CRM	Recognizing customer requirements and decision making	AI	Min, 2010; Dash et al., 2019

2.3 Challenges for the application of IoT and AI in SCM in Indian SME

Implementing and adapting IoT and AI in Indian SME can be very challenging, given the current situation of the SME sector in India and the prerequisite for the IoT architecture and AI implementation. From the literature studied, the combined list of barriers for implementation of AI and IoT has been shown in Table 2.

Table 6. List of combined barriers for AI and IoT in SCM

Barrier No.	Barriers	Description	References
1	IT infrastructure	Limited use of computers and IT tools like RFID, sensors, Scanners etc.	Mittal et al., 2017; Suresh et.al., 2018; Chun et al., 2019; Radanliev et.al., 2019; Gupta et al., 2020
2	Facility designs	Need to change the layouts, network design and facility design as per the implementation of AI and IoT	Dash et.al., 2019; Da Xu et.al., 2014
3	Information source	Limited information available online as many suppliers are not connected to the internet and information about the supplies are not updated	Yuan et.al., 2001; Duan et.al., 2005
4	Data storage platform	Need of better data storage like cloud platform services	Suresh et.al., 2018; Radanliev et.al., 2019
5	High IT skills required	AI and IoT need high level of IT and programming skills	Govindan and Hasanagic, 2018; Min, 2010
6	Cyber security	Vulnerability to cyber-attacks	Suresh et.al., 2018; Da Xu et.al., 2014; Radanliev et. al., 2019
7	Lack of Overall control	Many companies and parties are involved in the entire supply chain and hence there is a need of an entity to supervise the entire supply chain	Yuan et.al., 2001
8	High speed connectivity	The flow of information has to be uninterrupted and fast	Govindan and Hasanagic, 2018; Suresh et.al., 2018
9	New business models	Need to update the business models as there will be a major influence of AI	Gupta et al., 2020; Dash et.al., 2019
10	Advancement in HR	HR policies and strategies has to be changed considering the virtual labor- AI	Plastino & Purdy, 2018; Suresh et al., 2018
11	Understanding of Decision makers	The decision makers and top management needs to understand the working of AI systems, it's benefits and limitations	Min, 2010; Plastino & Purdy, 2018

12	Awareness and willingness	SMEs need to be aware about the AI and IoT and must be willing to implement them in their companies	Suresh et.al., 2018
13	Trust and delegation	Appropriate task has to be delegated to AI and the decision made by AI has to be reviewed as per the importance of the decision and as per how good the AI system is.	Yuan et.al., 2001
14	Ontology	The technical terms and jargons used in the entire SC must be same to reduce ambiguity for the system	Yuan et.al., 2001; Duan et.al., 2005; Da Xu et.al., 2014
15	Horizontal and vertical integration	For better flow of information, different departments at different levels of organization must be connected	Suresh et.al., 2018

3. Research Methodology

To achieve the objectives, the benefits from IoT and AI have been then classified into different domains of SC. The mathematical models have been applied to the identified benefits and challenges to derive appropriate conclusions. For the benefits, analytical hierarchy process (AHP) technique has been adopted to find out, which field of SC will be benefitted the most. For the challenges, interpretive structural modelling has been used to find out which is the most important challenge, which the companies have to tackle first in order to make the integration easier.

3.1. AHP Methodology

AHP uses numerical judgements to make relative scales between different factors, by using pairwise comparison matrices. The different factors are arranged in a hierarchical structure, with the objective on the top, different layers of criteria below it, and at the bottom, different alternatives. The decision makers make judgements by comparing the criteria and alternatives in pairs and derive a matrix, called pair-wise comparison matrix (Satty, 1987).

AHP is considered to be an effective multi-criteria decision making tool and has been used in many research papers like Precin (2008) used fuzzy AHP to evaluate the benefits of information sharing in supply chain and logistics management. The steps involved in the analysis are shown in Figure 1:

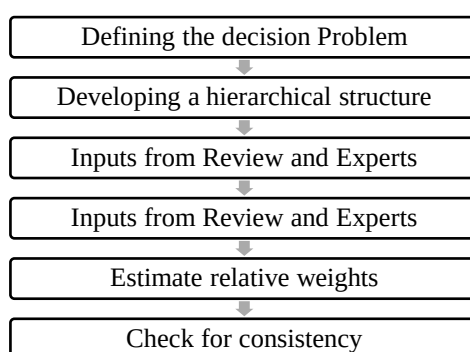


Fig. 21. Steps of AHP Methodology (Satty, 1987)

3.2. ISM Methodology

First proposed by J. Warfield (1974), Interpretive Structural Modelling (ISM) is a method to analyze complex relations among different elements of a system. It is an interactive learning process which transforms a poorly constructed model of system into a well-defined structural model, using the judgement of human being and identifying the relations between the elements of the system. The newly formed model depicts the structure

of the problem in a well-defined pattern including graphical representation. (Raj et al., 2008; Singh et al., 2008)

ISM has been used in many literature to analyze barriers. Raj et.al. (2008) used ISM to analyze the mutual interaction of key enablers in the implementation flexible manufacturing system. Gupta et al. (2018) studied the challenges of Indian logistics service provider using ISM model and, used the ISM model to analyze the interaction of different challenges in implementation of e-governance (Gupta et al., 2019). Luthra et.al. (2011) presented a study of challenges to implement green supply chain management in Indian automobile industries. As explained in various literature (Raj et al., 2008; Singh et al., 2008; Luthra et.al., 2011; Gupta et al., 2019), the steps involved in this technique as shown in Figure 2.

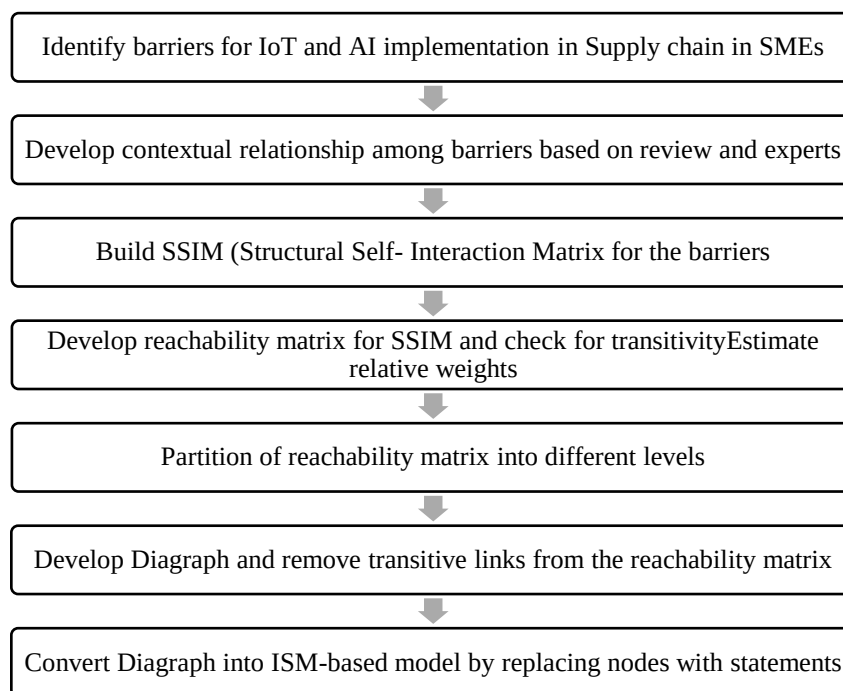


Fig. 22. Steps of ISM methodology Source: Warfield (1974).

4. Analysis and Discussion

4.1. AHP analysis for the benefits of AI and IoT

AHP is applied on level-1 and the principal vector so formed will depict the local weight of level-1 criteria. The AHP analysis is then applied on all the sub-criteria (level 2 criteria) i.e. all the criteria for each of the level-1 criteria. The principal vector of each of the level-2 matrix will give the local weight for level-2. A total of 6 pair-wise comparison matrices is formed (1 matrix for level-1 and 5 matrices for level-2) after taking inputs from experts. Four experts from manufacturing and online retail sector with rich experience working in supply chain department have been considered for the study. Multiplying the local weight of level-1 with the local weight of level-2, will give the total weight of the sub-criteria (Figure 3). The results of AHP calculations are shown in Table 3 and the local and global weights of all benefits are discussed in Table 5.

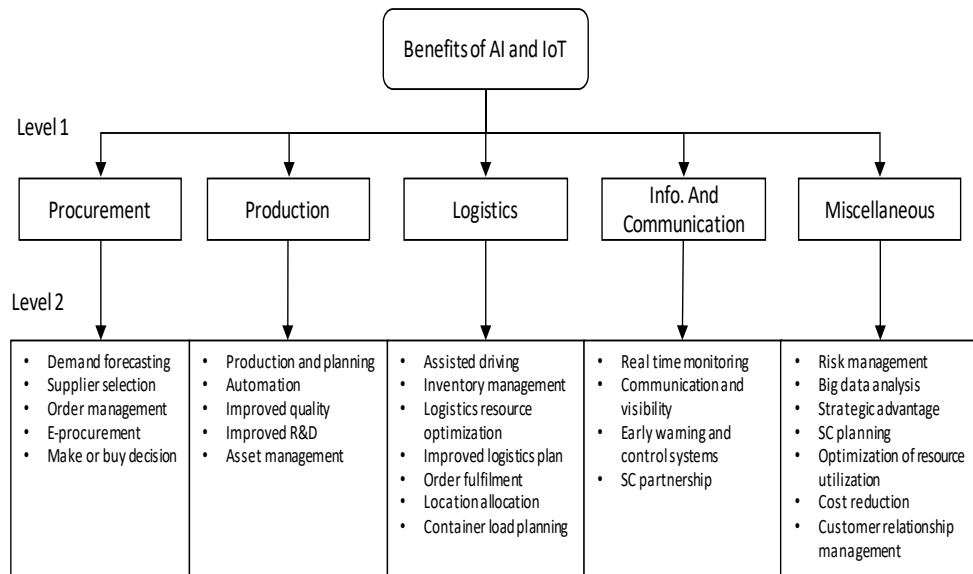


Fig. 23. AHP level classification for benefits

Table 7. Pair-wise comparison matrix for Level-1 Criteria

Level 1 Criteria Matrix							
	C1	C2	C3	C4	C5		
Procurement (C1)	1	1/3	1/5	3	5		
Production (C2)	3	1	1/3	3	7		
Logistics (C3)	5	3	1	5	9		
Communication (C4)	1/3	1/3	1/5	1	3		
Miscellaneous (C5)	1/5	1/7	1/9	1/3	1		
Procurement criteria matrix							
	C1	C2	C3	C4	C5		
Demand Forecasting (C1)	1	3	5	9	7		
Supplier selection (C2)	1/3	1	3	7	5		
Order management (C3)	1/5	1/3	1	5	3		
e-procurement (C4)	1/9	1/7	1/5	1	1/3		
Make or Buy decision (C5)	1/7	1/5	1/3	3	1		
Production criteria matrix							
	C1	C2	C3	C4	C5		
Production planning (C1)	1	1/5	7	3	1/4		
Automation (C2)	5	1	9	7	3		
Better Quality (C3)	1/7	1/9	1	1/3	1/7		
Improved R & D (C4)	1/3	1/7	3	1	1/5		
Asset management (C5)	4	1/3	7	5	1		
Logistics criteria matrix							
	C1	C2	C3	C4	C5	C6	C7
Assisted driving (C1)	1	1/7	1/5	1/9	4	1/3	1/5
Inventory management (C2)	7	1	2	1/3	8	7	4
Resource optimization (C3)	5	1/2	1	1/4	6	3	2
Improved logistics plans (C4)	9	3	4	1	7	6	4
Order fulfilment (C5)	1/4	1/8	1/6	1/7	1	1/4	1/3
Location allocation (C6)	3	1/7	1/3	1/6	4	1	1/3
container load planning (C7)	5	1/4	1/2	1/4	3	3	1
Communication criteria matrix							
	C1	C2	C3	C4			
Real time monitoring (C1)	1	5	1/5	3			

Communication and visibility (C2)	1/5	1	1/9	1/3			
Early warning and control (C3)	5	9	1	7			
SC partnership (C4)	1/3	3	1/7	1			
Miscellaneous criteria matrix							
	C1	C2	C3	C4	C5	C6	C7
Risk Management (C1)	1	6	5	1/3	1/2	3	7
Data analysis (C2)	1/6	1	1/2	1/6	1/5	1/3	3
Strategic advantage (C3)	1/5	2	1	1/5	1/3	1/2	4
SC planning (C4)	3	6	5	1	2	4	9
Resource utilization (C5)	2	5	3	1/2	1	5	6
Cost reduction (C6)	1/3	3	2	1/4	1/5	1	4
CRM (C7)	1/7	1/3	1/4	1/9	1/6	1/4	1

Table 8. Final Local and Global weights of all main and sub criteria

S No.	Main Criteria	Local Weight	Sub Criteria	Local Weight	Global weight
1	Logistics	0.49	Improved logistics plans	0.38	0.18601
			Inventory management	0.25	0.12185
			Resource optimization	0.14	0.06954
			container load planning	0.10	0.04958
			Location allocation	0.06	0.03083
			Assisted driving	0.04	0.02026
			Order fulfilment	0.03	0.01310
2	Production	0.25	Automation	0.49	0.12030
			Asset management	0.27	0.06631
			Production planning	0.14	0.03458
			Improved R & D	0.07	0.01607
			Better Quality	0.03	0.00806
3	Procurement	0.15	Demand Forecasting	0.50	0.07300
			Supplier Selection	0.26	0.03778
			Order management	0.13	0.01950
			Make or Buy	0.07	0.00984
			e-Procurement	0.03	0.00506
4	Communication	0.08	Early warning and control	0.64	0.05319
			Real time monitoring	0.21	0.01724
			SC partnership	0.10	0.00836
			Communication and visibility	0.05	0.00397
5	Miscellaneous	0.04	SC planning	0.34	0.01214
			Resource utilization	0.23	0.00834
			Risk Management	0.19	0.00694
			Cost reduction	0.09	0.00320
			Strategic advantage	0.07	0.00245
			Data analysis	0.05	0.00162
			CRM	0.03	0.00091

Discussion:

According to the final weights from the AHP analysis, it is clear that the logistics sector will be most benefited from the implementation of AI and IoT, followed by production and then by other domains. New technologies like GPS monitoring and RFID enabled transportation vehicles will help monitor the shipments status. Barcodes and RFID will help in inventory management. All this information, combined with the analytical power of AI, has a lot of scope in the improving the logistics performance. There have been new inventions and development in the field of self-driving vehicles and it is believed that these vehicles will improve the SC performance of Indian SCM also.

Production is a very essential part of supply chain and the scope of AI and IoT in the field of manufacturing and production can be seen in many organizations (Mangla et al., 2017). Automation with the help of robotics and better processes can help improve the production capabilities of the Indian SMEs. Using AI in production will not only reduce errors but will also reduce the disruption in the supply of finished goods, which in turn will keep the entire supply chain running smoothly. Better production planning will also help in achieving lean manufacturing and reduce wastes. The data received by the use of IoT technologies and analyzing them with the help of AI can help in improving the R&D, which will in turn improve the quality of the products.

Procuring the right material, at the right time and of the right quantity, has always been a challenge for SC professionals. Not only this, there are many other variables which affect the procurement process, from a business point of view especially in the Indian SME sector, like, the cost of the material, supplier relations, proximity of the supplier, etc. Having the information of all the potential suppliers and the materials they supply, combined with the accurate demand forecasting, can help the SMEs save a lot of money and material. This will not only reduce the cost, but will indirectly also affect the inventory and lead to lean manufacturing. All of the above mentioned tasks can be achieved by the implementation of AI and IoT technologies.

Apart from the mentioned domains, there has to be a proper communication between different SC partners to keep the SC running properly. The flow of information about the supplier, the manufacturer, the logistics provider, distributors and any other SC partner, has to be in real time and uninterrupted. If there is any issue with any of the SC partner, and the information has been communicated, the concerned party can immediately make other arrangements to tackle the situation. AI and IoT can lead to a better communication and help in developing an early warning and control system. This will also increase the SC visibility and help in real time monitoring.

The scope of development of SC practices in the Indian SMEs is huge, and also the rate at which Data Science and Artificial Intelligence is growing. It is believed that integrating the two, many challenges can be achieved in the Indian SMEs supply chain.

4.2. ISM Analysis for the barriers of AI and IoT implementation

The ISM analysis has been done as the steps mentioned in Figure 2. The first step is to develop the contextual relationship between different challenges and have been established by the judgement and opinion of same SC experts. The self-structural interaction matrix has been discussed in Table 6.

The relations are denoted by symbols, which are explained below:

- V – Barrier i will lead to barrier j
- A – Barrier j will lead to barrier i
- X – Barrier i and j will lead to each other
- O – Barrier i and j are unrelated

It is required to fill only half of the matrix as the relation of the other half can be interpreted (if cell (i,j) has a relation of 'V', then it is implied that cell (j,i) will have a relation of 'A').

Table 9. Self-Structural Interaction Matrix (SSIM)

Barrier No.	Challenges	15	14	13	12	11	10	9	8	7	6	5	4	3	2	1
1	IT infrastructure	V	O	O	A	O	A	X	X	V	X	A	V	V	X	X
2	Facility designs	V	O	O	A	A	A	X	A	V	O	O	O	V	X	
3	Information source	O	A	V	A	A	O	X	A	O	V	A	A	X		
4	Data storage platform	O	O	O	O	O	A	X	X	V	V	A	X			
5	High It skills required	O	O	V	A	A	A	V	X	V	V	X				
6	Cyber security	O	O	O	A	A	O	V	X	V	X					
7	Lack of overall control	A	O	A	A	A	A	A	A	X						
8	High speed	V	O	O	A	A	O	X	X							

	connectivity																
9	New business models	A	A	V	A	A	A	X									
10	Advancement in HR	V	O	V	X	X	X										
11	Understanding of Decision makers	V	A	X	X	X											
12	Awareness and willingness	X	X	V	X												
13	Trust and delegation	X	O	X													
14	Ontology	O	X														
15	Horizontal and vertical integration	X															

The SSIM is now converted to reachability matrix by substituting the symbols (V,A,X,O) with 1 and 0, by following certain rules described below and check for transitivity (Table 7):

- If the cell (i,j) is V then replace it with 1; If the cell (j,i) is V then replace it with 0;
- If the cell (i,j) is A then replace it with 0; If the cell (j,i) is A then replace it with 1;
- If the cell (i,j) is X then replace it with 1; If the cell (j,i) is X then replace it with 1;
- If the cell (i,j) is O then replace it with 0; If the cell (j,i) is O then replace it with 0;

Table 10. Final Reachability Matrix

	Challenges	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Driving Power
1	IT infrastructure	1	1	1	1	1*	1	1	1	1	0	0	1*	1*	0	1	12
2	Facility designs	1	1	1	1*	0	1*	1	1*	1	0	0	1*	1*	0	1	11
3	Information source	1*	1*	1	1*	0	1	1*	1*	1	0	1*	0	1	0	0	10
4	Data storage platform	1*	1*	1	1	1*	1	1	1	1	0	0	0	1*	0	1*	11
5	High It skills required	1	1*	1	1	1	1	1	1	1	0	1*	0	1	0	1*	12
6	Cyber security	1	1*	1*	1*	1*	1	1	1	1	0	0	0	1*	0	1*	11
7	Lack of overall control	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
8	High speed connectivity	1	1	1	1	1	1	1	1	1	0	0	0	1*	0	1	11
9	New business models	1	1	1	1	1*	1*	1	1	1	0	1*	0	1	0	1*	12
10	Advancement in HR	1	1	1*	1	1	1*	1	1*	1	1	1	1	1	1*	1	15
11	Understanding of Decision makers	1*	1	1	1*	1	1	1	1	1	1	1	1	1	1*	1	15
12	Awareness and willingness	1	1	1	1*	1	1	1	1	1	1	1	1	1	1	1	15
13	Trust and delegation	0	1*	1*	0	1*	1*	1	1*	1*	1*	1	1*	1	0	1	12
14	Ontology	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15

4		*	*		*	*	*	*	*		*			*		*	
1 5	Horizontal and vertical integration	1 *	1 *	1 *	1 *	1 *	1 *	1	1 *	1	1 *	1 *	1	1	1 *	1	15
	Dependence Power	1 3	1 4	1 4	1 3	1 2	1 4	1 5	1 4	1 4	6	9	8	1 4	5	1 3	178

From the final reachability matrix, reachability set and antecedent set for each attribute is derived. Reachability set contains the attribute itself and the attributes which it will help in achieving. The antecedent set consist the attribute itself and the attributes which will help achieving it. In other words the reachability set consists of all the attributes which has 1 along the row, and the antecedent set consists of all the attributes which has 1 along the column. From these two sets, an intersection set is derived which consists of the attributes present in both the reachability and antecedent sets.

The attribute(s) which has the same reachability and antecedent set, is considered as level-1 attribute(s) i.e. they occupy the top most level in the ISM model. Once the top most level is identified, the attribute is not considered for the remaining calculations and it is separated from the remaining variables. The same process is implemented to the remaining variables to find out the attributes in the next level. This process is repeated until all the variables or barriers are allocated to a level (Table 8).

Table 11. Level partitioning with all levels

Diagram formulation:

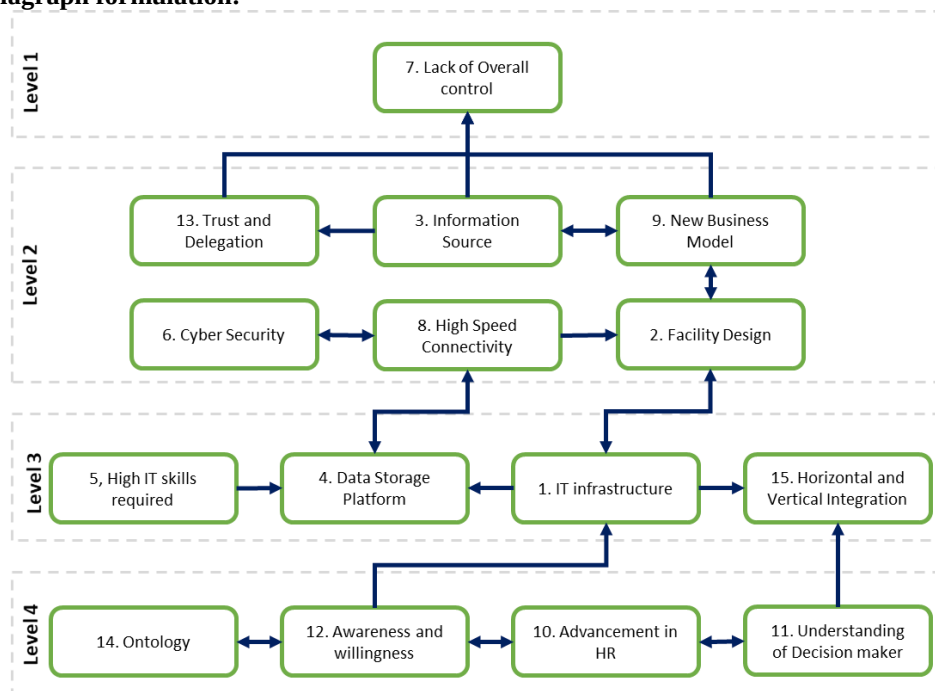


Fig. 24. Diagram - ISM Model

MICMAC Analysis

MICMAC analysis is done to analyze the driver and dependence power of different attributes or variables. It helps to identify the key variables which affects the system in different ways. As per the driver and dependence power, the variables are classified among 4 categories, described below:

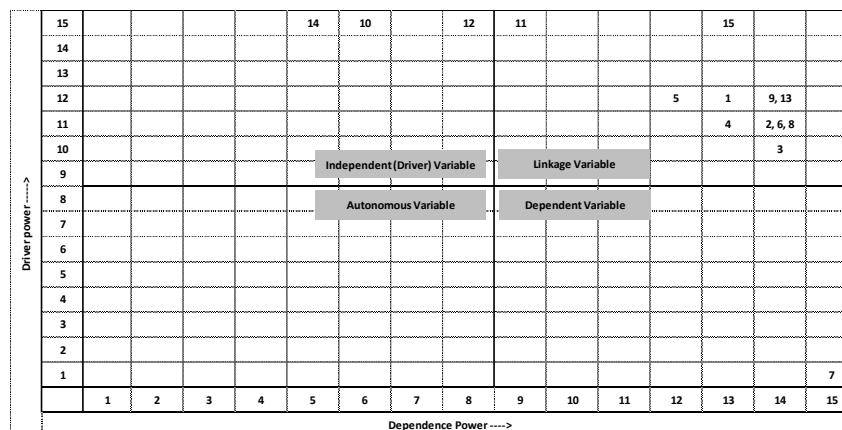


Fig. 25. Classification of Barriers

- **Autonomous Variables:** Variables with low driver and low dependence power falls in this category. They are relatively disconnected from the system and have few links with other variables.
 - **Dependent Variables:** These variables have low driver power and high dependence power and are majorly dependent on other variables.
 - **Linkage Variables:** These variables have high driver and high dependence power, and are considered unstable in nature as they have large number of linkages and any action on them will affect others.
- Independent (driver) Variables:** The variables in this category have high driver power and low dependence power. They are also known as “key enablers”.

1. Discussion

There is no doubt that there are many benefits of implementing AI and IoT in the SCM of Indian SMEs, but for them to function properly there are some conditions which needs to be met. These conditions acts like a barrier or challenge to the implementation of AI and IoT, which are discussed in the above sections. According to the ISM analysis, four levels of barriers have been identified.

Firstly, there is a strong need to make the SMEs aware about the application, advantages and most importantly the need of implementing the advanced technologies like AI and IoT. Most of the Indian SMEs have been operating on traditional manufacturing technologies, and since it is getting their job done, they don't feel the need of introducing IoT and AI in their manufacturing process. And it is believed because of this reason, even though most of the SMEs are aware about the technologies, they are not willing to invest in it. It may seem like an unnecessary investment in the beginning, but if we look at the SC of the entire SME and from the countries prospective, it is the need of the hour. The top management and the decision makes have to understand the working of these technologies in order to make necessary changes in the current status and to get the benefits of this integration. Apart from this, there has to changes in HR of the SMEs and people with good understanding of IT tools are required, to setup the basic platform for the integration of AI and IoT with SCM.

Once the basic requirements are met and the industries are set to the path of I4.0, the next level of challenge has to be tackled. All the IoT and AI systems need to have a fast and up to date computers and IT infrastructure to perform well. And to operate those IT tools people with good IT and programming skills are needed. There is also a need to integrate all the departments at different levels of the organization, so that more information can be procured for making a better decision and also for better communication. It is now implied that the AI and IoT systems works on lots of data, which is gathered from many different sources both inside and outside the organization. To manage this data, proper data storage platforms like cloud storage are required.

The next level of challenge is to improve the performance of the system. Since the IoT and AI will need to build a new business model, which will accommodate the new work culture. Many decisions will be made by AI systems and we have to delegate appropriate tasks to the AI. Not all the decisions can be made by the AI, but the suggestion by the system can be considered by human decision makers. This issue of trust and delegation will arise once the AI system are in place. This issue will also depend on the information source

i.e. where does the information about different variables affecting the decision are gathered from and how authentic it is. There is also the need of high speed connectivity to make the information flow uninterrupted and fast. And to protect all the data and the network from any kind of cyber threat like hacking, information leakage, data theft etc., the organization will have to invest and establish a robust cyber security system. Another challenge is to modify the designs of the facility to accommodate the AI and IoT systems. Inclusion of robotics as a part of automation and AI in manufacturing and inventory, will require designing a new layout to accommodate these machines.

The entire process of supply chain is very complex once we observe it at the ground level. There are lot of companies, SC partners and third parties which are a part of a supply chain, and any change in one node of the SC, eventually affect the entire SC (also known as the Bull-whip effect). Even though, studies show that one company may have relatively larger influence over the coordination of SC, there is no defined body to supervise the entire SC without any bias. As per the ISM results, this lack of overall control will only be addressed once the bottom level barriers are tackled.

2. Conclusion

Considering the Industry 4.0, and the Indian SME sector, there is a need to focus on the SCM of the businesses and also there is a huge potential of SCM of Indian SMEs to grow. There is also a rapid growth seen in the field of data science and AI. From the literature review we have found many benefits of AI and IoT in the field of SCM. For the Indian SMEs to cope up to the Industry 4.0 standards, it is believed that focusing on SC practices and integrating AI and IoT with SCM, is one of the best ways. This paper has shown different aspects of AI and IoT which can help in different domains of SCM.

Achieving this goal and vision of an AI integrated Supply chain in Indian SMEs is not going to be easy and many challenges has to be tackled for the same. Indian SMEs are way behind the requirements and the world standards. From the literature studied, this paper has addressed many challenges and it is implied that a lot of changes has to be made in the current working scenario in order to achieve the said goals.

3. Practical Implications and Future Scope

Almost all the companies in every sector is utilising the data to extract useful information and insights, which help the company to make better decisions. This study is aimed to analyse the scope of implementation of AI and IoT in the SCM. This study shows different possible application of AI and IoT in different sub-domains of SCM, and what challenges the companies may face while doing so. For instance, Demand forecasting can help in maintaining the inventory and logistics planning, an early warning system can give warning proactively to prevent any unfavourable situations, etc. Using simple algorithms, like regression and classification, SMEs can benefit a lot and make improved decisions. This study shows different opportunity areas where the SMEs can be benefited from these systems in place.

The future scope of research and application on this field is very vast. This paper addresses the benefits and challenges of AI and IoT, but the study of how these benefits will be achieved and how the challenges will be overcome is out of the scope of the project. There are so many specific field in the SCM which can be explored, where different algorithms of AI and IoT can be implemented. As it is known that data can be gathered from anywhere and any domain of SC, and the AI system work on these data, hence AI can be used in almost every field of SC. Deep learning algorithms like image recognition can help in developing self-driving vehicles suited to Indian traffic and road conditions. Different domains of AI can help in making more accurate predictions about demand. Many other example like these are already mentioned in the paper but the technicality of how to achieve it in the ground reality not discussed here. Apart from this, there has to be a model of SC, especially designed from an Indian prospective and for the Indian SMEs. There is a need to make a robust plan to prepare the Indian SMEs for the integration of AI and IoT with their SC.

References

- [1] Abdel-Basset, M., Manogaran, G., & Mohamed, M. (2018). Internet of Things (IoT) and its impact on supply chain: A framework for building smart, secure and efficient systems. *Future Generation Computer Systems*, 86, 614-628.

- [2] Ben-Daya, M., Hassini, E., & Bahrour, Z. (2019). Internet of things and supply chain management: a literature review. *International Journal of Production Research*, 57(15-16), 4719-4742.
- [3] Calatayud, A., Mangan, J., & Christopher, M. (2019). The self-thinking supply chain. *Supply Chain Management: An International Journal*.
- [4] Cardoso, M. T. C., & de Sousa, C. (2019). Artificial Intelligence: the new source of productivity in the fashion supply chain (Doctoral dissertation).
- [5] Chun, K.W., Kim, H., Lee, K., (2019). A study on research trends of technologies for industry 4.0; 3D printing, artificial intelligence, big data, cloud computing, and internet of things: advanced multimedia and ubiquitous engineering. *Lect. Notes Elect. Eng.* 518 (397-), 403.
- [6] Da Xu, L., He, W., & Li, S. (2014). Internet of things in industries: A survey. *IEEE Transactions on industrial informatics*, 10(4), 2233-2243.
- [7] Dash, R., McMurtrey, M., Rebman, C., & Kar, U. K. (2019). Application of Artificial Intelligence in Automation of Supply Chain Management. *Journal of Strategic Innovation and Sustainability*, 14(3).
- [8] Duan, Y., Edwards, J. S., & Xu, M. X. (2005). Web-based expert systems: benefits and challenges. *Information & Management*, 42(6), 799-811.
- [9] Govindan, K. and Hasanagic, M. (2018), "A systematic review on drivers, barriers, and practices towards circular economy: a supply chain perspective", *International Journal of Production Research*, Vol. 56 No. (1-2), pp. 278-311.
- [10] Gupta, A., & Singh, R. K. (2020). Managing operations by a logistics company for sustainable service quality: Indian perspective. *Management of Environmental Quality: An International Journal*.
- [11] Gupta, A., Suri, P. K., & Singh, R. K. (2019). Analyzing the interaction of barriers in e-governance implementation for effective service quality: interpretive structural modeling approach. *Business Perspectives and Research*, 7(1), 59-75.
- [12] Gupta, A., Singh, R., & Suri, P. (2018). Analysis of challenges faced by Indian logistics service providers. *Operations and Supply Chain Management: An International Journal*, 11(4), 214-225.
- [13] Luthra, S., Kumar, V., Kumar, S., & Haleem, A. (2011). Barriers to implement green supply chain management in automobile industry using interpretive structural modeling technique: An Indian perspective. *Journal of Industrial Engineering and Management (JIEM)*, 4(2), 231-257.
- [14] Mangla, S.K., Govindan, K. and Luthra, S. (2017). Prioritizing the barriers to achieve sustainable consumption and production trends in supply chains using fuzzy Analytical Hierarchy Process. *Journal of Cleaner Production*, 151, 509-525.
- [15] Min, H. (2010). Artificial intelligence in supply chain management: theory and applications. *International Journal of Logistics: Research and Applications*, 13(1), 13-39.
- [16] Mittal, S., Khan, M.A., Romero, D., Wuest, T., (2017). Smart manufacturing: characteristics, technologies and enabling factors. *J. Eng. Manuf.* 223 (5), 1342-1362.
- [17] Perçin, S. (2008). Use of fuzzy AHP for evaluating the benefits of information sharing decisions in a supply chain. *Journal of Enterprise Information Management*.
- [18] Plastino, E., & Purdy, M. (2018). Game changing value from artificial intelligence: Eight strategies. *Strategy & Leadership*, 46(1), 16-22.
- [19] Radanliev, P., De Roure, D. C., Nurse, J. R., Burnap, P., Anthi, E., Ani, U., & Montalvo, R. M. (2019). Cyber risk from IoT technologies in the supply chain-discussion on supply chains decision support system for the digital economy. no. March.
- [20] Raj, T., Shankar, R., & Suhaib, M. (2008). An ISM approach for modelling the enablers of flexible manufacturing system: the case for India. *International Journal of Production Research*, 46(24), 6883-6912.
- [21] Saaty, R. W. (1987). The analytic hierarchy process-what it is and how it is used. *Mathematical modelling*, 9(3-5), 161-176.
- [22] Singh, R. K. (2011). Developing the framework for coordination in supply chain of SMEs. *Business Process Management Journal*.
- [23] Suresh, N., Hemamala, K., & Ashok, N. (2018). Challenges in implementing industry revolution 4.0 in INDIAN manufacturing SMES: insights from five case studies. *International Journal of Engineering & Technology*, 7(2.4), 136-139.
- [24] Thakkar, J., Kanda, A., & Deshmukh, S. G. (2012). Supply chain issues in Indian manufacturing SMEs: insights from six case studies. *Journal of Manufacturing Technology Management*.
- [25] Thakkar, J., Kanda, A., & Deshmukh, S. G. (2011). Mapping of supply chain learning: a framework for SMEs. *The Learning Organization*.
- [26] Tripathy, S., Aich, S., Chakraborty, A., & Lee, G. M. (2016). Information technology is an enabling factor affecting supply chain performance in Indian SMEs. *Journal of Modelling in Management*.
- [27] Vaaland, T. I., & Heide, M. (2007). Can the SME survive the supply chain challenges?. *Supply chain management: an International Journal*.
- [28] Verdouw, C. N., Wolfert, J., Beulens, A. J. M., & Rialland, A. (2016). Virtualization of food supply chains with the internet of things. *Journal of Food Engineering*, 176, 128-136.
- [29] Webster, C., & Ivanov, S. (2020). Robotics, artificial intelligence, and the evolving nature of work. In *Digital Transformation in Business and Society* (pp. 127-143). Palgrave Macmillan, Cham.

- [30] Witkowski, K. (2017). Internet of things, big data, industry 4.0-innovative solutions in logistics and supply chains management. *Procedia Engineering*, 182, 763-769.
- [31] Yuan, Y., Liang, P., & Zhang, J. J. (2001). Using agent technology to support supply chain management: potentials and challenges.

Smart Breast Cancer Diagnosis using Machine Learning Techniques

¹ Amit Kumar Jakhar, ² Mrityunjay Singh*, ³ Shivam Pandey

^{1,2} Jaypee University of Information Technology, Waknaghat, H.P- 173234 India

³ Bundelkhand Institute of Engineering and Technology, Jhansi-284128 India

amitjakhar69@gmail.com, mrityunjay.cse045@gmail.com, vastav.shiv@gmail.com

Abstract. During the past few decades, a substantial amount of cancer cases have been reported throughout the world. The breast cancer is the widely acknowledged category of cancer disease in most of the women due to lack of awareness about this deadly disease. According to world cancer survey report 2018, there are about two million cases has been reported all over the world. As, the patient-doctor-ratio (PDR) is very high around the world; consequently, there is a need of an utmost machine-based smart breast cancer diagnosis system that is able to detect cancer at its early stage. The plan is to assemble scientists in both the restorative and the machine learning field to progress toward this clinical application. This paper presents a framework to detect of breast cancer at the early phase from the histopathological images of tumor cells with computer aided diagnosis tool. In this work, we use the dataset that consists of 7909 histopathological images obtained on 82 patients. The experimental framework reveals the acceptable results during analysis of several performance parameters of machine learning techniques.

Keywords: Machine learning, Breast Cancer, Classification, Benign, Magnification

1. Introduction

Nowadays, the machine learning techniques are extremely admired in most of the fields in the data analysis and prediction. Therefore, these techniques are mainly used for the prediction, classification, clustering, etc. The machine learning techniques requires the huge amount of historical data for learning and to provide a better prediction on the basis of learning [1]. The science of enabling computers programmable by analysis of historical data is called machine learning [2]. The healthcare also believes in large amount of data from diagnosis and pathology to drug discovery and epidemiology. Machine Learning (ML) techniques are used to facilitate in analysis of these data and produce revolutionary information for medical society.

As far as dangerous diseases are concern, the cancer is a one of the most precarious disease that causes due to uncontrolled growth of abnormal cells in the body. In cancer, the old cells do not decay and their growth gets uncontrolled and it is the second most deadly disease in the world [14, 19, 21]. The breast cancer is a kind of cancer that forms mass of cells such as tissues on breast. Most treatments, among the several treatments offered to cancer patients, depend on early diagnosis of cancer symptoms which effects survival of cancer patients. Further and continuous growth of cancer tissues start integrating on others normal human body parts, called metastasis [13]. When a patient is diagnosed with tumor, the very first step considered by the doctors is identifying whether it is malignant or benign. Malignant type of tumor is cancerous whether benign is non-cancerous. Therefore, differentiating tumor type is very important for better and effective cure and also required in the plan of treatment. Though, doctor requires a reliable mechanism to differentiate these two types of tumors. In almost all countries, cancer is the root cause of death around eight million people annually. Every possible treatment plan for cancer patient requires a deep study of behavioral changes in cells; approximately 85% of breast cancer cases come from the women [13].

Machine learning algorithms can be more reliable, effective and easier, and it will provide a better solution for automatic cancer disease detection. The formulation of learning algorithms having ability to adopt new data independently and iteratively; the machine learning techniques prerequisite are continuous learning from previous data. Learning helps machine to identify pattern available in data and to produce reliable and informed results for decision making. The main goal of machine learning is to make a system learn without human intervention which helps to design an automated system for decision making. The main consequence of machine learning algorithm is to predict an output variable, termed as dependent variable from the set of independent variables; these independent variables termed as predictors. With the help of machine learning algorithms, we model a function which maps predictors to desired output variables. The machine has been trained continuously to obtain a certain precision from our example data. The tumor type identification is the

essential phase to suggest better treatment plan to a patient that increase the probability of their survival. The cause of breast cancer disease is the presence of either benign tumor or malignant tumor. The type of tumor can be classified by using machine learning algorithms in an optimal way and can also suggest proper treatment to the needy patient. Therefore, an approach should be developed by using machine learning algorithms to predict the cancer at higher precision. This work is dedicated for getting better results by diagnosis of tumor type. In this paper, we apply the support vector machine, random forest, and decision tree algorithms on cancer data sets to predict the tumor type that supports better treatment.

The rest of the paper is organized as follows: Section 2 discusses the related work of the domain and Section 3 presents our proposed work. Then, Section 4 contains the discussion on experimental setup and results. The conclusion and future work is discussed in Section 5.

2. Related works

The combination of medical science and artificial intelligence has a great significance and several researchers have been working in the same domain and come up with extraordinary outputs. In the literature, the researchers have applied a number of classifiers to exploit the set of features; these classifiers are 1-nearest neighbor (1-NN), Support Vector Machines (SVM), Random forests (RF) and Quadratic Discriminant Analysis (QDA) [1, 15, 16, 20]. Support Vector Machine is a popular classification algorithm. This algorithm separates the given data objects into multiple classes using an optimal hyper plane. On the other hand, Random Forest classifier utilizes the decision tree predictors by combining them into one, and produces good results even without hyper-parameter tuning [16]. The advantages of the random forest method are: *efficiently works on unbalanced datasets and very fast*.

The other machine learning algorithms are taken into account for the breast cancer diagnosis using image classification [1, 2, 4, 15]. In image classification, the system divides the initial cancer image into patches; then we break patches from high-goals fixes of images that measure coordinates of the CIFAR informational index [3]. The results of fixes are linked with the complete image. At that moment, the system experiences the model and links with consequences. The ultimate result was achieved by ignoring all plausible patches from input images; however, this process is highly expensive in terms of computation. Moreover, a comprehensive response technique has been reported better with the newly proposed system, i.e., CSDDCNN [4]. CSDD under utilize the qualities of neural network and presume a necessary task in regulating mean improvement. The general multi-rating exactness technique has high precision and dropped dependable execution. For every one of the components, the typical patient's dimension correctness is 93.2%, whereas, the normal image exactness is 93.8%. Both test packs and approval packs have almost similar performance.

In this work, we have used the Local Binary Patterns (LBP) [5], Gray-Level Co-occurrence Matrices (GLCM) [6], Completed Local Binary Patterns (CLBP) [11], Local Phase Quantization (LPQ) [12], Parameter free threshold adjacency statistics (PFTAS) [8], Oriented FAST and Rotated BRIEF (ORB) [7] as feature extractors. The better result has been obtained by utilizing the feature extractor parameter free threshold adjacency statistics (PFTAS) combined with Support Vector Machine (SVM) [2]. In addition to it, the recognition is done at the patient level instead of image level. The receiver operating characteristics (ROC) curve is another important apparatus for execution investigation, particularly in unbalanced data sets [10]. The ROC curve is not sensitive; it does not work on variations in class distribution along with the positive (+ve) to negative (- ve) variations on testset. In the following section the tumor type is detected among the images of the dataset.

3. Proposed Work

This section discusses the work to diagnosis breast cancer using machine learning techniques. This work focuses on the problem of detecting tumor type from the given cancer image to predict the level of breast cancer present in a woman and to suggest better treatment to the patient. To apply a learning algorithm on a image data set, we first need to extract the features from given images, then we pass the extracted features to machine learning algorithm as input; to extract the various features from the given images a feature selection and extraction technique can be used. In this work, we perform the following steps to classify tumor type from the given cancer images: *first*, Tuning algorithm's parameters, *second*, feature selection and extraction,

third, Train the adopted machine learning models, and *finally*, predict the tumor type for the new images. Now, we explain each step in detail.

The machine learning algorithm gives better results in terms of prediction or classification when we tune its parameters. In order to train our machine learning model, we set the model's parameters; we refined the weightage of each learning parameters continuously to find out hyper parameters which result in better prediction. This process is called parameter tuning. The parameter tuning can be done using either manually or grid search strategy. In this work, we adopt the grid search algorithm strategy to tune the parameters of our learning algorithms [17]. In grid search strategy, we checked information from end goals to find out ideal parameters with respect to adopted learning model.

Feature selection is process of selecting the best parameters from a given data set to perform the machine learning task. In this work, we work on image dataset, therefore, we use OpenCV [14] library for image processing and local binary patterns descriptor to extract a set of features from the images. Local Binary Pattern (LBP) is a texture operator that gives result in a binary pattern and represents image pixels by threshold each pixel neighbors [5, 9]. The neighboring cells have fixed number of neighbors (P) that are present within the radius (R). We form a texture descriptor using the histogram of LBPs. LBP method has uniform patterns to make the different LBP codes; these codes are used to build the histogram containers, and generally decreases from 36 to 10. We assume that the number of neighbors (P) is 8; the produced feature vector is 10-D.

$$LocalBinaryPattern(LBP) = \sum_{p=0}^{p-1} s(g_p - g_c) 2^p \quad (1)$$

Once we select the optimal parameters from the dataset, then we train our learning model from training data. In this work, we have adopted support vector machine (SVM) classifier, random forest classifier, decision tree (DT) classifier, and neural network classifier to build our classification model. In the previous step, we have obtained the data as histogram. These histograms are given as an input to train our classifier. In this process, we tune work parameters to get more accurate results.

Once the training gets completed, we test our model on new data set, and measure the accuracy of our classifier. We consider the confusion matrix, accuracy score, and f-measure matrices to measure the performance of our algorithms. In this work, we use an exploratory strategy for model improvement and ensure the classification of malignancy with better accuracy. As our data set consist of class imbalance, e.g, the Adenosis, Phyllodes Tumor, Tubular Adenoma classes consisting of very less number of images, therefore, we have adopted the image data augmentation techniques for class balancing. Every image in the training data set of these classes is rotated, flipped and scaled. Rotation of images is done at 90 degrees, 180 degrees, 270 degrees. These values are taken so as to obtain the correct rotated image.

4. Experiment and Results

This section presents the experiment setup used in our work, and then we discuss the results of our work. In this work, we have used two data sets that consist of total 7909 number of images [1, 18]. Table 1 exhibit that the images are categorized into magnifications and benign. From Table 1, we can observe that there is huge class imbalance in the dataset. The numbers of malignant images are almost double to that of benign images. Malignant is further divided into Ductal Carcinoma (DC), Mucinous Carcinoma (MC), Lobular Carcinoma (LC), and Papillary Carcinoma (PC); the Benign is divided into Tubular Adenoma (TA), Adenosis (A), Fibroadenoma (F), and Phyllodes Tumor (PT) as shown in Table 2 and Table 3 respectively. Table 3 exhibits that the number of DC is very large than others thereby leading to class imbalance. So, many data augmentation techniques have been applied in order to deal with the class imbalance problem. We have a supervised learning problem where we need to classify among these 8 classes.

Table 1: Image distribution for Malignant and Benign data sets [1]

Magnification	Malignant	Benign
40X	1370	625

100X	1437	644
200X	1390	623
400X	1232	588
Total	5429	2480

Table 2: Image Distribution for Benign data set [1]

Magnification	A	F	PT	TA
40X	114	253	109	149
100X	113	260	121	150
200X	111	264	108	140
400X	106	237	115	130
Total	444	1014	453	569

We consider the Accuracy, precision, recall and F-1 scores as performance metrics to report the results of the machine learning algorithms on both datasets. The Accuracy, precision, recall and F-1 scores are defined as:

$$Accuracy = (TN + TP) / (TN + TP + FN + FP)$$

$$Recall = TP / (TP + FN)$$

$$Precision = TP / (TP + FP)$$

$$F1-Score = 2 * ((Precision * Recall) / (Precision + Recall))$$

Where, *TP*, *FP*, *TN*, *FN* denotes *True Positive*, *False Positive*, *True Negative*, and *False Negative* respectively.

Table 4 and Table 5 exhibit the results of our proposed framework. From there, it is being observed that the performance of the SVM is better than the other algorithms like random forest and decision tree. The same can be verified from the experimental results which are given in Table 4 and Table 5. We achieved the better results after data augmentation. LBP gave the best results on account of 40x amplification. The large execution of the descriptor was best on account of PFTAS. Future investigations may give a few inputs to the pathologist, to help him examining these images and characterizing a procedure to recognize areas to be investigated. At the point when machine learning procedures are contrasted with expert based frameworks or conventional factual frameworks, it is discovered that machine learning methods generally improve the execution with more precision.

Table 3: Image distribution for malignant data set

Magnification	DC	LC	MC	PC
40X	864	156	205	145
100X	903	170	222	142
200X	896	163	196	135
400X	788	137	169	138
Total	3451	626	792	500

After data augmentation, we achieved the better results. LBP technique produced the better outcomes on account of 40x amplification. On account of PFTAS, the descriptor also produced the better results on its

large execution. Future investigations may give a few inputs to the pathologist, to help him examining these images and characterizing a procedure to recognize areas to be investigated. At the point when a machine learning procedures are contrasted with expert based frameworks or conventional factual frameworks, it is discovered that machine learning methods generally improve the execution with more precision.

TABLE 4: Results using LBP as feature extractor on dataset 1

Classifier	Accuracy (without grid search)	Accuracy (with grid search)	Precision score	Recall score	F1 score
SVM	70%	88.3%	95.03%	94.3%	94.66%
Random Forest	70%	70%	91.42%	80%	85.33%

TABLE 5: Results using LBP as feature extractor on dataset 2

Classifier	Accuracy	Precision	Recall	F-1
SVM	90%	95.347%	90%	92.6%
Decision	97.22%	90.59%	87.22%	88.87%

5. Conclusion

This paper presents a framework to diagnose the breast cancer from different types of images using the existing machine learning algorithms. We have applied our proposed framework on two datasets for cancer prediction and evaluated the results. Our results show that the performance of the existing machine learning models is within acceptable level. In the near future, the detection of breast cancer become easier and faster if the machines are capable to made prediction correctly that would reduce the chances of false treatment, and even trim down the death rate because of breast cancer. This work can be extended by exploring the other existing machine learning algorithms so that the prediction becomes better and more efficient.

References

- [1] Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2015). A dataset for breast cancer histopathological image classification. *Ieee transactions on biomedical engineering*, 63(7), 1455-1462..
- [2] Hamilton, N. A., Pantelic, R. S., Hanson, K., & Teasdale, R. D. (2007). Fast automated cell phenotype image classification. *BMC bioinformatics*, 8(1), 1-8.
- [3] Spanhol, F. A., Oliveira, L. S., Petitjean, C., & Heutte, L. (2016, July). Breast cancer histopathological image classification using convolutional neural networks. In *2016 international joint conference on neural networks (IJCNN)* (pp. 2560-2567). IEEE.
- [4] Han, Z., Wei, B., Zheng, Y., Yin, Y., Li, K., & Li, S. (2017). Breast cancer multi-classification from histopathological images with structured deep learning model. *Scientific reports*, 7(1), 1-10.
- [5] Ojala, T., Pietikainen, M., & Maenpaa, T. (2002). Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on pattern analysis and machine intelligence*, 24(7), 971-987.
- [6] Haralick, R. M., Shanmugam, K., & Dinstein, I. H. (1973). Textural features for image classification. *IEEE Transactions on systems, man, and cybernetics*, (6), 610-621.
- [7] Rublee, E., Rabaud, V., Konolige, K., & Bradski, G. (2011, November). ORB: An efficient alternative to SIFT or SURF. In *2011 International conference on computer vision* (pp. 2564-2571). Ieee.
- [8] Coelho, L. P., Ahmed, A., Arnold, A., Kangas, J., Sheikh, A. S., Xing, E. P., ... & Murphy, R. F. (2010). Structured literature image finder: extracting information from text and images in biomedical literature. In *Linking Literature, Information, and Knowledge for Biology* (pp. 23-32). Springer, Berlin, Heidelberg.
- [9] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.
- [10] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, 27(8), 861-874.
- [11] Guo, Z., Zhang, L., & Zhang, D. (2010). A completed modeling of local binary pattern operator for texture classification. *IEEE transactions on image processing*, 19(6), 1657-1663.

- [12] Päivärinta, J., Rahtu, E., & Heikkilä, J. (2011, May). Volume local phase quantization for blur-insensitive dynamic texture classification. In *Scandinavian Conference on Image Analysis* (pp. 360-369). Springer, Berlin, Heidelberg.
- [13] Boyle, P., & Levin, B. (2008). *World cancer report 2008*. IARC Press, International Agency for Research on Cancer. [Online]. Available: http://www.iarc.fr/en/publications/pdfs-online/wcr/2008/wcr_2008.pdf
- [14] Culjak, I., Abram, D., Pribanic, T., Dzap, H., & Cifrek, M. (2012, May). A brief introduction to OpenCV. In *2012 proceedings of the 35th international convention MIPRO* (pp. 1725-1730). IEEE.
- [15] George, Y. M., Zayed, H. H., Roushdy, M. I., & Elbagoury, B. M. (2013). Remote computer-aided breast cancer detection and diagnosis system based on cytological images. *IEEE Systems Journal*, 8(3), 949-964.
- [16] Breiman, L. (2001). Random forests. *Machine learning*, 45(1), 5-32.
- [17] Syarif, I., Prugel-Bennett, A., & Wills, G. (2016). SVM parameter optimization using grid search and genetic algorithm to improve classification performance. *Telkomnika*, 14(4), 1502.
- [18] Mangasarian, O. L., Street, W. N., & Wolberg, W. H. (1995). Breast cancer diagnosis and prognosis via linear programming. *Operations Research*, 43(4), 570-577.
- [19] Gayathri, B. M., Sumathi, C. P., & Santhanam, T. (2013). Breast cancer diagnosis using machine learning algorithms-a survey. *International Journal of Distributed and Parallel Systems*, 4(3), 105.
- [20] Arshdeep Singh, Mohak Goel and Mrityunjay Singh. "Breast cancer classification by random forest classifier with feature decomposition using principal component analysis" in the Proceeding of the *International Conference on Data and Information Sciences (ICDIS 2021)*, May 14-15, 2021, Agra, published in Springer Book Series "Lecture Notes in Networks and Systems (LNNS).
- [21] Wild, C. (2014). *World cancer report 2014* (pp. 482-494). C. P. Wild, & B. W. Stewart (Eds.). Geneva, Switzerland: World Health Organization.

Progressive Strategies For Identification Of Straggler Node In Hadoop Mapreduce Environment

Ajay Kumar Bansal¹, Ashu Gupta², Manmohan Sharma³,

^{1,3} School of Computer Science & Engineering, Lovely Professional University

² Prince Sultan University, Riyadh, Kingdom of Saudi Arabia

Abstract: In today's era, Big data processing capabilities is a challenging task and bigger than this is finding the node which is performing slow(est) in the distributed environment. This paper goes through the various technological as well algorithmically development to proactively identifying the straggler nodes in the Hadoop MapReduce environment. Various algorithms have been reviewed on homogeneous as well as on heterogeneous cluster and with varying loads at various nodes in the cluster. It has been identified that there does not exist a single strategy or algorithm which address all the issues because of which a node becomes a straggler node rather it is a tradeoff between various parameters responsible for straggler node. Also, the varied software and hardware configuration of the processing environment also impact the performance strategies or techniques used to locate the straggler node in advance.

Keywords: Straggler, MapReduce, Hadoop, Speculation, Data center, partitioner

1. Introduction

In the recent years, information technology has evolved and penetrated in every aspect of the life. Digital devices have become a major source of capturing, storing and processing the data. [12] **Fig. 1** showed that companies, corporate, academia and even government have started using huge amount of data to gather intelligence and make informed social, business and strategic decisions. Almost half of the organizations out of 560 believe that Big Data techniques or tools will optimize or improve operational productivity and many more aspects to increase profits.

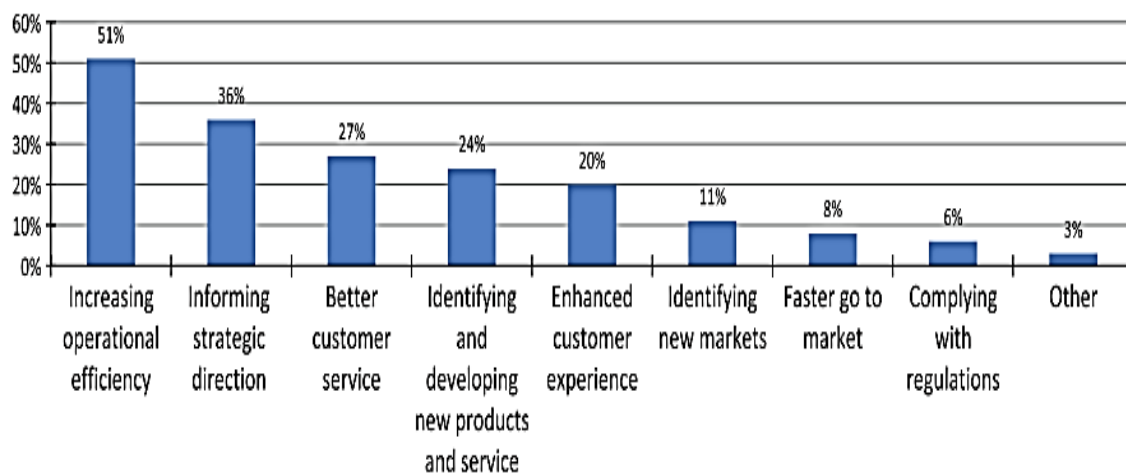


Fig. 1 Big Data usage

The Big Data primarily is collections of unstructured, heterogeneous, massive digital data that is difficult to handle and process by the most widely used structured data management tools and techniques like Oracle, SQL server etc.. There are many challenges in handling the Big data and related applications, difficulties lie in data capture, sharing, searching, analysis, storage and visualization [3]. The real challenge lies in using the data and it is important that the data should be available on time which is accurate and complete in nature [7]. Various tools and techniques in terms of hardware and software are available right from capturing to visualization of data.

[1] **Table 1** presents the various frameworks, programming languages and tools based on their Big Data applications.

Table 1 : Frameworks, programming languages and tools for Big Data applications [12,19]

Name	Advantages	Specified use	Programming language
Batch analysis			
Google MapReduce	Simple, scalable and fault tolerant	Data processing on large clusters	Java
Apache Hadoop	Scalable, reliability, extensibility, completeness	Infrastructure and platform	Java, Python, R, HiveQL, Pig Latin
Microsoft Dryad	Ease in programmability and fault-tolerant	Infrastructure and platform	DryadLINQ, SCOPE
Apache Mahout	Scalable, good maturity	Machine learning algorithms	Java
Stream analysis			
Apache Storm	Simple, Fault-tolerant, efficient, scalable, easily usable and operationally efficient	Real-time computation system	All
Apache S4	Scalable, fault-tolerant, extensible and a proven model	Stream computing platform	Java
Apache Spark	Fast, general, easy to use	Large-scale data processing engine	Java, Scala, Python
MOA	Extensible, scalable	data stream mining framework	Java
Interactive analysis			
Apache Drill	Agile, flexible, familiar	Hadoop and NoSQL (SQL query engine)	SQL
SpagoBI	Agile, real time Business Intelligence on Big Data streaming	Business Intelligence	Java
D3.js	Scalable	Interactive	JavaScript

MapReduce is a programming framework used for effectively addresses such wide variety of complex queries. MapReduce has been widely used and acclaimed by academia and industry because of its characteristics like Simplicity, capability, fault tolerance and scalability. [20] Fig. 2 and [18] fig. 3 represents the tremendous growth in the research work and scholarly articles on big data on year to year basis and by renowned publishers respectively.



Fig. 2 Research papers published year wise on Bigdata



Fig. 3 Research papers published by different publishing house

MapReduce is a good programming model for Big Data processing and applications (**Fig. 4**) [22]. One node from the available node is selected as master node and responsible for scheduling the task and remaining nodes are treated as map or worker nodes. The input data are divided into chunks of almost equal size and the master node assigns chunks to Map worker nodes. Each worker node processes the assigned input chunk of data and produces key/value pairs and these key/value pairs are then stored in files for further processing. These files are called intermediate files. The master node informs the node which hold the Reduce function about the link or address of the intermediate files and the Reduce node read the data from the intermediate files and process it according to functions and the result produced stored to output files.

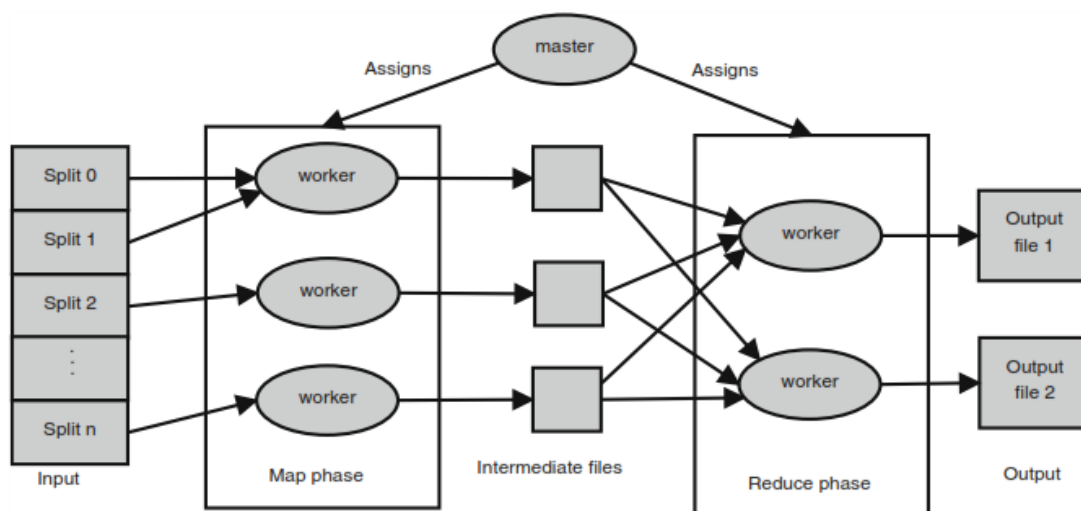


Fig. 4 MapReduce strategy

The problem lies in when one of the map workers fails or very slow in processing, the reduce function is either not able to produce the desired result in real time or fails to deliver the result. The map worker node which is very slow is called a straggler node.

2. Literature Analysis

Google MapReduce [2] is a Java-based framework which is a parallel and batch-oriented computing model distributed framework designed for a cluster of machines. Analytics with Hadoop [5] involves providing files containing data to the distributed file system, and processing it by using parallel MapReduce tool on the data. MapReduce processing in Hadoop can be coded using “general-purpose programming languages like Java or Python”, specific applications programming language R, or using Pig Latin or HiveQL to automatically generate declarative statements which look like SQL-queries. The large collection of various popular tools or programming languages supported by Hadoop makes it more popular for deep analytics. Various core components of Hadoop [8] like input/output data formats, storage subsystem, scheduler, compression algorithms, data partitioner, monitoring and caching layer can be customized, reprogrammed and can be replaced very easily [19].

The Hadoop system has very rapidly become a most popular tool in industry as a scalable with high data-intensive *MapReduce* platform [10], and now a day widely used for many case specific applications like clickstream, Web indexing, log analysis, machine learning jobs and certain large volume information extraction [4]. *Hadoop* is used by very renowned high-tech companies such as Facebook, LinkedIn, Amazon and Twitter [9,13].

The author discussed that most of the data resides on the map node itself while this data locality feature is not considered while creating a speculative task for the straggler node but the same is not applicable in visualized data centers [6]. The author proposed a model by placing the data on various nodes to balance the load on various data processing nodes (map or reduce nodes).

The author discussed the reasons of poor performance of MapReduce implementations on heterogeneous clusters “(e.g., on a 90-node cluster that contains 10 Xeon-based servers and 80 Atom-based servers, Hadoop performs worse than on 10-node Xeon-only or 80-node Atom-only homogeneous sub-clusters for many of our benchmarks)” [6]. The author proposed the *Tarazu*, a collection of tools for optimizing and improving the MapReduce performance on heterogeneous clusters. *Tarazu* consists of “*Communication-Aware Load Balancing of Map computation (CALB)*” on map nodes which are involved in processing and “*Communication-Aware Scheduling of Map computation (CAS)*” to avoid large volume data traffic on network and “*Predictive Load Balancing of Reduce computation (PLB)*” across the nodes.

Author discussed about the execution of MapReduce function on a cluster with homogeneous nodes [17]. The master node checked the failure node by pinging every map node periodically. If map node does not give response no response in a fixed interval of time, then the master node treats that node as a failure node and marked it as a failed node. The author also proposed to back up the task completed by nodes in case if the task is very close to complete its designated work.

In the paper [14], author focus on the tasks which are smaller and interactive in nature. The impact of straggler nodes is calculated for these kinds of jobs which become a bottleneck for the analyzing the data available on data centers. The author talked about the Microsoft Bing and Facebook’s production clusters. The author concludes that the impact of stragglers by using the existing latest straggler mitigation techniques, the overall task becomes on an average 8 times slower if compared with the median time to complete the task. The average job completion time is increased by 47% times due to stragglers. Author proposes to completely clone these small straggler jobs instead of waiting and speculating after the identification of straggler with the existing algorithm. Author accepts the fact that these extra clones can create congestion for intermediate data. Author proposed a technique which can efficiently avoid such congestion and it is called delay assignment. The evolution showed that by just using the 5% extra resources for cloning, the small jobs are increased by 36% to 46% after applying the latest state-of-the-art mitigation techniques.

In this paper [15], the author discussed issue of statically assigning the task to workers in Hadoop MapReduce environment and problems caused due to this. The author considered the resource utilization of every task tracker node and based on this utilization proposed an Automatic control technique to dynamically assign the task to various map workers. The results from experiment showed that this automatic assigning of tasks adjust the best optimal settings during execution.

The author discussed [16] that Speculating and creating clone for the stragglers are the existing techniques to help remove the issues of stragglers and reducing the performance in heterogeneous environment for big data processing. The author proposed to profile the results analysis and system parameters like processing power of nodes, current progress rate and utilization of various resources named as TPC-H. The author proposes a technique named MrHeter [23] in which MapReduce task is divided into two parts as mapshuffle and MapReduce stage. Author added that different optimization model is then used for these two different tasks and tasks are allocated keeping in mind the processing capabilities of individual node. Experiment shows that MrMeter and D-MrMeter can reduce the overall completion time of the task from 30% to 70% heterogeneous environment as compared to existing Hadoop computing model.

The algorithm suggests the dynamic adjustment [21] of load to various nodes of different computing powers in run time keeping in mind the existing work load on the node. The experimental results show completion time of the task is reduced very drastically and helps in optimizing balance of work load of the cluster as compared to Hadoop's default algorithm.

Table 2 : Comparative analysis of important Straggler finding algorithms

Algorithm	Findings	Challenges/Future Scope
HADOOP NATIVE SCHEDULER [5]	* Equal Weightage for Map and Reduce Tasks * Calculate average PR (Progress rate) and treat all nodes as straggler having 0.2 PR than the average PR	* Homogeneous in nature * Average based calculations
LATE [6]	* Predict Finish Time based on current progress rate * Identify straggler based on future finish time	* Not true heterogeneous in nature * End time estimation is based on current advancement rate which can be error prone * Starting assessment time is much higher (1 minute)
TARAZU [11]	Causes Detection and Solution	* Truly not intended for heterogeneous situation * Can cause normal node to be treated as straggler
DOLLY [14]	Simple Cloning	* Not good for long jobs * Run multiple speculative jobs which is an overhead on the system
WRANGLER [15]	* Using the job logs and snapshot of resource usage counters, a statistical model is built for each node. * Node's load is calculated periodically using this model. * If a node is heavily loaded, task on that node is speculated.	* Time consuming for short (in terms of time) jobs. * Time consuming for big jobs (in terms of volume of data and nodes)
Esplash [24]	* Uses History table and compare the progress rate with the fastest/completed node. * Speculate the task which are slowest ones	* does not consider the current status of the node. * Slowest node is compared with the fastest node.

Author discussed the problems of increase in the job completion time and wastage of resource in handling stragglers [24]. The reason is not proactively identifying stragglers rather focusing on reactive measures after a node is identified as stragglers. Author proposed a prognostic system based on predictive models. In first phase, the author uses the distributed machine learning model to predict the straggler nodes before allocating the task to a node and this way guiding the task scheduler to assign task. The proposed model is compared with Hadoop's native scheduler.

3. Conclusion

It has been observed after studying the various papers on stragglers node that there is no particular model which can effectively identify the straggler node in the early stage. Identifying the straggler node is a challenge in distributed environment and speculating the straggler on another node is another challenge. Different strategies are used for heterogeneous node vs homogeneous nodes, smaller jobs vs longer jobs or varying loads on worker nodes. No single strategy is effective in all environments. Further, the changes in hardware as well as software architecture in distributed or cloud environment also leads to innovate/modify the new/existing technique or algorithm.

References

- [1] Dean J, Ghemawat S et al (2008) "MapReduce: Simplified Data Processing on Large Clusters", COMMUNICATIONS OF THE ACM 51(1) p. 107-113.
- [2] Dean J, Ghemawat S (2008) "MapReduce: Simplified Data Processing on Large Clusters", ACM 51(1) p. 107-113.
- [3] Gordon Bell, Tony Hey, Alex Szalay (2009) "Beyond the data deluge", Science 323 (5919) p. 1297-1298.
- [4] Tony Hey, Stewart Tansley, Kristin Tolle (2009) "The fourth paradigm: data-intensive scientific discovery", Microsoft Research.
- [5] White T (2009) "Hadoop : the definite guide", 1st edn. O'Reilly Media Inc, Sebastopol.
- [6] Jiong X, Shu Y, Xiaojun R, Zhiyang D, Yun T, James M, Adam M, Xiao Q (2010) "Improving MapReduce Performance through Data Placement in Heterogeneous Hadoop Clusters", IEEE
- [7] Herodotou H, Lim H, Luo G et al (2011) "Starfish: a self-tuning system for big data analytics", Proceeding of the 5th biennial conference on innovative data systems research (CIDR 11), p. 261-272.
- [8] Borkar V, Carey MJ, Li C (2012) "Inside "Big Data Management": ogres, onions, or parfaits?", Proceeding of EDBT/ICDT joint conference. ACM
- [9] Lin J, Ryaboy D (2012) "Scaling big data mining infrastructure: the twitter experience", SIGKDD Explorations 14(2) p. 6-19.
- [10] Begoli E, Horey J (2012) "Design principles for effective knowledge discovery from big data", Proceeding of the joint working IEEE/IFIP conference on software architecture (WICSA) and European conference on software architecture (ECSA), p. 215-218.
- [11] Ahmad F, Chakradhar S, Raghunathan A, Vijaykumar T N (2012) "Tarazu: Optimizing MapReduce On Heterogeneous Clusters", ACM p. 61-74.
- [12] James M, Michael C, Brad B, Jacques B, Richard D, Charles R, Angela H B (2012) "Big data: The Next Frontier for Innovation, Competition, and Productivity", McKinsey Global Institute.
- [13] O'Driscoll A, Daugelaite J, Sleator RD (2013) "Big Data", Hadoop and cloud computing in genomics", J Biomed Inform 46(6) p. 774-781.
- [14] Ananthanarayanan G, Ghodsi A, Shenker S, Stoica I (2013) "Effective Straggler Mitigation: Attack of the Clones", 10th USENIX Symposium on Networked Systems Design and Implementation, p. 185-198.
- [15] Weina W, Kai Z, Lei Y (2013) "A Throughput Optimal Algorithm for Map Task Scheduling in MapReduce with Data Locality", Performance Evaluation Review 40(4) p. 33-42.
- [16] Xia Z, Kai K, YuZhong S, Yin S, Minhao X, Tao P (2013) "Insight and Reduction of Map Reduce Stragglers in Heterogeneous Environment", IEEE.
- [17] Demchenko Y, Grosso P, de Laat C, Membrey P (2013) "Addressing Big Data Issues in Scientific Data Infrastructure", International Conference on Collaboration Technologies and Systems (CTS). IEEE Computer Society.
- [18] Talia D (2013) "Clouds for scalable big data analytics", Computer 46(5) p. 98-101
- [19] Katal A, Wazid M, Goudar RH (2013) "Big data: issues, challenges, tools and good practices", Sixth international conference on contemporary computing (IC3), p. 404-409
- [20] Zhang X, Liu C, Nepal S et al (2014) "A hybrid approach for scalable sub-tree anonymization over big data using MapReduce on cloud", Journal of Computer and System Sciences 80(5) p. 1008-1020
- [21] Chen PCL, Zhang C-Y (2014) "Data-intensive applications, challenges, techniques and technologies: a survey on big data", Inf Sci Elsevier 275 p. 314-347.
- [22] Grolinger K, Hayes M, Higashino WA et al (2014), "Challenges for MapReduce in big data", Proceeding of the 2014 IEEE world congress on services (SERVICES) p. 182-189.
- [23] Zhang X, Wu Y, Zhao C (2016) "MrHeter: improving MapReduce performance in heterogeneous environments", Cluster Computing 19(4) p. 1691-1701.
- [24] Deshmukh S., Aghav J., Rao K.T., Rao B.T. (2017) "Avoiding Slow Running Nodes in Distributed Systems", Computer Communication, Networking and Internet Security. Lecture Notes in Networks and Systems 5 p. 411-420

Security and Privacy preservation in blockchain based IoT systems

Ms. Priyanka¹, Dr. Hemant Verma², Dr. Ritu Makani³

¹Research Scholar, Guru Jambheshwar University of Science and Technology, Hisar, Haryana

²Assistant Professor, Ch. Bansi Lal University, Bhiwani, Haryana

³Associate Professor, Guru Jambheshwar University of Science and Technology, Hisar, Haryana
Priyankaarora2844@gmail.com, hemantverma21@gmail.com, ritunagpal1973@gmail.com

Abstract. This paper presents a case study based on utilization of BC technology in the IoT systems and also presents how the blockchain technology provides security, integrity and privacy. From the invention of blockchain by Satoshi Nakamoto for cryptocurrency, mainly for Bitcoin, blockchain is also used in different applications by different organizations. IoT is mostly used by smart automated devices, smart homes, smart healthcare, smart grids, where achieving security and privacy is a big challenge. For providing the security, integrity and privacy in IoT we emerged Blockchain-based peer-to-peer, distributed, and decentralized technology. In this paper, we describe the need for Blockchain for the Internet of things (IoT) system, for achieving the security and privacy of data. We also present a basic view of IoT and Blockchain with their working and types of blockchain and describe the various applications of IoT with their Pros, cons in different areas.

Keywords: IoT, Blockchain, Security, Privacy, Authentication, Smart Devices.

1 Introduction

BC is a peer-to-peer, open decentralized and distributed digital ledger structure which store all the details of transactions in a transparent way and anyone can see all the transaction details whenever required. Every time addition of new block needs full mining and verification of new node, if the block is valid then it is added to blockchain otherwise rejected. In Blockchain based system there is no need of third party verification and control, the existing nodes perform this verification task and get reward for this job. Now a days, the internet of things has been used in various automation systems where all automatic devices and sensor networks are connected with each other and providing services. Internet of things has made life easier with smart city, smart agriculture, smart transportation, smart home. Various devices are connected in this network, where providing security and integrity of data is a great concern. The current research work shows that employing Blockchain seems a better way for security concerns regarding IOT. The current IOT (Internet of Things) based system provide centralization of devices, where all the devices are controlled by a central device and needs third party interaction for communication and transaction processing which is time consuming and also generate scalability problem. To solve these problems, we can use distributed and decentralized blockchain based IOT (Internet of things) system that also provide security and privacy of data.

These are some objectives of this paper:

- To discuss the overview of blockchain with their types and working procedure.
- To Describe IoT (Internet of things) with various Applications.
- To describe the Blockchain involvement in IoT based system.
- Limitations of Blockchain based IoT system.

2. IoT and Blockchain: An Overview

An overview of IoT:

The term IoT refers to Internet of Things and mostly used where we need less human intervention in between IoT devices to generate, exchange and processing of data where sensors and internet play a very important role in automation. The IoT (Internet of things) based devices can be controlled and sensed in a large area over the current network system, that creates opportunities for direct combination of physical world & computer based system and result is better efficiency, correctness & economic profit when IoT (Internet of things) is improved with sensors,

automated devices & actuators, this technology becomes an example of basic class of imitated actual physical systems, which includes smart technologies like smart transportation, smart grids, smart home & smart cities.

Fig: 1 IoT [11]



Strategies Data to Implement IoT (Internet of things) Data Management

- A. Identify use cases: Before adopting IoT (Internet of things) solutions, business leaders must identify their potential IoT (Internet of things) use cases for various business procedures.
- B. Hire skilled professionals: Skilled professionals can help in developing and executing effective IoT (Internet of things) adaptation strategies.
- C. Allocate sufficient budget: Business leaders need to carefully understand several necessities for successful IoT (Internet of things) data management such as infrastructure and resources and allocate funds accordingly.
- D. Understand data requirements: Before beginning the data collection process, organizations have to understand which types of data will be required and how much data storage they own.
- E. Implement Security Protocols: Business needs to implement a multifaceted approach for data security that includes data encryption and restrict sensitive data access.
- F. Educate and train employees: Business must educate their employees about IoT solutions and IoT data management.

Some Applications of IoT:

These are some possible areas where we can take advantage of IoT to clarify day to day problems.

- A. Industrial Automation: By applying this technology, we can automate the processes of production remotely. It is beneficial in optimizing the processes of production.
- B. Health Monitoring: Health problems can be identified by using this technology. If there are any deficiency in heart rate pattern, pulse, digestive system & blood pressure, we can keep an eye on that.
- C. Smart Environment: A very crucial application of Internet of things is to disclose pollution and natural calamities. We can control radiation from factories & emission of smoke from vehicles to minimize pollution with air.
- D. Smart Cities: The Internet of things can be used to monitor vibrations of bridges, buildings and monuments in case building material are threatened or it may be overloaded. Sound smog can be guarded around schools and hospitals. It can be used to handle traffic especially during peak hours, traffic jams, rains, accidents.



Fig: 2 Smart Cities [12]

- E. Home Automation: The Internet of things can be used to program and control home appliances remotely. It can be advantageous in finding & avoiding thefts home automation is process of controlling home appliances automatically using various techniques of control system.



Fig: 3 Home Automation [13]

- F. Industrial internet: Industrial Internet is one of the special Internet applications of IoT. While many market research like Gartner or Cisco sees the industrial internet as the most powerful IoT concept, its current popularity is not reaching as many people as smart home or portable. The industry sector gets more people on Twitter (~ 1,700 tweets per month) compared to other non-consumer IoT concepts.
- G. Connected car: The concept of connected car is growing slowly as compared to other IoT devices, because it takes about 2-4 years in development in automotive industry. There is not so much development in connected cars till now. But slowly we are reaching there. Some big car companies like Tesla, Google, Apple, BMW are continuously working in this area.

2 An overview of Blockchain

Background

Blockchain is firstly developed by Satoshi Nakamoto in 2008, through the digital cryptocurrency known as Bitcoin. It is based on distributed and peer-to-peer network- based technology that is called blockchain, the information of transitions is registered and maintained in a distributed public ledger across a number of computers that are directly connected to each other. In normal way of money transfer, we have faced various

problems like double spending and third-party interaction and transaction fee. To solve all these problems, we involve blockchain with cryptocurrency that not only solve all these problems but also provide security of data. It is a collection of nodes or blocks where each block consists of the group of block transactions and all the blocks are connected in a linked list manner. If we want to add new Block then it will be added to chain in sequential manner after a complete verification process. Blockchain requires a strict verification procedure to record a transaction which leads to a delay due to time necessary to confirm a block and it requires extensive use of resources like CPU's and GPU's power. At present it takes around 10 minutes for a transaction to be confirmed. When a new block is added or alteration of existing block is required then the full node or miners are used for verification procedure. By using various consensus or verification algorithms like POW, POS etc. the new node is verified, if the node is valid then it is added to the blockchain otherwise rejected. After addition of new node this information is transferred to all existing blocks with updated ledger.

General Structure of Blockchain

The Basic structure of Blockchain systems is common in various applications. Mostly Blockchain contain a group of blocks which are connected in a linked list manner where hash of previous block is directly connected to the next block and helps to generate new hash value and no one can tamper the data of block due to this sequential connected hash value, if anyone changes the data of block the hash of all subsequent blocks will also change and we can easily find out these changes of node. In which the very first block is represented as the Genesis Block which contains hash value zero. All subsequent block is added by performing complex verification or consensus algorithms like POW and POS, hash and timestamp values are also contained by blocks. Each block has their different Key data that is based application to application.

Working Procedure

Fig. 4 Describes the working of blockchain transactions process.

1. First of all, a transaction is Requested
2. After requesting transaction, the new Block for that transaction is created
3. This Block is circulated over the network and each block received this block
4. Full Node Validate the transactions
5. After the validation process of new block, the full node gets rewards.
6. If the new Block proved valid then it is added at the last of existing Blockchain
7. At the end Transaction process is completed.

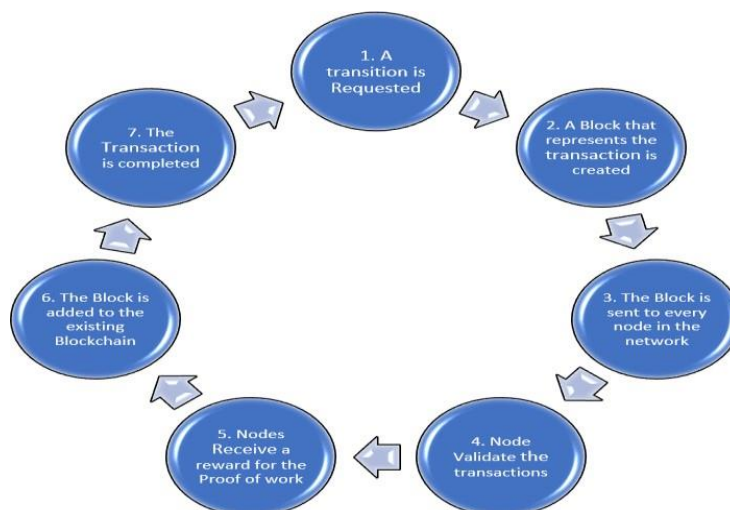


Fig.4 Working of Blockchain

Types of Blockchain

Currently there are three types of Blockchain systems which are mostly used:

1. Public Blockchains: Anyone can join and become the part of public Block-chain and there are no restrictions to access the Blockchain. Example: Bitcoin and Ethereum etc.

2. Private Blockchains: Private Blockchain have some restrictions in their joining process. This type of Blockchain is mostly used by private organization where they don't want to share their private data with public.
3. Consortium Blockchain: When two or more companies wants to work and control the operations of blockchain together then this type of blockchain can be used.

4. Security and Privacy of Blockchain based IoT system:

Now a days we use IoT (Internet of things) based devices in different places, it may be used in smart grid and smartcity, smart home, smart environment and in organizations we also used automated devices for providing easy services. These IoT based devices play a very important role in various applications where security is major concern for protecting the data in all areas. For providing the security and privacy we can use Blockchain. Blockchain not only provide security and privacy, we can also achieve some more privacy properties like confidentiality, integrity, authentication, availability, authorization and access control. For providing the privacy of data in which we can use various asymmetric and symmetric algorithms like ECC, RSA, ASE etc. and Hashing algorithms are used for providing integrity to transactions. Elliptic Curve Digital signature and Proof of Trust (PoT) also used for providing the trust and integrity. For Discarded the malicious re-quest and availability of data in Internet of things (IoT) is totally depends on handling authorized nodes and discarded unauthorized blocks [4] [8]. Authentication is provided already by default because the complete system is distributed and decentralized where every node is verified before addition in blockchain. The property of non-repudiation also achieving by public ledger of blockchain, in which anyone can see all the transactions so that the user can't deny after performing the transactions.

5. Limitations

The application opportunity and scope of Blockchain is very high due to the convenience of decentralization and distributed management and transparency. The Blockchain is used in various applications but it needs more research to remove some limitations and to make it more better technology. The main limitation of Blockchain is high cost of maintenance and needs lots of CPU's and GPU's Power and time consumption in mining process [2]. Sometime scaling issue also faced due to decentralization then we think centralization is better than decentralized maintenance or hybrid use of centralization and decentralization is also a better solution. Blockchain also faced some security attacks like 51% attack and transparency. In Blockchain mostly data are transparent and anyone who is part of blockchain can see this transaction information. The 51% attack is also a serious threat, in this attack if 6 out of 10 blocks or 51% blocks are malicious nodes then they can temper the data of blocks and control the complete blockchain. With these limitations blockchain provide various feature to IoT (Internet of things) like flexibility and various security properties like non repudiation, availability, authentication and authorization.

6. Conclusion

This paper described how blockchain involved in IoT (Internet of things) based system where Blockchain is distributed and decentralized technology and provide security and IoT (Internet of things) provide automation in various applications. IoT (Internet of things) and Blockchain independently used in various application with their features and limitations. By combining together both technology we overcome some limitations and can develop a secure and automated environment. In Blockchain various consensus algorithms are used that protect data from deliberate alteration and also protect from accidental loss of data. With combined advantages of both these technologies we require further research attentions.

7. Some Abbreviations used in this paper:

BC: Blockchain
POW: Proof of work
POS: Proof of Stack
IoT: Internet of things

References

- [1] Y. Rahulamathavan, R. C.-W. Phan, M. Rajarajan, S. Misra, and A. Kondo, "Pri-vacy-preserving blockchain based iot ecosystem using attributebased encryp- tion," in 2017 IEEE International Conference on Advanced Networks and Tele- communications Systems (ANTS). IEEE, 2017, pp. 1–6.
- [2] C. S. Kouzinopoulos, G. Spathoulas, K. M. Giannoutakis, K. Votis, vP. Pandey, D. Tzovaras, S. K. Katsikas, A. Collen, and N. A. Nijdam, "Using blockchains to strengthen the security of internet of things," in International ISCIS Security Workshop. Springer, 2018, pp. 90–100.
- [3] M. A. Khan and K. Salah, "Iot security: Review, blockchain solutions, and open challenges," Future Generation Computer Systems, vol. 82, pp. 395–411, 2018.
- [4] Dorri, S. S. Kanhere, R. Jurdak, and P. Gauravaram, "Lsb: A lightweight scala-ble blockchain for iot security and privacy," arXiv preprint arXiv:1712.02969, 2017.
- [5] M. T. Hammi, B. Hammi, P. Bellot, and A. Serhrouchni, "Bubbles of trust: A de-centralized blockchain-based authentication system for iot," Computers & Se- curity, vol. 78, pp. 126–142, 2018.
- [6] V. Buterin et al., "A next-generation smart contract and decentralized applica-tion platform," white paper, 2014.
- [7] C. Machado and A. A. M. Fröhlich, "Iot data integrity verification for cyber- physical systems using blockchain," in 2018 IEEE 21st International Symposium on Real-Time Distributed Computing (ISORC). IEEE, 2018, pp. 83–90.
- [8] Dorri, S. S. Kanhere, R. Jurdak, and P. Gauravaram, "Blockchain for iot secu- rity and privacy: The case study of a smart home," in Pervasive Computing and Communications Workshops (PerCom Workshops), 2017 IEEE International Conference on. IEEE, 2017, pp. 618–623.
- [9] T. He, J. Stankovic, C. Lu and T. Abdelzaher, A SpatioTemporal Communication Protocol for Wireless Sensor Networks, IEEE Transactions on Parallel and Dis- tributed Systems, Vol. 16, No. 10, Oct. 2005, pp. 995-1006.
- [10] M. Huang, J. Li, X. Song, and H. Guo, Modeling Impulsive Injections of Insulin: Towards Artificial Pancreas. SIAM Journal of Applied Mathematics 72, 5, 2012, pp. 1524–1548.
- [11] https://www.google.com/search?q=IOT&sxsrf=AOaemvLb0FiGlnOLo93t-503CCh3zZXVDA:1635499535390&source=lnms&tbm=isch&sa=X&ved=2ahUKEwj4tjGpu_zAhWK7XMBH ai9AvIQ_AUoAnoECAEQBA&biw=1920&bih=937&dpr=1
- [12] https://www.google.com/search?q=smart+cities+intelligent+nation&sxsrf=AOaemvLY_qCZEfbLC16YZq6xQ01d eGl9qw:1635499680794&source=lnms&tbm=isch&sa=X&ved=2ahUKEwjmsOmp-_zAhVRjeYKHa9_ApwQ_AUoA3oECAIQBQ&biw=1920&bih=937&dpr=1#imgrc=wOZ2Q-j6ibDW7M
- [13] https://www.google.com/search?q=smart+house+iot&tbm=isch&ved=2ahUKEwi4qsSyp-_zAhWwGbcAHRPRA3EQ2-cCegQIABAA&oq=smart+house+iot&gs_lcp=CgNpbWcQAzIFCAAQgAQyBggAEAoQGDoHCCMQ7wMQJzo ECAAQQ1CtBFj_DGC_E2gAcAB4AIAB2AGIAbgGkgEFMC40LjGYAQCgAQGqAQtd3Mtd2l6LWltZ8ABA Q&scclient=img&ei=8b57YfjtILCz3LUPk6KPiAc&bih=937&biw=1920#imgrc=eZiiDz7vWLS8zM
- [14] Husamuddin Mohammed, Mohammed Qayyum, "Internet of Things: A Study on Security and Privacy Threats", March 2017.
- [15] Aytaj Badirova, "Security and Privacy in Internet of Things", July 2018.

An Automated Video-Password based Authentication System using Electrocardiogram

Shahina Anwarul
School of Computer Science
University of Petroleum and Energy Studies (UPES)
Bidholi, Dehradun, INDIA 248007
shahinaanwarul@gmail.com

Abstract: In traditional authentication systems, most of the applications were using text and image-based passwords. These passwords can be easily tricked and spoofed by intruders. With the advent of the attacks, the efficient and robust authentication system becomes the need of the hour. The recent developments in the technologies have diverted the interest of researchers from the traditional methods to biometric-based methods for authentication purposes. In recent researches, the electrocardiogram (ECG) is investigated as a significant biometric feature for the identification or verification of an individual. Therefore, the authors proposed a novel video-password authentication system using Electrocardiogram (ECG). The present paper introduces a new class of authentication scheme called “Video-Password using ECG”. This paper deals with the detection of the QRS complex of the ECG signal using the Pan Tompkins algorithm. First, the system captures the ECG signal of the user and calculates the QRS complex of the captured ECG signal. Then, the detected QRS complex is utilized for the division of the video frames that are used for the matching of the stored frames for the authentication process. The authors conducted the experiments by using the ECG signals available in the MIT-BIH Arrhythmia Database <<http://www.physionet.org/physiobank/database/mitdb/>>.

Keywords: Electrocardiogram, Video Password, Authentication, QRS Complex, Automated

1. Introduction

Biometrics plays a significant role in authentication systems. It can be utilized for recognition as well as verification of an individual. In recent researches, a lot of biometric features have been used for authentication systems [1-3]. The characteristics of biometric can be categorized on the basis of behavioral and physiological features of the person as shown in Fig. 1. The behavioral and the physiological external observable characteristics like gait, keystroke, voice, iris, palm, fingerprint, and facial features can be easily duped without checking the liveness of the discussed features [4-10]. The ECG was used as a diagnostic tool for many years but in the last few years a large number of researches have been conducted for the use of the ECG in biometrics and it has been proven to be a potentially useful biometric tool for the identification of an individual [11, 12]. ECG estimates the electrical activity of the heart. A typical ECG tracing of the heart consists of a P wave, a QRS complex, and a T wave. Amongst the three discussed unique features of the ECG tracing, the QRS complex remains constant irrespective of the change of the heart rate. The elementary shape of the ECG signal is shown in Fig. 2 where the P wave reflects the sequential depolarization of the left and right atria whereas ventricular repolarization is contemplated by T wave. The QRS complex represents the depolarization of left and right ventricles [13].

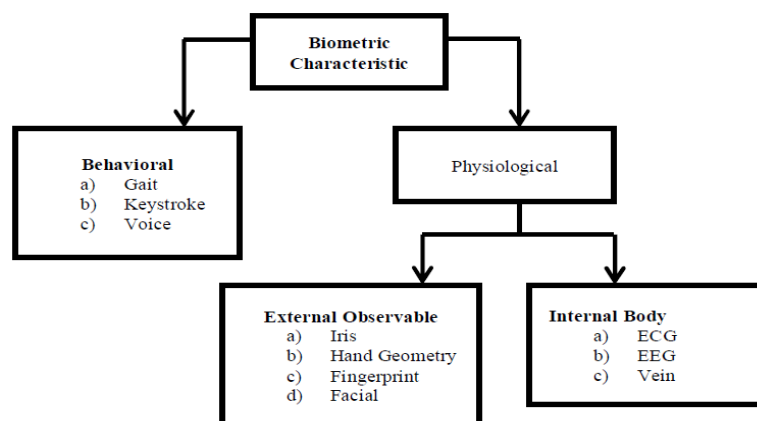


Fig. 1 Classification of Biometric Characteristics

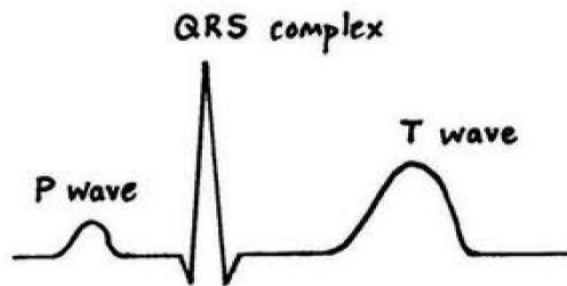


Fig. 2 The Basic Shape of the ECG Signal

The present approach utilizes the QRS complex in a video password authentication system. Traditionally, text and image passwords are used for authentication purposes. But the text passwords can be easily guessed and the image passwords can be stolen by the intruder [14]. These traditional methods do not provide the assurance of security to the individual. The proposed video password-based scheme eradicates these shortcomings by employing a QRS complex with video framing. The major contribution of the proposed research is summarized as:

- The author proposed a novel automated video-password-based authentication system using ECG.
- An alert generation system is also proposed to reduce human intervention.

The structure of the present paper is as follows; the first section confers the overview of the biometric-based security systems and the major contributions of the author in the proposed research. The second section discusses the research done in the addressed task in the present paper. The third section gives a detailed description of the proposed methodology containing various sub-sections to demonstrate the flow of the different modules of the authentication system. The experimental results and discussions are conducted in section 5. Last, the authors concluded the proposed work by providing the future perspective of the proposed research.

2 Related Works

A lot has been done on the video-password-based authentication system. The approaches used in video-password based authentication schemes are based on remembering and knowledge. Thorpe et al. [15] proposed a video password authentication system in which the user needs to memorize the video parts such as movements, scene sequence, or sound of the video. Traditional text passwords are also based on memorability but can be easily guessed and attacked by intruders. The graphical password scheme is proposed by Vaddeti et al. [16] that utilizes the images as passwords. But the used images for the password should also be memorized in the given approach. Anjitha et al. [17] proposed the video-based captcha for the enhanced security services. The attackers having the knowledge of pixels will not be able to retrieve any information of the object due to the movement of the objects. Various unique features of ECG are incorporated in biometric-based researches [18]. The complexity-based approach is also introduced by Chen et al. [19] to deal with the abnormal ECG signals for the biometric-based authentication system. Fiducial points detection in heartbeat is also done by Israel et al. [20] to extract the unique attributes of the ECG signal.

3 Proposed Methodology

The proposed methodology consists of three modules namely database creation, the authentication process, and alert generation. The schematic flow of the modules in the proposed work is illustrated in Fig. 3 and Fig. 5.

3.1 Database Creation

In this module, the video frames are selected on the basis of the generated QRS complex values to be considered as a password. The complete flow of the module is demonstrated in Fig. 3.

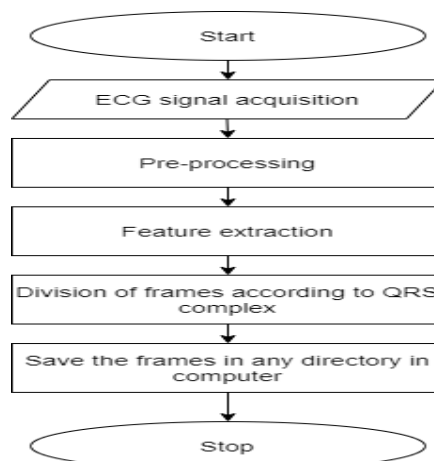


Fig. 3 Database Creation

ECG signal acquisition: The ECG signal of an individual will be captured to calculate the QRS complex. The authors conducted the experiments by utilizing the ECG signal images available in the MIT-BIH Arrhythmia database [21]. Fig. 4 illustrates the one such recording of the database.

Pre-processing: The acquired ECG data consists of various types of noises which corrupt the signals. The authors used band-pass filters to remove such noises.

Feature extraction: The unique features of the signals are extracted after the reduction of noise from the ECG data. There are various unique features of ECG signal are available that are persistent with the heart rate such as QRS complex, QT interval, etc. The authors employed the QRS complex of the ECG signal in the proposed authentication system for the identification of an individual. Pan Tompkins algorithm [22] is used to evaluate the QRS complex of the ECG signal.

Division of video frames according to calculated QRS complex: After the calculation of the QRS complex, the selected video is divided into frames on the basis of the calculated QRS complex value. The divided video frames are stored in a pre-defined location of the computer for the further authentication process.

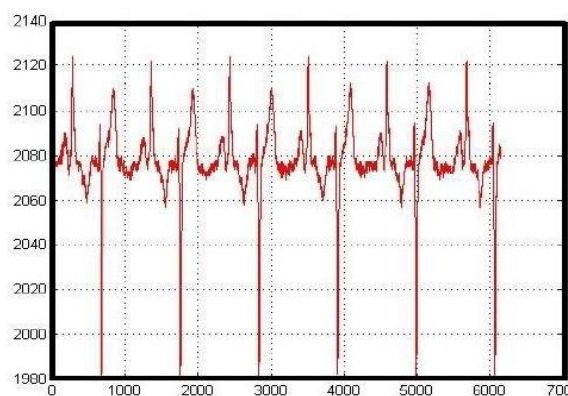


Fig. 4 ECG Waveform Available in MIT-BIH Arrhythmia Database

3.2 Authentication Process & Alert Generation

In the authentication process, the user passes the same video as used for database creation to the authentication system. The system divides the video frames by taking the calculated QRS complex of the ECG signal of an individual. The system matches the generated frames with the available frames that are stored in the database. If both the frames are same, then the user gets access to the system else the request will be rejected. If the attempt is done by the unauthentic user, then the proposed authentication system sends an alert in mail and message to the owner of the system. The complete process of the authentication and alert generation module is illustrated in Fig. 5.

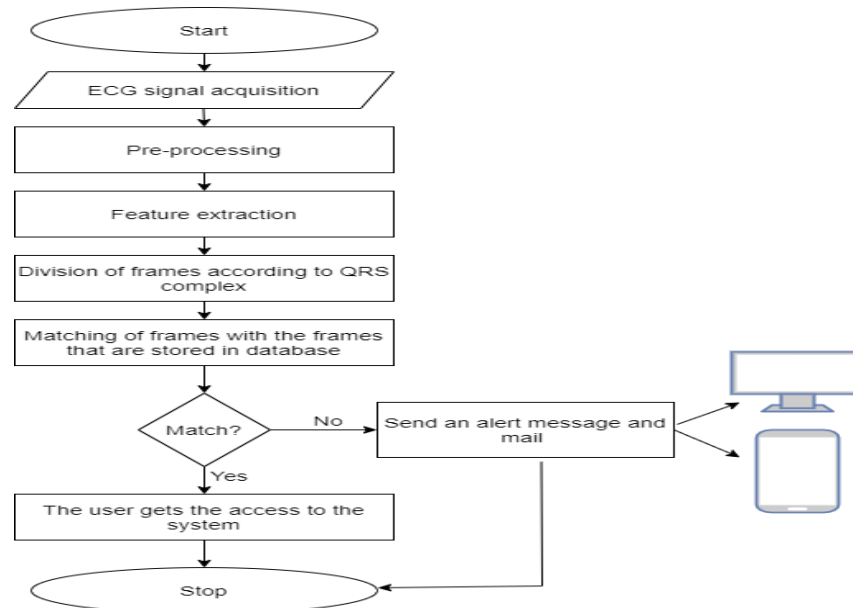


Fig. 5 Authentication and Alert Generation

4 Experimental Results And Discussion

The authors conducted all the experiments in MATLAB using the MIT-BIH Arrhythmia Database. The dataset consists of the ECG waveform images of patients of the age between 47 and 84 years. The present research results in an effective and robust automated authentication system that cannot be tricked and fooled by intruders. It also provides assurance to the user that his/her data is safe even after death. The QRS complex is calculated using the Pan Tompkins algorithm which utilizes the derivatives of the ECG signal.

Derivatives: The three derivatives of the ECG signal are computed using equations (1-3). Assuming the first, second, and cumulative derivatives of the ECG signal be Γ , $\dot{\Gamma}$, and $\S$, respectively, and shown in Fig. 6.

$$\Gamma = \text{diff}(\text{ECG}) \quad (1)$$

$$\dot{\Gamma} = \text{diff}(\Gamma) \quad (2)$$

$$\S = 1.3 * \Gamma(1:6500) + 1.1 * \dot{\Gamma}(1:6500) \quad (3)$$

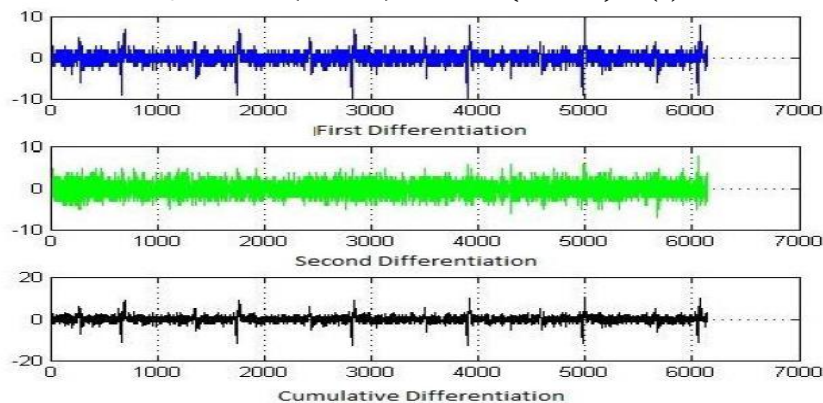


Fig. 6 First, Second, and Third Derivatives of ECG Signal

Setting the threshold value: The authors selected the peak points from the cumulative differentiation of the ECG signal as a threshold value to calculate the QRS complex. Fig. 7 illustrates the detected QRS waveform while Fig. 8 shows the detected QRS complex against time. As nine peak points have been selected at a given time, therefore, the video frames of that particular instance of time will be selected as a password for the authentication process.

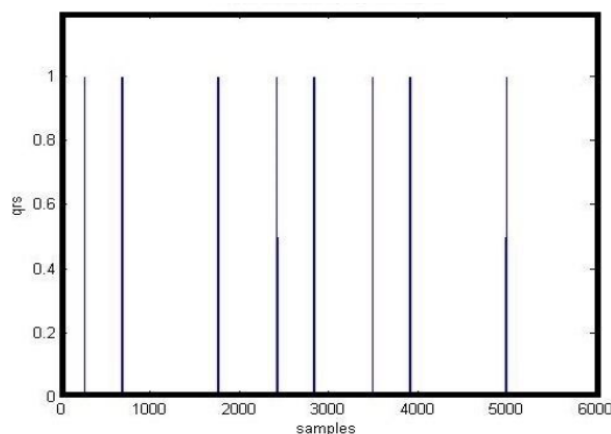


Fig. 7 Detected QRS Waveform

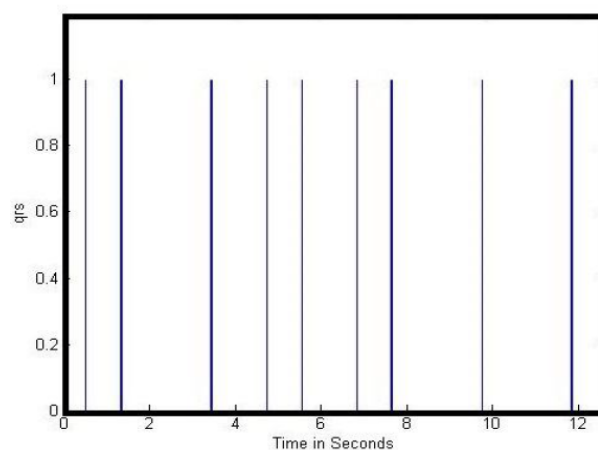


Fig. 8 QRS Complex Against Time

5 Conclusion And Future Perspective

The present paper utilized the uniqueness of the QRS complex of the ECG signal into a video password scheme and provides a new class of authentication system. The proposed research provides a higher class of security system which is impossible to deceive. The analysis on security indicates that video passwords have the potential to be an assuring class of authentication systems. Also, it is proved that the QRS complex of the ECG signal is stable and invariant towards the change of heart rate. The association of these two makes the proposed security system vigorous and sturdy. The proposed system can be more robust by adding the ancillary unique features of the ECG signal. The system can be more pliable by making it compatible with different video parameters like video size and video type.

References

- [1] Xiao, K., Tian, Y., Lu, Y., Lai, Y., & Wang, X. (2021). Quality assessment-based iris and face fusion recognition with dynamic weight. *The Visual Computer*, 1-13.
- [2] Ota, H., Aoyama, S., Watanabe, R., Ito, K., Miyake, Y., & Aoki, T. (2013). Implementation and evaluation of a remote authentication system using touchless palmprint recognition. *Multimedia systems*, 19(2), 117-129.
- [3] Min-Allah, N., Jan, F., & Alrashed, S. (2021). Pupil detection schemes in human eye: a review. *Multimedia Systems*, 1-25.
- [4] Cola, G., Avvenuti, M., Musso, F., & Vecchio, A. (2016, November). Gait-based authentication using a wrist-worn device. In *Proceedings of the 13th International Conference on Mobile and Ubiquitous Systems: Computing, Networking and Services* (pp. 208-217).
- [5] Obaidat, M. S., Krishna, P. V., Saritha, V., & Agarwal, S. (2019). Advances in key stroke dynamics-based security schemes. In *Biometric-Based Physical and Cybersecurity Systems* (pp. 165-187). Springer, Cham.
- [6] Singh, N., Agrawal, A., & Khan, R. A. (2018). Voice biometric: A technology for voice based authentication. *Advanced Science, Engineering and Medicine*, 10(7-8), 754-759.

- [7] Dey, S., & Samanta, D. (2008). Improved feature processing for iris biometric authentication system. *Int. J. Electr. Electron. Eng.* 4, 127-34.
- [8] Al-Juboori, A. M., Wu, X., & Zhao, Q. (2013, December). Biometric authentication system based on palm vein. In *2013 International Conference on Computer Sciences and Applications* (pp. 52-58). IEEE.
- [9] Roy, A., Memon, N., & Ross, A. (2017). Masterprint: Exploring the vulnerability of partial fingerprint-based authentication systems. *IEEE Transactions on Information Forensics and Security*, 12(9), 2013-2025.
- [10] Anwarul, S., & Dahiya, S. (2020). A comprehensive review on face recognition methods and factors affecting facial recognition accuracy. *Proceedings of ICRIC 2019*, 495-514.
- [11] Shen, T. W., Tompkins, W. J., & Hu, Y. H. (2002, October). One-lead ECG for identity verification. In *Proceedings of the second joint 24th annual conference and the annual fall meeting of the biomedical engineering society* [engineering in medicine and biology (Vol. 1, pp. 62-63). IEEE.
- [12] Biel, L., Pettersson, O., Philipson, L., & Wide, P. (2001). ECG analysis: a new approach in human identification. *IEEE Transactions on Instrumentation and Measurement*, 50(3), 808-812.
- [13] Sörnmo, L., & Laguna, P. (2005). *Bioelectrical signal processing in cardiac and neurological applications* (Vol. 8). Academic Press.
- [14] Weir, M., Aggarwal, S., Collins, M., & Stern, H. (2010, October). Testing metrics for password creation policies by attacking large sets of revealed passwords. In *Proceedings of the 17th ACM conference on Computer and communications security* (pp. 162-175).
- [15] Thorpe, J., Salehi-Abari, A., & Burden, R. (2012, September). Video-passwords: Advertising while authenticating. In *Proceedings of the 2012 New Security Paradigms Workshop* (pp. 127-140).
- [16] Vaddeti, A., Vidiyala, D., Puritipati, V., Ponnuru, R. B., Shin, J. S., & Alavalapati, G. R. (2020). Graphical passwords: Behind the attainment of goals. *Security and Privacy*, 3(6), e125.
- [17] Anjitha, K., & Rijin, I. K. (2015, October). Captcha as graphical passwords-enhanced with video-based captcha for secure services. In *2015 International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)* (pp. 213-217). IEEE.
- [18] Porée, F., Gallix, A., & Carrault, G. (2011, September). Biometric identification of individuals based on the ECG. Which conditions?. In *2011 Computing in Cardiology* (pp. 761-764). IEEE.
- [19] Chen, S. W. (2006). Complexity-measure-based sequential hypothesis testing for real-time detection of lethal cardiac arrhythmias. *EURASIP Journal on Advances in Signal Processing*, 2007, 1-8.
- [20] Israel, S. A., Irvine, J. M., Cheng, A., Wiederhold, M. D., & Wiederhold, B. K. (2005). ECG to identify individuals. *Pattern recognition*, 38(1), 133-142.
- [21] Moody, G. B., & Mark, R. G. (2001). The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3), 45-50.
- [22] Sedghamiz, H. (2014). Matlab implementation of Pan Tompkins ECG QRS detector. Code Available at the File Exchange Site of MathWorks.

A Novel Intuitionistic Fuzzy Inference System for Sentiment Analysis

Makrand Dhyani¹, Sanjay Kumar², Govind Singh Kushwaha³

^{1&2} Department of Mathematics, Statistics and Computer Science
G. B. Pant University of Agriculture & Technology, Pantnagar, India

³ Dept. of Social Sciences and Humanities,
G. B. Pant University of Agriculture & Technology, Pantnagar, India
makdhyani23@gmail.com¹ skruhela@hotmail.com²
drgovindsingh@yahoo.co.in³

Abstract. Sentiment analysis (SA) is important part in psychology that helps to predict the attitude or personality trait of human. In the present study, we propose novel intuitionistic fuzzy inference system (IFIS) for the sentiment analysis. The research paper does the sentiment analysis of twitter data and tries to predict the personality trait characteristics of the tweeting individual through proposed IFIS. Twitter data was analysed using Natural Language Processing Toolkit (NLTK) through TextBlob in Google Colaboratory for their subjectivity and polarity to predict personality traits of the tweeting individuals using proposed novel IFIS. Performance of proposed IFIS is compared with mamdani FIS in terms of RMSE. Reduced amount of RMSE confirms outperformance of proposed IFIS over mamdani FIS.

Keywords: Sentiment Analysis, Fuzzy set, Natural Language Processing, Intuitionistic Fuzzy Set.

1. Introduction

Emotions are defined as distinct feelings of consciousness of joy or sadness and have substantial influence on the cognitive process of perception, attention, learning, reasoning and problem solving in humans. Sentiment is an attitude or opinion and is also influenced by emotions very often. SA is a natural language processing technique that determines whether text is neutral, negative or positive. However, many researchers use sentiment analysis and emotional intelligence as synonyms of each other but they are different. While SA depends mainly on information to classify expression as neutral, negative or positive, emotional intelligence further identifies whether information articulated by comment leads to sadness, joy, dissatisfaction. Sentiment analysis is not only important in the study of human behaviour but also being applied in the analysis of feedback and monitoring of a specific consumer product.

In sentimental analysis information from the raw data is extracted and is transformed in an organized and structured way for further analysis.[1] Developed a framework to integrate online social network and business intelligence using analytical tools for the dissemination and analysis of knowledge. Aspect-based, comparative sentiment analysis, sentence-level, document-level, and sentiment lexicon acquisition are the various methods found in the literature of sentiment analysis. Detail and specific problem with aforementioned method of sentiment analysis can be found in the work of [2]. Over the time various techniques have been developed for sentiment analysis. [3] Proposed machine learning based classification algorithms of Semantic Orientation based WordNet, SVM and Maximum entropy along with the Naive Bayes for the sentiment analysis.[4] proposed parts of speech (PoS) tagging and tree kernel based novel approach for the sentiment analysis using twitter data. Recently [5] and [6] proposed methods for SA using deep learning techniques optimized clustering based algorithms, respectively.

From the literature, it has been observed that sentiment analysis is important for the simulation of human behaviour that depends upon the number of parameters. Behaviour pattern recognition is still the unfolded area of Psychology. In Psychology, empirical models are developed using the questionnaire [7–9] for the simulation of human behaviour. Database collected using these questionnaires may include uncertainty and ambiguity due to biasness of the participants [10, 11]. Motivated with this many researchers [12, 13] proposed use of fuzzy set (FS) theory [14] for the modelling of simulation of human behaviour. [15–18] and [19] also developed fuzzy set theory based models for the study of psychological parameters that affects the human behaviour. Uncertainty due to linguistic representation of information makes fuzzy set theory as ideal tool for sentiment analysis. Earlier few models of personality traits were developed that included hostile, realistic, selfish, sensitive, neuroticism, extroversion and fuzzy logic based modelling of emotions [20–22]. [23, 24] defined intuitionistic fuzzy set (IFS) as extension of fuzzy set to model non-determinism represented by “can’t say” kind of situation and were proved more ideal than fuzzy set.

In the present study we propose a novel intuitionistic fuzzy inference system to analyse the attitude of web users using data of a social networking site (SNS) of twitter. Few researchers [25, 26] have also anticipated few psychological criteria using data of SNSs of twitter and facebook with help of questioners.

SNSs and increasing user-generated web are speedily becoming one of the most trendy tools for information exchange and social interaction. Previous studies of [25] and [26] have shown that how users data on web and SNSs can be used to predict their personality and behaviour. Motivated with the applications of IFS in decision making problems, we propose a novel intuitionistic fuzzy inference system to analyze the attitude of web users using their comments on SNS of twitter.

2. Fuzzy Set and Intuitionistic Fuzzy Set

Let $X = (x_i : i = 1 \text{ to } n)$ be a finite crisp reference set. Fuzzy set and IFS on X can be defined as follows.

A FS A on X is a mathematical object of the form $A = \{x, \mu_A(x) : \forall x \in X\}$. Here $\mu_A(x)$ is the grade of membership of x in A and $\mu_A : X \rightarrow [0, 1]$.

An IFS, I on X is a mathematical object defined as $I = \{(x, \mu_I(x), \nu_I(x)) : \forall x \in X\}$ satisfying the condition that for all x in X $0 \leq \nu_I(x) + \mu_I(x) \leq 1$. Here $\nu_I, \mu_I : X \rightarrow [0, 1]$ and $\nu_I(x)$ & $\mu_I(x)$ represent the non-membership and membership grade of x in I . $\pi_I(x) = 1 - \nu_I(x) - \mu_I(x)$ is called the degree of hesitancy/ non-determinacy.

3. Materials and methods

This section explains the material used for the development of IFIS along with the detailed methodology.

• Collection of data and pre-processing

From a site named vicinetas.com, 2000 tweets about the protest of farmer bills in various languages were collected. 100 tweets out of them in English language were selected anonymously after eliminating retweets, links, date, time and other non-meaningful symbols. Selected tweets were analyzed using “TextBlob” tool of a Natural language toolkit (NLTK) for their polarity and subjectivity which are taken as input variables in proposed IFIS. A brief description of polarity and subjectivity associated with tweets in terms of psychology is given as follows.

Polarity: Polarity is usually defined in terms of negative or positive connotations that people have with an individual word [27] Fundamentally, polarity is a mixed construct referring to the fact that in natural language, some words are negative while some are positive.

Subjectivity: Subjectivity of a sentence expresses some personal feelings, views, or beliefs [28, 29].

It is important to note that a purely objective sentence is sentiment free (without any sentiment), while a sentence purely subjective in nature leans towards a negative or a positive sentiment.

• Novel Intuitionistic Fuzzy Inference System

Process of developing an IFIS includes following steps.

1. Intuitionistic fuzzification of input and output data.
2. Construction of intuitionistic fuzzy rule base.

3. Inferencing of intuitionistic fuzzy rules.

4. Intuitionistic defuzzification.

Stepwise procedure of proposed IFIS for prediction of positivity is explained in following steps.

Step 1: Intuitionistic fuzzification. After analyzing tweets using “TextBlob” tool of a natural language toolkit (NLTK), raw scores of polarity (x), subjectivity (y) are computed. Raw score of positivity (z) associated with tweets is computed using “NaiveBayesClassifier” available in “TextBlob”. These raw scores of polarity, subjectivity and positivity are shown in Table 3. Raw scores of polarity, subjectivity and positivity are fuzzified using triangular fuzzy sets A_i ($i = 1$ to 5), B_j ($j = 1$ to 4) and P_k ($k = 1$ to 4), respectively with following parameters.

Table1: Parameters for triangular fuzzy sets for polarity, subjectivity and positivity.

S.No	Poalrity	Subjectivity	Positivity
1	$A_1 = [-1 -1 0.5]$	$B_1 = [0 0 .0.5]$	$P_1 = [0 0 0.5]$
2	$A_2 = [-1 -0.5 0]$	$B_2 = [0.25 0.5 0.75]$	$P_2 = [0.25 0.5 0.75]$
3	$A_3 = [-0.5 0 0.5]$	$B_3 = [0.5 0.75 1]$	$P_3 = [0.5 0.75 1]$
4	$A_4 = [0 0.5 1]$	$B_4 = [0.75 1 1]$	$P_4 = [0.75 1 1]$
5	$A_5 = [0.5 1 1]$		

Fuzzy sets A_i ($i = 1$ to 5), B_j ($j = 1$ to 4) and P_k ($k = 1$ to 4) are converted into the IFSs IA_i ($i = 1$ to 5), IB_j ($j = 1$ to 4) and IP_k ($k = 1$ to 4), respectively using following conversion method of [30].

Let A be the FS on X and mapping, $f : [0,1]^2 \times [0,1] \rightarrow L$ given by

$$f(x, \pi, \delta) = (f_\mu(x, \pi, \delta), f_\nu(x, \pi, \delta))$$

Where,

$$f_\mu(x, \pi, \delta) = x(1 - \delta\pi)$$

$$f_\nu(x, \pi, \delta) = 1 - x(1 - \delta\pi) - \delta\pi$$

Satisfies following.

- 1) If $y_1 \leq y_2$ then $\pi(f(x, y_1, \delta)) \leq (f(x, y_2, \delta))$ for all $x, \delta \in [0,1]$
- 2) $f_\mu(x, \pi, \delta) \leq x \leq 1 - f_\nu(x, \pi, \delta)$ for all $x \in [0,1]$
- 3) $f(x, 0, \delta) = (x, 1 - x)$
- 4) $f(0, y, \delta) = (0, 1 - \delta y)$
- 5) $f(x, y, 0) = (x, 1 - x)$
- 6) $\pi(f(x, y, \delta)) = \delta y$

Let $A \in FSS(X)$, ($FSS(X)$ is collection of fuzzy sets over X). Let $\pi, \delta : X \rightarrow [0,1]$ be two mappings.

Then $A = \{(x_i, f(\mu_{A_F}(x_i), \pi(x_i), \delta(x_i))) \mid x_i \in X\}$

Construction of IFS from a FS is demonstrated in the example below.

Example: Consider a reference Fuzzy Set $A = \{(u_1, 0.5), (u_2, 1), (u_3, 0.8), (u_4, 0.2)\}$. Choose, δ and π for every μ_i such that:

$$\delta(\mu_i) = \text{maximum}(\mu_i; i = 1 \text{ to } 4) = 1.0$$

$$\pi(\mu_i) = \text{minimum}(\mu_i; i = 1 \text{ to } 4) = 0.2$$

Degree of membership for u_1 is computed as follows:

$$f_\mu(\mu(u_1), \pi(u_1), \delta(u_1)) = f_\mu(0.5, 0.2, 1.0) = 0.5(1 - 1.0 \times 0.2) = 0.4$$

Degree of non-membership for u_1 is computed as follows:

$$f_v(\mu(u_1), \pi(u_1), \delta(u_1)) = f_v(0.5, 0.2, 1.0) = 1 - 0.5(1 - 1.0 \times 0.2) - 0.2 \times 1.0 = 0.4$$

Membership and non-membership grade for other u_i are also calculated in similar way and IFS, $I = \{(u_1, 0.4, 0.4), (u_2, 0.8, 0.0), (u_3, 0.64, 0.16), (u_4, 0.16, 0.64)\}$ is obtained corresponding to fuzzy set A.

Step 2: Intuitionistic fuzzy rule base and their inferencing.

Proposed IFIS uses the rule base which consists of IF-Then rules. IF-THEN rules formulate the conditional statements that comprise intuitionistic fuzzy logic. Since numbers of IFSs for polarity (x), subjectivity (y) and positivity (z) are 5, 4 and 4 respectively, total 80 rules of the following form can be constructed.

$$\text{If } x \text{ is } IA_i \text{ and } y \text{ is } IB_j \text{ then } z \text{ is } IP_k \text{ (} i = 1 \text{ to } 5, j = 1 \text{ to } 4, k = 1 \text{ to } 4) \Leftrightarrow [IA_i, IB_j : IP_k]$$

In the research paper, intuitionistic fuzzy rule is represented as $(IA_i, IB_j : IP_k)$. Set of rules is constructed by studying the available data to carry out a relationship between the input and output variables. The intuitionistic fuzzy rules that are used in the present study are given in following Table.

Table2: Intuitionistic fuzzy rule base matrix

R₁	(IA ₃ , IB ₄ : IP ₃)	R₉	(IA ₄ , IB ₁ : IP ₄)
R₂	(IA ₃ , IB ₁ : IP ₄)	R₁₀	(IA ₅ , IB ₄ : IP ₂)
R₃	(IA ₂ , IB ₂ : IP ₄)	R₁₁	(IA ₃ , IB ₃ : IP ₂)
R₄	(IA ₃ , IB ₂ : IP ₃)	R₁₂	(IA ₄ , IB ₄ : IP ₂)
R₅	(IA ₂ , IB ₃ : IP ₄)	R₁₃	(IA ₅ , IB ₁ : IP ₄)
R₆	(IA ₃ , IB ₃ : IP ₂)	R₁₄	(IA ₅ , IB ₂ : IP ₃)
R₇	(IA ₄ , IB ₂ : IP ₁)	R₁₅	(IA ₂ , IB ₁ : IP ₂)
R₈	(IA ₅ , IB ₄ : IP ₁)		

Conflict resolution decides what control action should be taken as the result of the firing of several rules.

The following example illustrates inference process of multiple rules. Let us assume that following two rules are used:

Rule 1: [IA₁, IB₁: IP₁]

Rule 2: [IA₂, IB₂: IP₂]

Now, suppose if x_0 and y_0 are input values for x and y with their membership and non-membership grades as $\mu_{IA_1}(x_0), \mu_{IB_1}(y_0)$ and $\nu_{IA_1}(x_0), \nu_{IB_1}(y_0)$ respectively for **Rule 1**. Similarly for **Rule 2**, ν_{IA_2}, ν_{IB_2} and μ_{IA_2}, μ_{IB_2} are the grades of non-membership and membership respectively.

The combined output of **Rule 1** and **Rule 2** is computed by applying the following matching strengths α_1 & α_2 .

$$\alpha_1 = \frac{\max(\mu_{IA_i}(x_o), \mu_{IB_j}(y_o)) + \min(\mu_{IA_i}(x_o), \mu_{IB_j}(y_o))}{2}, i, j = 1, 2$$

$$\alpha_2 = \frac{\max(v_{IA_i}(x_o), v_{IB_j}(y_o)) + \min(v_{IA_i}(x_o), v_{IB_j}(y_o))}{2}, i, j = 1, 2$$

Step 4: Intuitionistic Defuzzification.

Defuzzification is process of producing a quantifiable result and maps an intuitionistic fuzzy number to a crisp value. In proposed IFIS, degree of non-determinacy is included in defuzzification process. If **Rule 1**: $[IA_1, IB_1: IP_1]$ and **Rule 2**: $[IA_2, IB_2: IP_2]$ are fired for some input, then defuzzified crisp output is computed using the following algorithm.

For $i = 1$ to n (number of distinct IFSs in rules fired for inputs)

Construct the set of elements $\tilde{o} = \{z : \mu_{IP_i}(z) \geq \alpha_1 \text{ and } v_{IP_i}(z) \leq \alpha_2, \forall z \in IP_i\}$

Compute following

$$Z_m = \frac{\sum_{j=1}^{\text{Number of elements in } \tilde{o}} z_j \left(\mu(z_j) + \frac{\pi_j(z_j)}{2} \right)}{\sum_{j=1}^{\text{Number of elements in } \tilde{o}} \left(\mu(z_j) + \frac{\pi_j(z_j)}{2} \right)}$$

$$Z_{nm} = \frac{\sum_{j=1}^{\text{Number of elements in } \tilde{o}} z_j \left(v(z_j) + \frac{\pi_j(z_j)}{2} \right)}{\sum_{j=1}^{\text{Number of elements in } \tilde{o}} \left(v(z_j) + \frac{\pi_j(z_j)}{2} \right)}$$

$$Z_i = \frac{Z_m + Z_{nm}}{2}$$

End for

$$\text{Compute defuzzified output as } Z = \frac{\sum_{i=1}^n Z_i}{i}$$

Sample Calculation for determining positive score of a participant who tweeted following text in protest of farmer bill in India is as follows.

“Meet the veer who has been doing sewa - “Free Stitching Service” at #SinghuBorder. He hardly speaks and rarely leaves his tent. I have seen him stitching even at 1 am. He has dharmic songs playing every day after Rehras Sahib and spends most of his time in silence #FarmersProtest ”.

After analyzing above tweet using “TextBlob” of NLTK, Polarity and subjectivity of above tweet are computed as 0.1 and 0.4 and are used as inputs to predict the score of positivity of the participant who posted the tweet. Polarity lies in IFS IA_3 and IA_4 while subjectivity lands in IFS IB_1 and IB_2 . Values of α_1 and α_2 are calculated as follows.

$$\alpha_1 = \{\max(0.6767, 0.1641, 0.1742, 0.5878) + \min(0.6767, 0.1641, 0.1742, 0.5878)\} / 2 = 0.42$$

$$\alpha_2 = \{\max(0.169, 0.6566, 0.68568, 0.38572) + \min(0.169, 0.6566, 0.68568, 0.3857)\} / 2 = 0.43$$

Following are corresponding rules fired by the intuitionistic inference engine for the scores of polarity and subjectivity taken in the sample computation.

Rule 1: (IA₃, IB₁: IP₄)

Rule2: (IA₃, IB₂: IP₃)

Rule 3: (IA₄, IB₁: IP₄)

Rule4: (IA₄, IB₂: IP₁)

Now for IFS P₁ we choose all the elements which have their degree of membership value greater than or equal to α_1 i.e. 0.42 and all the elements which have their non-membership grade less than or equal to α_2 i.e. 0.43. We calculate z_1 and z_2 using algorithm mentioned in step4 which were found out to be 0.1750 & 0.19340 respectively. Hence $Z_1 = 0.1842$. Similarly we calculate Z_2 & Z_3 for IP₃ and IP₄ which were found out to be 0.7578 and 0.9479 respectively. Hence the final crisp output is Z that is the required personality score of the individual is

$$Z = \frac{Z_1 + Z_2 + Z_3}{3} = 0.6343$$

Similar calculations were done for other pairs of polarity and subjectivity and the results of are presented in Table 3.

4. Results and Discussion

After following the computations from step 1 to step 4 of section 3, scores of positivity of the participants are computed using polarity and subjectivity associated with their tweets. Raw scores of polarity and subjectivity which were collected using NLTK are intuitionistic fuzzified using triangular IFSs. Parameters and rule based used in proposed IFIS are shown in Table 1 and Table 2. Predicted scores of positivity using Mamdani FIS and proposed novel IFIS are shown in Table 3.

Table 3: Predicted score of positivity using Mamdani FIS and proposed IFIS.

S.No	Polarity	Subjectivity	Predicted Score of positivity	
			FIS	IFIS
1	0	0	0.92	0.9963
2	0.25	0.2	0.905	0.961055
3	0.1	0.4	0.599	0.6343
4	-0.015	0.205	0.909	0.8047
5	-0.5	0.5	0.644	0.96
6	0.141	0.66833	0.495	0.63795
7	0	0.75	0.5	0.5079
8	0.147	0.74107	0.499	0.612
9	0.552083	0.40833	0.326	0.68482
10	0.119	0.3095	0.545	0.621
11	0.15	0.55	0.552	0.73294
12	0.2	0.8	0.567	0.6594
13	0.15	0.35	0.53	0.876325
14	0.095238095	0.1785	0.912	0.96463
15	0.119047619	0.3095	0.751	0.87784
16	0.225	0.425	0.495	0.6548
17	0.29765625	0.64166	0.469	0.73286
18	0.5	0.5	0.163	0.22822
RMSE			0.399388	0.307847

Performance of proposed IFIS is compared using root mean square error (RMSE). Proposed novel IFIS does not only handle non-determinacy associated with tweets but also predicts score of positivity more accurately than Mamdani FIS.

5. Conclusion

This study proposes an IFIS for sentiment analysis to analyze the attitude of the participants using linguistic information provided by them. The proposed IFIS is knowledge-driven predictive model that is not very common for the study of sentiment analysis, but can be very useful to overcome subjective indistinctness and doubts in emotional intelligence. The major advantage of proposed IFIS is that it enables the use of uncertainty measures along with non-determinacy to quantify the ambiguity associated with linguistic information provided by the participants. Since IFISs are considered more ideal than fuzzy set for the modelling of non-determinacy in the system, proposed IFIS has an edge over fuzzy rule based and other prediction models used for sentiment analysis. Proposed IFIS is efficient in handling uncertainty along with non-determinacy and produces more rationale score of the positivity than NLTK. Reduced amount of RMSE involved in the prediction of personality score also confirms the outperformance of proposed IFIS over FIS

References:

- [1] Costa, PRS, FF Souza, ... VC Times - Proceedings of the, and Undefined 2012. 2012. Towards integrating online social networks and business intelligence. *researchgate.net*.
- [2] Feldman, Ronen. 2013. Techniques and applications for sentiment analysis: The main applications and challenges of one of the hottest research areas in computer science. *Communications of the ACM* 56: 82–89. <https://doi.org/10.1145/2436256.2436274>.
- [3] Gautam, Geetika, and Divakar Yadav. 2014. Sentiment analysis of twitter data using machine learning approaches and semantic analysis. *2014 7th International Conference on Contemporary Computing, IC3 2014*. Institute of Electrical and Electronics Engineers Inc.: 437–442. <https://doi.org/10.1109/IC3.2014.6897213>.
- [4] Sahayak, V, V Shete, ... A Pathan - Research in Advanced Engineering, and undefined 2015. 2015. Sentiment analysis on twitter data. *academia.edu* 1.
- [5] Chakraborty, Koyel, Surbhi Bhatia, Siddhartha Bhattacharyya, Jan Platos, Rajib Bag, and Aboul Ella Hassanien. 2020. Sentiment Analysis of COVID-19 tweets by Deep Learning Classifiers—A study to show how popularity is affecting accuracy in social media. *Applied Soft Computing* 97. Elsevier: 106754. <https://doi.org/10.1016/J.ASOC.2020.106754>.
- [6] Suhaib Bin Younis, et. al. 2021. Opinion Mining On Web-Based Communities Using Optimised Clustering Algorithms. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)* 12: 438–447. <https://doi.org/10.17762/TURCOMAT.V12I9.3099>.
- [7] Eysenck, H. J. 1959. Learning Theory and Behaviour Therapy. *Journal of Mental Science* 105. Cambridge University Press: 61–75. <https://doi.org/10.1192/BJP.105.438.61>.
- [8] Price, Donald D., James E. Barrell, and James J. Barrell. 1985. A quantitative-experiential analysis of human emotions. *Motivation and Emotion* 9. Kluwer Academic Publishers-Plenum Publishers: 19–38. <https://doi.org/10.1007/BF00991548>.
- [9] AKP Sinha. 1995. Manual for sinha's comprehensive anxiety test. *Agra: National Psychological Corporation*.
- [10] Smithson, Michael. 1982. Applications of fuzzy set concepts to behavioral sciences. *Mathematical Social Sciences* 2. North-Holland: 257–274. [https://doi.org/10.1016/0165-4896\(82\)91081-2](https://doi.org/10.1016/0165-4896(82)91081-2).
- [11] Smithson, Michael. 1988. Fuzzy set theory and the social sciences: The scope for applications. *Fuzzy Sets and Systems* 26. North-Holland: 1–21. [https://doi.org/10.1016/0165-0114\(88\)90002-4](https://doi.org/10.1016/0165-0114(88)90002-4).
- [12] Oren, T. I., & Ghasem-Aghaee, N. 2003. Personality representation processable in fuzzy logic for human behavior simulation. *Citeseer*: 11–18.
- [13] Chen, J. Y. 2005. A study on college students anxiety of career decision. *Educ Psycho* 28: 745–771.
- [14] Zadeh, Lotfi A. 1965. Fuzzy set, Information and Control. In , ed. Zadeh LA, 338–353. https://doi.org/10.1142/9789814261302_0021.
- [15] Kushwaha, GS, S Kumar - Europe's Journal of Psychology, and undefined 2009. 2009. Role of the fuzzy system in psychological research. *ejop.psychopen.eu* 2: 123–134.
- [16] Stoklasa, Jan, Tomáš Talášek, and Jana Musilová. 2014. Fuzzy approach - a new chapter in the methodology of psychology? *Human Affairs* 24. Slovak Academy of Sciences - Social Biological Comm: 189–203. <https://doi.org/10.2478/S13374-014-0219-8/PDF>.
- [17] Devi, Satwanti, Sanjay Kumar, and Govind Singh Kushwaha. 2016. An adaptive neuro fuzzy inference system for prediction of anxiety of students. *Proceedings of the 8th International Conference on Advanced Computational Intelligence, ICACI 2016*. Institute of Electrical and Electronics Engineers Inc.: 7–13. <https://doi.org/10.1109/ICACI.2016.7449795>.
- [18] Srivastava, Shilpa, Millie Pant, and Namrata Agrawal. 2016. Psychology of adolescents: a fuzzy logic analysis.

- International Journal of System Assurance Engineering and Management* 2016 9:1 9. Springer: 66–81. <https://doi.org/10.1007/S13198-016-0472-9>.
- [19] Pandey, Deepak Chandra, Govind Singh Kushwaha, and Sanjay Kumar. 2020. Mamdani fuzzy rule-based models for psychological research. *SN Applied Sciences* 2020 2:5 2. Springer: 1–10. <https://doi.org/10.1007/S42452-020-2726-Z>.
- [20] El-Nasr, Magy Seif, John Yen, and Thomas R. Ioerger. 2000. FLAME—Fuzzy Logic Adaptive Model of Emotions. *Autonomous Agents and Multi-Agent Systems* 2000 3:3 3. Springer: 219–257. <https://doi.org/10.1023/A:1010030809960>.
- [21] Rousseau, Daniel. 1996. Personality in Computer Characters. *AAAI Workshop on Entertainment and AI / A-Life*: 38–43.
- [22] Smithson, Michael, and Gregg C. Oden. 1999. Fuzzy Set Theory and Applications in Psychology. Springer, Boston, MA: 557–585. https://doi.org/10.1007/978-1-4615-4601-6_17.
- [23] Atanassov. 1986. K.T. Atanassov, Intuitionistic fuzzy sets. *Fuzzy sets Syst.* 20: 87–96.
- [24] Atanassov, KT. 1999. Intuitionistic fuzzy sets. *Springer*.
- [25] Correa, Teresa, Amber Willard Hinsley, and Homero Gil de Zúñiga. 2010. Who interacts on the Web?: The intersection of users' personality and social media use. *Computers in Human Behavior* 26. Pergamon: 247–253. <https://doi.org/10.1016/J.CHB.2009.09.003>.
- [26] Hughes, David John, Moss Rowe, Mark Batey, and Andrew Lee. 2012. A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage. *Computers in Human Behavior* 28. Pergamon: 561–569. <https://doi.org/10.1016/J.CHB.2011.11.001>.
- [27] Van Tiel, Bob, and Elizabeth Pankratz. 2021. Adjectival polarity and the processing of scalar inferences. *Glossa* 6. Ubiquity Press. <https://doi.org/10.5334/GJGL.1457/>.
- [28] Liu, Bing. 2012. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies* 5. Morgan and Claypool Publishers: 1–184. <https://doi.org/10.2200/S00416ED1V01Y201204HLT016>.
- [29] Sindhu, C., Binoy Sasmal, Rahul Gupta, and J. Prathipa. 2021. Subjectivity detection for sentiment analysis on Twitter data. *Lecture Notes in Networks and Systems* 130. Springer: 467–476. https://doi.org/10.1007/978-981-15-5329-5_43.
- [30] Jurio, A., D. Paternain, H. Bustince, C. Guerra, and G. Beliakov. 2010. A construction method of Atanassov's intuitionistic fuzzy sets for image processing. In *2010 IEEE International Conference on Intelligent Systems, IS 2010 - Proceedings*. <https://doi.org/10.1109/IS.2010.5548390>.

Leveraging Machine Learning for Ransomware Detection

Nanda Rani¹ and Sunita Vikrant Dhavale²

Defence Institute of Advanced Technology, Pune, India

¹ nanda_mcse19@diat.ac.in

² sunitadhavale@diat.ac.in

Abstract. The current pandemic situation has increased cyber-attacks drastically worldwide. The attackers are using malware like trojans, spyware, rootkits, worms, ransomware heavily. Ransomware is the most notorious malware, yet we didn't have any defensive mechanism to prevent or detect a zero-day attack. Most defensive products in the industry rely on either signature-based mechanisms or traffic-based anomalies detection. Therefore, researchers are adopting machine learning and deep learning to develop a behaviour-based mechanism for detecting malware. Though we have some hybrid mechanisms that perform static and dynamic analysis of executable for detection, we haven't any full proof detection proof of concept, which can be used to develop a full proof product specific to ransomware. In this work, we have developed a proof of concept for ransomware detection using machine learning models. We have done detailed analysis and compared efficiency between several machine learning models like decision tree, random forest, KNN, SVM, XGBoost and Logistic Regression. We obtained 98.21% accuracy and evaluated various metrics like precision, recall, TP, TN, FP, and FN.

Keywords: Ransomware, Machine Learning, Malware, Cyber security.

1. Introduction

The US colonial pipeline was recently attacked by ransomware named DarkSide and paid \$ 4.4 million ransom, which was requested in ransom note [1]. The attackers stole approx. 100GB data and threatened to leak if the ransom was not being paid. Due to the attack, the country faced fuel shortages and disruption in oil-related services. Nowadays, Attackers are primarily targeting critical infrastructure to disrupt the services that can harm and affect a country population or even a human life if healthcare is being targeted. In history, ransomware has also compromised health care, which causes even human life in danger. Not only critical infrastructure but some giant companies are also being targeted by ransomware, and they are pressured to pay the ransom to get their system access and data back. Attackers generally encrypt victim's data with their key, and they return the key or decrypt the data after victims pay them demanded ransom. Earlier, ransomware was only disrupting the services and making the data unavailable to the victim; they were still safely stored only on their system. But nowadays, data is also being stolen by attackers, threatening the victim to pay the ransom. Though it is not guaranteed that data can get back after paying the ransom or they didn't leak the stolen data.

Ransomware is a class of malware that encrypts the victim data and make them not accessible to the victim. An attacker generally displays a ransom note after encrypting the victim data, which usually states that the attack has been performed and demanded ransom. The steps to make the payment they also provided with that ransom note only. Some of the attackers also create a negotiation window for ransom negotiation. Ransoms are typically demanded in bitcoins because of their anonymity, so attackers cannot get tracked from payment gateway [2]. Ransomware is categorized into two types [3]: (1) Locker ransomware, (2) Crypto ransomware.

Locker ransomware only locks the system, and the victim won't be able to use the system. Ransom notes usually displayed by attackers consist of payment information. This attack usually happens by locking the victim's desktop or displaying a ransom note that never closes. Victim data is not compromised in this attack, and in some cases, there is a recovery option to restart the system in safe mode and perform countermeasures. Sometimes locker ransomware, also known as Screen-Locker ransomware [4]. Crypto ransomware encrypts the user files and adds their extension to the victim files, which makes them inaccessible to use. In some cases, data is also being stolen during communication between the victim system and command control. After encryption, a ransom note usually displays consisting of payment information and demanded ransom. Sometimes crypto-ransomware is also known as File-Locker ransomware [4]. If we think from the CIA triad perspective, availability compromises in any locker ransomware attack. But in the case of crypto-

ransomware, Integrity and availability compromise typically, and in some cases, confidentiality also compromises.

Numerous ransomware variants have been introduced yet, such as WannaCry [9], Cryptowall [5], Gancrab [6], REvil [10], Jigsaw [11], Ryuk [12], Petya [13], Cryptolocker [7], Teslacrypt [8]. WannaCry ransomware attack was a notorious attack that occurred worldwide in 2017 and affected 150 countries [14]. Many healthcare services and hospitals were affected during this attack period. Attackers have exploited the EternalBlue vulnerability in the SMB (Server Message Block) protocol of Windows frameworks [15]. Wannacry, Cryptolocker ransomware came into the picture in 2013, and it was the highest ransom of its time and infected millions of windows systems [9]. Like cyber kill chain process, One of the Exabeam threat research report generalized ransomware attack flow kill chain process in six stages shown in Fig. 26 [16]:

1. *Distributed Campaign*: Victims are tricked by the attacker to download their dropper code using various techniques such as social engineering.
2. *Malicious Code Injection*: Dropper code download and ransomware executable introduce to the victim machines.
3. *Malicious Payload Staging*: Insert ransomware into the victim system and create a persistence connection with the command control server.
4. *Scanning*: Compromised system and network reachable resources being scanned by installed executables for encryption.
5. *Encryption*: All targeted data will be encrypted in this stage.
6. *Payday*: After encryption, a ransom note will be displayed, and data will be inaccessible to use for the victim.

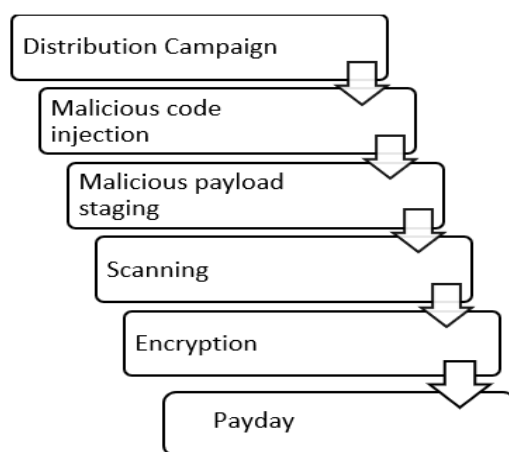


Fig. 26. Ransomware attack flow kill-chain stages

Encryption techniques used by ransomware are (1) Symmetric encryption, (2) Asymmetric encryption, and (3) Hybrid encryption. Attackers use a single key for encryption and decryption in symmetric encryption. They share a common key while communicating with the C&C server or encode the key with dropper code. Forensic experts can extract the common key if it is encoded with dropper code. While attackers use different keys for encryption and decryption in asymmetric keys, it would be challenging to extract them using forensic methods. But communication made during key distribution between the victim system and C&C server can be captured by monitoring traffic. But it fails when attackers use proxy servers to communicate with a compromised system. Afterwards, attackers started using a hybrid encryption technique, in which files would be encrypted by symmetric key only, but another asymmetric key will encrypt that symmetric key. So, if forensic teams extract the symmetric key, they won't be able to use it as it is encrypted with asymmetric encryption, whose decryption is with the attacker only.

The attacker also uses various obfuscation techniques to evade themselves in the system [17]. Malware tries to check inconsistencies in hardware like number of CPU cores to investigate sandbox existence. If the malware finds its execution in the sandbox environment, they hide their malicious indication or behaviours during analysis. Nowadays, most sandbox merchants try to hide their genuine configuration to trick the malware into confidence that it is being executed within the actual host. Several static analysis tools can unpack the packed malware and then unpacked malware can be used for analysis.

2. Literature Review

To analyze ransomware, we have different ways to evaluate any executables: (1) Static Analysis - In this method, executables are being examined without running them, and it is useful to detect malicious infrastructure, library or packed files. Technical indicators are recognized, like file name, hash, a string such as IP address, domain name, and file's header data, which can get utilized to decide whether that file is malicious or not. In addition, tools like disassemblers and network analyzers can spot the malware without actually running it to collect evidence on how that executable malware works. (2) Dynamic Analysis: - In dynamic malware analysis, it executes suspected malicious code or programs in a safe environment called a sandbox. This closed system allows security professionals to investigate the malware in action without the risk of letting it infect the targeted system or escape into the enterprise network. Cuckoo sandbox is the most useful tool when it comes to analyzing malware dynamically. (3) Hybrid Analysis: It delivers the security team the best of both approaches by merging static and dynamic analysis techniques. It can detect malicious programs that hide themselves and then extract many indicators of compromise (IOCs) by statically and previously unseen programs. It also helps to detect unknown threats, even those from the utmost sophisticated malware.

The signature-based protection works on the concept of having a signature covering the families of ransomware. We can consider a signature as a "fingerprint". The detection system matches files signatures from the stored signature database. This system works relatively well for known malware but does not cover threats they know about, i.e., zero-day threats. Abnormal traffic detection is considered as one step up from signature-based detection. Anomalous traffic is detected based on many different parameters like network intrusion detection (NID) and other traffic detections recognized as malicious. The main shortcoming of using the abnormal traffic detection method is its high false-positive rate. Hence, there is a decent chance that genuine network traffic can get identified inaccurately. Thus, we need to monitor process activity and the host's backend operations to decide ransomware infection presence. Due to the emerging Machine Learning technology, it's proving a vital role in monitoring endpoint properties and identifying patterns of system activities for any malware. When machine learning is implemented in file behaviour detection in any system, this can create an effective solution for detecting ransomware. Machine Learning (ML) is similar to human learning in a sense. Legitimate program execution and application present a specific type of behaviour. Over time, ML "learns" how genuine, regular programs act by taking in enormous amounts of data facts through dedicated investigation that may include interactive debugging. In this detailed and inclusive investigation of legitimate program execution, ML develops very good at recognizing programs that are impersonators and have malicious intentions.

There are different methods and techniques to detect ransomware [18,19,20]. Static analysis-based methods disassemble source code without executing it. But they suffer from high false-positive rates and fails to detect obfuscated ransomware. New variants are coming up frequently, and attackers use different packing techniques to mutate their codes. Researchers are moving towards dynamic behaviour analysis-based methods that monitor the interactions of the executed code in a virtual environment for fighting these issues. However, these detection methods are slower and require considerable memory resources. The machine learning approach is best suited for the analysis of the behaviour of any process or application.

Various researchers have focused on either host-based artefact analysis or network-based artefacts analysis to develop an effective defensive mechanism. Host-based artefacts like process behaviour, file system activities, API calls, I/O request packets logs, Registry Key Operations and many more are usually considered to monitor. To monitor network-level artefacts traffic can get monitored when the compromised system communicates with the C&C server. The network-level artefacts such as source and destination IP address, protocol type, source and destination port number, the total number of bytes per packet per conversation can be considered for analysis. In the network-based analysis, there is an assumption that malware will communicate with their C&C server for purposes like getting an encryption key. But this technique may fail if malware doesn't contact C&C servers or used any local key. Also, malware might use proxy servers while contacting C&C servers, and then the network-based analysis may not be feasible to detect. Researchers focused on process anomaly detection to enhance their detection rate. The technique discussed in the paper uses Windows API calls [21,22,23], I/O request Packets (IRP) logs, File system operations, set of operations performed per file extension, directories operations, dropped files, registry key operation, strings [24,25,26,27,28] for detection. Researchers recorded artefacts like folder listing, Files written, Files Renamed, files read, write entropy, and _le type coverage as file system activities [27]. The various

researchers recorded IRP open, IRP write, IRP create for IRP logs [26]. Chen et al. [29] generated API CFG (Call Flow graph) by API sequence extracted from the API monitor tool. They converted CFG into a feature vector for further processing.

3. Implementation

Our work has been carried out with a publicly available dataset prepared by Sgandurra et al. [24]. The dataset includes 582 ransomware samples and 942 goodware samples. Dataset authors have collected good ware and ransomware samples from various sources and used Cuckoo sandbox for analyzing them. Goodware samples include browsers, drivers, multimedia tools, file and network utilities, etc. Ransomware includes its various families tabulated in Table 12.

This dataset consists of a total 1524 records and 30970 features [24]. Features are set of various system events and activities, and they are identified with codes tabulated in

Table 13. Feature values are in binary form, representing whether that event occurred during ransomware and goodware executables execution or not. If that event has occurred during execution, then the value of that feature column is 1 for that specific sample; otherwise, the value for that feature column is 0 for that specific sample. Like this, we have obtained a collection of binary values as a data frame for further analysis.

Table 12 Ransomware family names with correspondence ID

Family Name	ID
Goodware	0
Critroni	1
CryptLocker	2
CryptoWall	3
KOLLAH	4
Kovter	5
Locker	6
MATSNU	7
PGPCODER	8
Reveton	9
TeslaCrypt	10
Trojan-Ransom	11

Table 13 Features ID with description

ID	Description
API	API invocations
DROP	Extensions of the dropped files
REG	Registry key operations
FILES	File operations
FILES_EXT	Extension of the files involved in file operations
DIR	File directory operations
STR	Embedded strings

Each column value of the dataset is mapped with several features shown in the above table, such as API invocation, registry key and _les operations (write, read, create, delete), dropped _le information, embedded strings, etc.

Feature Selection

The next we have followed is feature selection. As we had 30970 features, we have implemented a Correlation-based Feature Selection (CFS) technique to select relevant features from a set of features. CFS evaluates the worth of a subset of attributes by considering each feature's predictive ability and the degree of redundancy between them. Subsets of features that are highly correlated with the class while having low intercorrelation are preferred. So, we decided to use the Mutual Information (MI) feature selection method. It is a more robust option for correlation estimation, which measures mutual dependence between variables. Mutual Information classification algorithm takes target features, and features set and evaluate a MI score for each feature in the set for the target feature. MI score shows how much that feature is dependent on the target feature. MI relies on entropy estimation from K-nearest neighbour's distances. If the MI score of any feature equals zero, the feature is independent of the target feature. So, these features are none of our use, and we can neglect the same. If the MI score of any feature is greater than 0, it means that the feature is dependent on the target feature, and we can consider those features for our model. Higher the MI score, features are more dependent on the target feature, which is a good option. MI score resulted for our chosen dataset features are shown in Figure 27. Next, we have used the SelectKBest feature selection technique to select a set of top MI scored features, i.e., a set of most dependent features to provide input to the ML models.

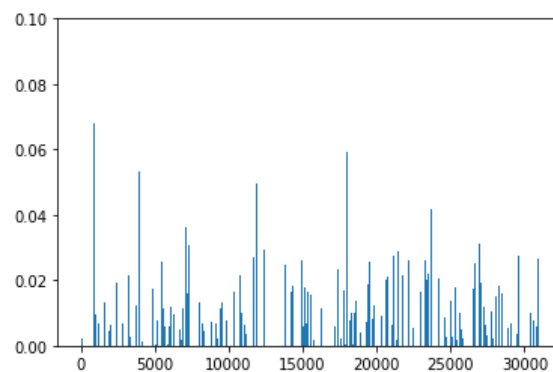


Figure 27 MI score evaluated for each feature space

Model Implementation

After selecting relevant features, we have moved forward towards model implementations. We have implemented various machine learning models on the selected feature set of chosen datasets, and those ML models are Decision Tree, Random Forest, KNN, SVM, XGBoost and Logistic Regression.

The decision tree is based on the concept of decision-making in each step. It forms a tree-like structure with possible outcomes, including resource cost, chance event consequences, and utility. This algorithm demonstrates that it contains conditional control statements. We have trained our decision tree model on our dataset. Next, we implemented random forest, a classification algorithm consisting of decision trees and uses feature randomness while building uncorrelated decision tree forest whose prediction is more accurate than any individual decision tree. The random forest has also given better results for our dataset. We have also implemented KNN (K-Nearest neighbour) classification method to classify whether an executable is malware or ransomware. This method classifies data based on the concept of how similar two data points are. Test data is used to make an 'educated guess', like class, an unclassified data point should be classified. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can quickly put the new data point in the correct category in the future. This best decision boundary is called a hyperplane. We trained our SVM model on our dataset and obtained good precision and recall; the result section discussed the same. Next, we have chosen XGBoost, an ensemble decision-tree-based mechanism based on a gradient boosting framework. When we implemented our dataset on XGBoost, we obtained the highest accuracy. At the same time, we have also implemented the same dataset

with the Logistic regression (LR) machine learning model, which is used to allocate observation to a discrete set of classes. LR uses sigmoid function and produce a probability value, and based on the threshold on probability value; we used to decide in which any sample can be classified. The hypothesis of LR tends to limit the cost function in the range of 0 to 1.

Results

As several popular ML models have been implemented in this work, we have made a comparative analysis between these model's efficiency. We have tested each implemented ML model on the test dataset and compared the results on various metrics, those are explained in detail below:

1. **Accuracy:** It means the fraction of predictions our model got right, i.e., the ratio of no of correct prediction to the total number of samples. Comparative accuracy analysis is shown in Table 14.

$$Accuracy = \frac{\text{Number of Correct Prediction}}{\text{Total number of predictions made}} .$$

Table 14 Accuracy obtained by each model

Model Name	Accuracy
Decision Tree	95.63
Random Forest	96.02
KNN	93.64
SVM	96.42
XGBoost	98.21
Linear Regression	98.21

2. **Confusion Matrix:** It provides a matrix as output and describes the complete performance of any ML model. For our case, we are solving binary class classification, and we had a total of 503 samples in our test data. There are four most relevant terms: (1) True Positive (TP) - The number of cases that are predicted as ransomware and the actual expected output is also ransomware only, (2) True Negative (TN) - The number of cases that are predicted as goodware and actual expected output is also goodware, (3) False Positive (FP) - The number of cases which are predicted as ransomware and expected output is goodware, (4) False Negative (FN) - The number of cases which are predicted as goodware and expected output is ransomware. The score of each of these terminologies are shown in Table 15.

Table 15 TP, TN, FP and FN obtained by each model

Name	TP	TN	FP	FN
DT	184	297	16	6
RF	186	297	16	4
KNN	181	299	23	9
SVM	185	300	13	5
XGB	188	306	7	2
LR	187	307	6	3

3. **Precision & Recall:** Precision shows the number of correct positive outputs in total positive results predicted by the classifier. Recall demonstrates no correct positive outputs in all samples that should have been resulted as positive. Comparative analysis is shown in Table 16.

$$Precision = \frac{TP}{TP + FP}$$

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

Table 16 Precision and Recall obtained by each model

Name	Precision	Recall
DT	0.92	0.97
RF	0.92	0.98
KNN	0.89	0.95

SVM	0.93	0.97
XGB	0.96	0.99
LR	0.97	0.98

4. Conclusion & Future Works

This work aimed to design and develop an ML-based system that identifies executables whose activity matches ransomware activities. This research works upon working on well-defined behaviour, and the intention of executables are being analyzed. This framework consists of steps: Data collection, processing of those data, features extraction and implementation of ML model. The same has been achieved by deploying Cuckoo Sandbox, Ransomware sample analysis and ML model implementation using python script.

The implemented ML model can get extended to implement an ML-based detection system sole product to detect upcoming ransomware threats for the industry. For enterprise-level implementation, we need to monitor logs generated by the organization's distributed systems. We can feed the status of API calls log, registry key operations logs, and file system operations logs to our saved model to decide whether the activity belongs to ransomware or not.

To record these logs activities, enterprises generally use EDR (Endpoint Detection and Response) and SIEM (Security Information and Event Management). EDR is an endpoint security solution that combines endpoint data and real-time continuous log monitoring with a rule-based automated response. SIEM is a security tool, which collects data(logs) from devices, domain controllers, networks, servers. It stores those data and applies analytics on stored data to detect threats, discovers trends, and enables organizations to examine any alerts. Several EDR and SIEM solutions are available in the industry, and we can integrate this ML model with any of these solutions to detect ransomware activities.

Also, we can use ELK (Elasticsearch Logstash Kibana) [30] to integrate our ML model and develop a ransomware detection system as a product. ELK is an open-source tool, and many SIEMs are built on top of ELK only. We can use ELK for elastic search to filter and collected desired logs for our detection system. The implemented model can get ingested the collected logs to decide on suspicious activity based on logs details. ELK stack is a collection of three open-source tools, i.e., Elasticsearch, Logstash, Kibana, which allow us to store several logs to visualize them easier. Elasticsearch is a search and analysis engine that allow

There are no sources in the current document.s us to store massive information in JSON format. Logstash is a data processor and collector, which allows us to collect, load, and transfer data in numerous architectures, and in the last, it can get forwarded to Elasticsearch. Kibana provides a visualization layer to a web browser to parse any query to Elasticsearch in several visualizations like histograms, linear graphs, pie charts, etc. We can use ELK directly to integrate our model with it as ELK is an open-source project.

References

- [1] Hobbs, A., The colonial pipeline hack: Exposing vulnerabilities in U.S. cybersecurity. In SAGE Business Cases. SAGE Publications, Ltd.,(2021)
- [2] Sultan, H., Khaliq, A., Alam, S.I, and Tanweer, S. A Survey on Ransomware: Evolution, Growth and Impact. International Journal of Advanced Research in Computer Science, 9(2) (2018).
- [3] Khan, C. Ncube, L. K. Ramasamy, S. Kadry, and Y. Nam: A Digital DNA Sequencing Engine for Ransomware Detection Using Machine Learning. In: IEEE Access, vol. 8, pp. 119710 119719, (2020), <https://doi.org/10.1109/ACCESS.2020.3003785>
- [4] S. K. Sahay et al.: A Survey on Ransomware Detection Techniques, Springer Nature Singapore Pte Ltd, 2020. (eds.): SKM 2019, CCIS 1186, pp. 55-68, (2020)
- [5] The current state of ransomware: Cryptowall. <https://news.sophos.com/en-us/2015/12/17/the-current-state-of-ransomware-cryptowall/> last accessed 2020/11/21.
- [6] GandCrab Ransomware. [https://www.symantec.com/security-center/writeup/2018-013106-5656- 99](https://www.symantec.com/security-center/writeup/2018-013106-5656-99). last accessed 2020/11/30.
- [7] Petters J. CryptoLocker: Everything You need to know <https://www.secureworks.com/research/cryptolockerransomware> last accessed 2021/02/21.

- [8] TeslaCrypt ransomware attack. <https://www.kaspersky.co.in/resource-center/threats/teslacrypt>. last accessed 2021/08/21.
- [9] Mohurle, S., Patil, M.: A brief study of WannaCry threat: ransomware attack 2017. *Int. J. Adv. Res. Comput. Sci.* 8(5), 1938–1940 (2017)
- [10] Sason D. REvil Ransomware Attack on Kaseya VSA: What You Need to Know. <https://www.varonis.com/blog/revil-msp-supply-chain-attack/> last accessed 2021/07/30.
- [11] Jigsaw ransomware wants to play a game, but not in a good way. <https://us.norton.com/internetsecurity-emerging-threats-jigsaw-ransomware-wants-to-play-a-game-but-not-in-a-good-way.html> last accessed 2016/11/21.
- [12] Ryuk Ransomware: https://www.trendmicro.com/en_in/what-is/ransomware/ryuk-ransomware.html last accessed 2021/09/10.
- [13] Petya (NotPetya): <https://www.proofpoint.com/us/threat-reference/petya> last accessed 2021/04/17.
- [14] WannaCry ransomware attack: https://en.wikipedia.org/wiki/WannaCry_ransomware_attack#Impact last accessed 2021/05/21.
- [15] Adams M. 2016: Year of the ransomware attacks. <http://techgenix.com/2016-ransomware/> last accessed 2021/01/07.
- [16] The Anatomy of a Ransomware Attack. (2016). The Threat Research Report: https://www.exabeam.com/wp-content/uploads/2017/07/Exabeam_Ransomware_Threat_Report_Final.pdf last accessed 2021/02/11.
- [17] Nachreiner C. How attackers hide their malware: Advanced Obfuscation. <https://www.darkreading.com/attacks-breaches/how-hackers-hide-their-malware-advanced-obfuscation/a/d-id/1329723> last accessed 2020/10/02.
- [18] Scaife, N., Carter, H., Traynor, P. & Butler, K.R. (2016). Cryptolock (and drop it): stopping ransomware attacks on user data. In 2016 IEEE 36th International Conference on Distributed Computing Systems (ICDCS) (pp. 303-312).
- [19] Kharaz, A., Arshad, S., Mulliner, C., Robertson, W., Kirda, E.: UNVEIL: a largescale, automated approach to detecting ransomware. In: 25th USENIX Security Symposium (USENIX Security 2016), Austin, TX, pp. 757–772. USENIX Association (2016)
- [20] Gomez-Hernandez, J.A., ´Alvarez-Gonzalez, L., Garcia-Teodoro, P.: R-Locker: thwarting ransomware action through a honeyfile-based approach. *Comput. Secur.* 73, 389–398 (2018)
- [21] Al-rimy, B.A.S., Maarof, M.A., Prasetyo, Y.A., Shaid, S.Z.M., Ariffin, AFM: Zero-day aware decision fusion-based model for crypto-ransomware early detection. *Int. J. Integr. Eng.* 10(6), 82–88 (2018)
- [22] D. Maiorca, F. Mercaldo, G. Giacinto, C. A. Visaggio, and F. Martinelli, “R-PackDroid: API package-based characterization and detection of mobile ransomware,” in *Proceedings of the symposium on applied computing*. ACM, 2017, pp. 1718–1723.
- [23] Vigneswaran, Rahul & R, Vinayakumar & Kp, Soman & Poornachandran, Prabakaran. (2018). Evaluating Shallow and Deep Neural Networks for Network Intrusion Detection Systems in Cyber Security. 1-6. 10.1109/ICCCNT.2018.8494096.
- [24] D. Sgandurra, L. Muñoz-González, R. Mohsen, and E. C. Lupu, “Automated dynamic analysis of ransomware: Bene_ts, limitations, and use for detection,” 2016, arXiv:1609.03020.
- [25] Al-rimy, B.A.S., Maarof, M.A., Prasetyo, Y.A., Shaid, S.Z.M., Ariffin, AFM: Zero-day aware decision fusion-based model for crypto-ransomware early detection. *Int. J. Integr. Eng.* 10(6), 82–88 (2018)
- [26] Mehnaz, S., Mudgerikar, A., Bertino, E.: RWGuard: a real-time detection system against cryptographic ransomware. In: Bailey, M., Holz, T., Stamatogiannakis, M., Ioannidis, S. (eds.) RAID 2018. LNCS, vol. 11050, pp. 114-136. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-00470-5_6
- [27] Continella, A., et al.: ShieldFS: a self-healing, ransomware-aware filesystem. In: *Proceedings of the 32nd Annual Conference on Computer Security Applications, ACSAC 2016*, pp. 336–347. ACM, New York (2016)
- [28] Harikrishnan, N.B., Soman, K.P.: Detecting ransomware using GURLS. In: 2018 Second International Conference on Advances in Electronics, Computers and Communications (ICAEECC).
- [29] Z.-G. Chen, H.-S. Kang, S.-N. Yin, and S.-R. Kim, “Automatic ransomware detection and analysis based on dynamic API calls flow graph,” in *Proceedings of the International Conference on Research in Adaptive and Convergent Systems*. ACM, 2017, pp. 196–201.
- [30] R. Stoleriu, A. Puncioiu and I. Bica, "Cyber Attacks Detection Using Open Source ELK Stack," 2021 13th International Conference on Electronics, Computers and Arti_cial Intelligence (ECAI), 2021, pp. 1-6, <https://doi.org/10.1109/ECAI52376.2021.9515120>.

Identification and removal of Code Smell

Jaswinder Singh¹, Kanwalvir Singh Dhindsa², Jaiteg Singh³

¹Chitkara University Institute of Engineering and Technology
Chitkara University Punjab, India

² Baba Banda Singh Bahadur Engineering College
Fatehgarh Sahib, Punjab, India

³Chitkara University Institute of Engineering and Technology
Chitkara University Punjab, India

Jaswinder.singh@chitkara.edu.in, kdhindsa@gmail.com
Jaiteg.singh@chitkara.edu.in

Abstract: Today's era is driven by the software. Successful software development and delivery is important and includes various phases, development strategies and implementation techniques. Software development also includes various verifications and validation methods to ensure its accuracy, completeness and confirmation to the customer requirement. If the standard practices will not be followed rigorously in the software development then development process may result in software failure. There is another possibility that there is no bug in the software but due to lack of improper design and coding standards the software is underperforming and may results in failure in future. This problem is called code smell. Thus code smell represent problem at the level of design which get coded and/or lack of coding practices used while performing coding. The objective of this paper is to identify various types of code smell that can degrade the quality of the software and also to analyze various code smell detection tools.

Keywords: Code smell, Detection tools, Software quality, code quality

1. Introduction

Software development is done using software development life cycle which includes specific phases to develop software without introducing any bug. There exist various development methodologies to govern the development process from conventional waterfall model to Agile methodologies to Devopps development. If the practices and standards will not be followed properly then the development may undergoes in failure. It may happen that the tester have not find any bug while performing testing but the development process lacks the standards that need to be followed while designing and coding. Term 'Code smell' coined in 1999[1] which indicates various issues due to improper designing and coding. The causes of problem may be due to lack of proper design or coding practices used while performing the coding. Code smells need to be handled with care for its identification and removal so that the software would have adequate quality. Code smell is not a bug but if allowed to be propagated in deployment then may results in software failure and thus results in huge loss. Table 1 below categorizes various code smell.

Couplers:

This classification shows the problem associated with more dependency (coupling) among the modules.

Object Oriented Abusers:

This bad smell shows the problem related to improper use of object oriented paradigms in the classes implementation.

Bloaters

It represents inadequate increase in the size of the code which then difficult to handle.

Dispensable:

This code smell related to that code whose presence is not required at all and the code can work without their presence.

Change Preventers

Such kind of code make difficult to update/change the code. Change at one place in the code may requires to change other parts of code also.

Sr. no.	Categories	Code smells
1	Couplers	- Feature Envy - Inappropriate Intimacy - Message Chain - Middle Man
2	Object-Oriented Abusers	- Switch Statements - Temporary Field - Refused Bequest - Alternative Classes with Different Interfaces - Parallel Inheritance Hierarchies
3	Bloaters	- Long Method - Large Class - Primitive Obsession - Long Parameter List - Data Clumps
4	Dispensables	- Lazy class - Data class - Duplicate Code - Speculative Generality
5	Change Preventers	- Divergent Change - Shotgun Surgery
6	Others	- Incomplete Library Class - Comments

Table 1: Code Smell classification by Martin Fowler and Kent Beck[1]

There may be problems due to extensive use of comments and including libraries that have not been used. Another classification Code smell can based on the code smell coverage. As listed in Table 2 below

Code smells between classes	Code smells within classes
- Alternative Classes with Different Interfaces - Primitive Obsession - Feature Envy - Lazy Class - Message Chains - Middle Man - Divergent Change - Shotgun Surgery - Parallel Inheritance Hierarchies - Incomplete Library Classes - Data Class - Data Clumps - Refused Bequest - Inappropriate Intimacy - Indecent Exposure	- Comments - Long Method - Long Parameter List - Duplicated Code - Conditional Complexity - Combinatorial Explosion - Large Class - Type Embedded in Name - Uncommunicative Name - Inconsistent Names - Dead Code - Speculative Generality - Oddball Solution - Temporary Field

Table 2: Classification based on coverage of code smell

2. Literature Review

Many researcher gave valuable contribution in the identification and classification of code smell. Paiva et al. [2] identifies various types and also analyze the accuracy of various code smell detection tools. Author identify accuracy of tools with respect to three code smell namely God Class, God Method and Feature Envy. Lanza et al. [3] proposed another code smell classes like God smell class and Brain smell class which represent high coupling and low cohesion classes. Steffen et al. [4] use tools to analyze the effect of God class and brain class on the open access software and concluded that these code smell do not harm the code until their size is extreme. Eposhi et al. [5] proposed refactoring approach to remove code smell problems that occurs due to lack of good software design. Yamashita et al. [6] relates code smell as maintainability indicators and concluded that code smell cannot be used as an maintainability indicator as system differs greatly in size. In another research [7], a method for detection and specification of Code and design smell has been proposed. This method is named as detection and correction Method (DÉCOR). Author also proposed detection technique to detect the code smell. In one of the empirical study [8] conducted on 200 open source software, It is found that code smell get introduced while creating software artifacts and if not handled properly then 80% of the code smell get survived in the system. Machine learning techniques [9] [10] are also used to detect the code smell existence in the code and detection accuracy upto 95% have been achieved. To identify defects in the code, Chau [11] proposed various metrics that perform static code analysis.

3. Identification of Code Smell

Identification of code smell: It is important to identify code smell in the source code. As stated their might be possible that the code is not generating any error but it might be not logically structured and may include flaw. Consider the code below to identify the code smell in the code and how to remove it

```

1.    int main()
2.    {
3.        int searcharr(int [],int,int);
4.        int m,i,arr[100],e,res;
5.        cout<<"Enter number of Elements:";
6.        cin>>m;
7.        cout<<"\nEnter Array Element sorted in Ascending order\n";
8.        for(i=0;i<m;++i) {
9.            i.cin>>arr[i]; }
10.       cout<<"\nEnter element to search:";
11.       cin>>e;
12.       res=searcharr(arr,n,e);
13.       if(res!=-1)
14.           i.cout<<"\nElement found at position "<<res+1;
15.       else
16.           i.cout<<"\nElement is not found....!!!";
17.       return 0;
18.    }
19.    int searcharr(int a[],int n,int e)
20.    {
21.        int f1,l1,m;
22.        f1=0;
23.        l1=n-1;
24.        d.    jump:
25.        e.    m=(f1+l1)/2;
26.        f.    while(f1<=l1) {
27.            i.if(e==a[m])
28.                a)    return(m);
29.            ii.else
30.            iii.if(e>a[m])

```

```

a)      f1=m+11;
iv.else
a)      l1=m-11;
g.      goto jump;
h.      }
i.      return -1;
6.      }

```

Now if we observe the code, it successfully determines the existent of element in the list. Code represents one of the searching techniques named 'Binary Search'. But this code is having code smell because it lacks the logical structuring. Difference between logical and illogical structuring is suggested by Farrell Joyce [12]. If we look upon statement number 5.g and 5.d, there is unconditional jump from inside of the loop from statement number 5.g to 5.d. The code is highly unstructured and require improvement. Code is restructured to remove this code smell and the updated code is given below

```

1.      int main()
2.      {
a.      int searcharr(int [],int,int);
b.      int m,i,arr[100],e,res;
c.      cout<<"Enter Number of Elements:";
d.      cin>>m;
a.      cout<<"\n Enter Array Element sorted in Ascending order\n";
e.      for(i=0;i<m;++i) {
i.cin>>arr[i]; }
f.      cout<<"\nEnter element to search:";
g.      cin>>e;
h.      res=search(arr,n,e);
i.      if(res!=-1)
i.cout<<"\nElement found at position "<<res+1;
j.      else
i.cout<<"\nElement is not found....!!!";
k.      return 0;
3.      }
4.      int search(int a[],int n,int e)
5.      {
a.      int f1,l1,m;
b.      f1=0;
c.      l1=n-1;
d.      while(f1<=l1) {
i.m=(f1+l1)/2;
ii.if(e==a[m])
a)      return(m);
iii.else
iv.if(e>a[m])
a)      f1=m+11;
v.else
a)      l1=m-11;
e.      }
f.      return -1;
6.      }

```

In the new code, code is made structure by removing the goto statement and the statement $m=(f+1)/2$; (earlier as 5.e) is shifted inside of while loop(at 5.d.i). Now the new code is more structured as compare to the previous code. So with proper identification of code smell and by removing it, we can ensure better quality software.

4. Results

Thus code smell represent problem at the level of design which get coded and/or lack of coding practices used while performing coding. As discussed in the above section, It is required to identify the code smell and to fix it. Use of goto statements results in unstructured code which should be replaced by structure coding. If the code remains unstructured, it is difficult to understand. Thus Performing changes, testing & debugging is difficult in unstructured code.

5. Conclusion

It is important to identify and locate Code smell. Code smell may introduce severe bugs in the later stages of the software development. As discussed in the section above, code smell do not stop the program to execute but degrade the program design. Good Design leads to good quality software so it is important to build code that is logically structured without introducing code smell. Code smell increases the code Complexity and readability of the code. This Research paper identifies the existence of code smell in the given code manually .Further automatic tools can be used for the identification of Code smell.

References

- [1] M. Fowler, *Refactoring Improving The Design Of Existing Code*. Addison Wesley, 1999.
- [2] T. Paiva, A. Damasceno, E. Figueiredo, and C. Sant'Anna, "On the evaluation of code smells and detection tools," *J. Softw. Eng. Res. Dev.*, vol. 5, no. 1, pp. 1–28, Dec. 2017.
- [3] M. Lanza and R. Marinescu, "Object-Oriented Metrics in Practice," in *Object-Oriented Metrics in Practice*, Springer Berlin Heidelberg, 2007.
- [4] S. M. Olbrich, D. S. Cruzes, and D. I. K. Sjøberg, "Are all Code Smells Harmful? A Study of God Classes and Brain Classes in the Evolution of three Open Source Systems," in *26th IEEE International Conference on Software Maintenance.*, 2010, pp. 1–10.
- [5] A. Eposhi, W. Oizumi, A. Garcia, L. Sousa, R. Oliveira, and A. Oliveira, "Removal of design problems through refactorings: Are we looking at the right symptoms?," in *IEEE International Conference on Program Comprehension*, 2019, vol. 2019–May, pp. 148–153.
- [6] A. Yamashita and L. Moonen, "Do code smells reflect important maintainability aspects?," in *IEEE International Conference on Software Maintenance, ICSM*, 2012, pp. 306–315.
- [7] N. Moha, Y. G. Guéhéneuc, L. Duchien, and A. F. Le Meur, "DECOR: A method for the specification and detection of code and design smells," *IEEE Trans. Softw. Eng.*, vol. 36, no. 1, pp. 20–36, 2010.
- [8] M. Tufano *et al.*, "When and Why Your Code Starts to Smell Bad (and Whether the Smells Go Away)," *IEEE Trans. Softw. Eng.*, vol. 43, no. 11, pp. 1063–1088, Nov. 2017.
- [9] N. Pritam *et al.*, "Assessment of code smell for predicting class change proneness using machine learning," *IEEE Access*, vol. 7, pp. 37414–37425, 2019.
- [10] F. A. Fontana, P. Braione, and M. Zanoni, "Automatic detection of bad smells in code: An experimental assessment," *J. Object Technol.*, vol. 11, no. 2, pp. 1–38, 2012.
- [11] C. Chou, "Metrics in Evaluating Software Defects," vol. 63, no. 3, pp. 23–29.
- [12] Farrell Joyce, *Programming Logic and Design Comprehensive*, Sixth Edition. Cengage Learning, Inc, 2011.

Data Security in Cloud Environment: A Comprehensive Survey & Open Challenges

Sartaj Singh¹, Dr. Amar Singh² Dr. Ashok Sharma³

¹Research Scholar, School of Computer Science and Engineering, LPU Phagwara

²Associate Professor, School of Computer Science and Engineering, LPU Phagwara

³Associate Professor, Dept of Computer Science and & IT, University of Jammu

sartaj.singh@lpu.co.in, drashoksharma@hotmail.com

amar.23318@lpu.co.in

Abstract: The advent of smart homes, smart cities, and smart everything has changed the world around us. The connected smart devices in Internet of Things environment are producing large volumetric data. With the enormous growth in data and need to access it from different devices, users are shifting their data to the cloud. Cloud computing has emerged as an area of incredible impact, potential, and growth. Cloud computing has enhanced the capabilities without investing in new infrastructure, training new personnel, or licensing new software. However, privacy and security of data remains the crucial concerns in cloud environment. Data stored in public cloud would face both outside attacks and inside attacks since public cloud provider themselves are untrusted. This paper reviews the possible attacks on the data and solutions to those issues which have been proposed in the literature.

Keywords: Attacks, Cloud Computing, Data Security, Privacy

1. Introduction

The rapid evolution in wireless and mobile communication in recent years has led to remarkable proliferation in the internet users. The advent of smart homes, smart cities, and smart everything has changed the world around us. The connected smart devices in Internet of Things environment are producing large volumetric data. The exponential rise in data is the prime driver for the Cloud paradigm. Cloud computing is defined by the National Institution of Standards and Technology as “a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g. networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction” [1]. According to Gartner [2] cloud computing is “a style of computing where massively scalable IT enabled capabilities are delivered ‘as a service’ to external customers using Internet technologies”. The significant features of cloud such as minimum cost involved, highly elastic, ubiquitous access to services, scalability and cheap data storage has shifted the mass towards cloud. Figure 1 shows the advantages offered by cloud. Deployment model includes three types, namely public cloud, private cloud, and hybrid cloud [3]. The service models for Cloud computing are Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). Figure 2 shows list of cloud service provider for cloud service models.

Cloud computing facilitates services to users at low cost. It offers virtual computing services such as processing, software, hardware and storage to users efficiently. The data generated from smart devices is stored on cloud. Users can retrieve their data at any time anywhere across the globe. By shifting their data to cloud offers advantages such as: a) permanent storage of their data; b) fast retrieval at any time and at any place; c) no need for actual purchasing of resources; d) minimal maintenance required for resources. This platform is not only suitable for organisations but also for individual users. They can easily access their data using smart phones, laptops, etc. sitting at any remote location.



Figure 1. Features of Cloud Computing

The immense popularity of cloud has created a profound marketplace for cloud providers. The cloud providers must consider security aspect as it is of vital importance. The biggest threat to this paradigm is data security [4]. Data security has emerged as major hindrance which is prohibiting users to shift to cloud as the data is stored and managed by third party. Cloud providers must ensure security of servers. There must not be leak of any secret data of the users which is stored on the servers. With the increasing computational power, attackers too have become more computationally efficient and try different attacks to breakthrough the servers to gain access of the stored data. This paper outlines key threats which are possible in cloud environment and potential security solutions to these threats.

SaaS	Google Apps, Cloud9 Analytics, IBM, Antenna Software, Solutions, Exoprise Systems, Host Analytics, Knowledge Tree, Reval, Taleo, NetSuite, Microsoft 365
IaaS	Amazon Elastic Compute Cloud, Rackspace, Bluelock, CSC, GoGrid, IBM, OpenStack, Rackspace, Savvis, VMware, Terremark, Citrix, Joyent, and BluePoint
PaaS	Amazon AWS, Google Apps, Microsoft Azure, SAP, Salesforce, Intuit, Netsuite, IBM, WorkXpress, and Joyent

Figure 2. Cloud Service Providers

1.1 Organisation of the paper

The paper is organised as follows. Section 2 discusses existing schemes in the field of cloud computing. Section 3 explains possible threats in cloud environment. Section 4 presents a survey of recent security solutions to overcome data confidentiality related security challenges and critical evaluation of literature has been presented in Section 5. Section 6 concludes the paper.

2. Related Work

Several schemes to ensure security of data in cloud environment have been proposed in the literature. In 2012, Chen & Zhao [15] analysed privacy and security concerns of data in cloud computing. The current solutions adopted to tackle these concerns have also been discussed. In Liu et al [16], a dynamic multi-owner data sharing scheme, *Mona*. Users can share their data with other participants using untrusted cloud. It employs bilinear maps and ensures efficient joining of new user and user revocation. Furthermore, it is computationally inefficient. In 2014, Wei et al. [17] put forth an auditing protocol named as *SecCloud*. It permits storage of user data on the cloud securely. It employs probabilistic sampling and verifier signature to perform storage and computation auditing. However, it is not computationally efficient as it has more cost than the existing schemes. Another scheme in same context was put forth by Dong et al. [18] in 2015, *SECO*. The scheme uses hierarchical identity based encryption for secure data collaboration services in cloud. But this scheme too suffers from several issues. It needs synchronisation and fails to completely secure the data. The scheme has not been implemented on actual cloud to assure its performance.

In 2016, Yao et al. [19] came with a framework which integrated big data processing and semantic analysis to process and analyse data. It specified stages in semantic security, such as collection, storage, processing and analysis using semantic framework. The paper lacks actual results. The framework has neither been implemented using any tool nor has it discussed the cost involved in performing this semantic analysis. In the same year, Aldossary & Allen [20] put forth a survey paper which explained prominent issues faced in cloud platform such as data loss, virtualisation security, integrity of data, confidentiality and possible solutions have been discussed. In 2019, Lin & Liu [21] put forth the idea of using deep learning in processing and analysis of data. The authors has used this in context of industrial internet and discussed several steps for acquiring

and representation of data. In the same year, Sharma & Singh [9] presented the idea of employing “programming cryptography” instead of encrypting data using run-time schemes. It greatly minimises the cost as it has reduced the length of output thus, providing better compression ratio. It also enhances security.

Sun [22] in 2019 discussed privacy issues of cloud computing, searchable encryption techniques and mostly employed models for attributed-based encryption, proxy re-encryption and hierarchical encryption and put forth a framework for achieving privacy protection. Yadav et al. [23] employed blockchain to achieve two factor authentication system for real time monitoring. The authors claimed it trustworthy and efficient for storing data on cloud. It lacks effectiveness. In 2020, Singh & Sharma [8] discussed different asymmetric cryptosystem along with their applications, current usage. In 2021, Tahir et al. [24] presented *CryptoGA*, which employing Genetic Algorithm (GA) to produce keys for performing encryption and decryption. *CryptoGA* tackles privacy issues and data integrity. However, the scheme has significant memory overhead.

3. Possible Threats in Cloud Environment

This section discusses the possible threats in cloud environment. The threats in cloud environment can broadly be categorised as: threats pertaining to Data, Network and Cloud Environment.

a) Threats pertaining to Network

DOS attack: Denial of Service is one of most popular security attack in which the attacker repudiates the access of resources to the authorised users. The attacker makes the resource unavailable by flooding the network. He disrupts the services by flooding network with unwanted packets which consumes most of the cloud resources such as computation power, bandwidth of network and memory, thus users are unable to access the services [5].

Insider attack: The privileged user may try to access the data of some other users deliberately. The confidential data of the users is one of the vulnerable resource and it comes to high risk when some privileged insider gets access of it. The frequency of such attacks is high and is one of the most common attack.

Impersonation attack: The attacker impersonates as an authentic entity to gain access of services. He may try to transmit recorded messages such as identity of user so that he can successfully gain access of the secret data. If the cloud has strong cloud authentication mechanism, then this impersonated messages will not be authorized to access services.

Password guessing attack: The attacker initially listens to the ongoing communication of authentic entity and server which are in the process of mutually authenticating each other. He then uses information from the eavesdropped messages along with the guesses of probable password. This attack can be attacked in two ways, online and offline. The attacker may try to make guess in real time or later in offline mode. This is just like an exhaustive search where he executes all possibilities of the password [6].

Replay attack: The attacker captures the messages which are communicated between legitimate entities in a session and replays them later. The attacker secretly listens to the messages. He then resends the messages to gain access of resources. To avert this attack, the authentication mechanism must employ nonce and timestamps. If an attacker tries to start a fresh session using recorded messages, it will be terminated as permitted timestamp delay will be high.

Service/Account hijack: Cloud services are accessible through internet. This also brings risk to the account of the users. Users may loss control of their account thus, loosing their confidential data. Public networks brings several threats to the user account on cloud.

b) Threats pertaining to Data

Data breach: The attackers may resort to numerous ways to breach the security of the system and gets access to the data. The attackers find every loophole to get access of the data such as poor authentication mechanism, flaws in application design and infrastructure, and lack of audit control [7, 8]. This greatly affects companies such as Google, Yahoo and Microsoft.

Data Loss: One of the crucial concerns of cloud is loss of data. The secret data of users is stored on cloud. Several threats such as data deletion unintentionally, loss of keys used for data encryption [9], data corruption, natural disasters, network attacks, etc lead towards loss of data. Users migrate to cloud so that they can store, manage, and retrieve their data in future as and when required. However, loss of this data has a devastating effect. [10]

c) Threats pertaining to Cloud Environment

Transition in business model: Cloud brings transfer of control of data to third party. Additionally, there is vast change of business model as compared to traditional one. This model also brings significant risks which are new and attackers are trying every possible way to intrude into the system.

Abusive use: Cloud users may misuse cloud services. As the confidential data of the user is stored on the cloud, intruders resort to different ways to access data illegally or unethically. They may violate their agreement or perform attacks to get access.

API and browser vulnerabilities: Threat to the API of cloud provider or to interface is one of the crucial risks in cloud environment. The intruders may perform browser based attacks to breach the security.

Shared resources vulnerabilities: Cloud platform brings sharing of resources which is its one of the advantage as it reduces costs. This increase in leverage of resources create a single point of attack. For instance, sharing of technology is a hypervisor or cloud orchestration.

4. Countermeasures and Controls

In this section, the security countermeasures to avert the threats have been discussed.

a) **Access Management:** Strong access control mechanism must be devised so that the sensitive data of the user must only be accessed by legitimate users. There two aspects in access management; firstly, the physical systems where actual data is stored needs to be constantly monitored. Secondly, security mechanism must be implemented to avert traffic access to the data. To restrict access from untrusted parties, Intrusion Detection System (IDS) and firewalls must be employed. They can also observe malicious activities. Furthermore, authentication standards such as eXtensible Access Control Markup Language (XACML) and Security Assertion Markup Language (SAML) must be enforced to restrict access. XACML emphasize reaching on authorization selection and SAML emphasize on shifting authentication control between entities [11].

b) **Data Protection:** There can be breach of data security accidentally or deliberately. Some privileged insider or attacker sitting at any remote location can try to gain access of data. There is need to devise security tools for averting network attacks. Specific tools for averting data loss, preserving structure, detection of behavior and authentication techniques can be employed [12] for monitoring and real time detection of suspicious activities.

c) **End-to-end encryption:** As the data is stored at diverse geographical locations, the cloud model must ensure end-to-end encryption of the data. The data which is to be stored on cloud, must be encrypted to ensure its secrecy.

d) **Improvised security mechanisms:** Cloud providers must adopt strong authentication mechanism. The remote user must be authenticated before granting access. If the process of verifying user's identity is weak, the intruders will get advantage and perform malicious activities for instance, spam and performing attacks on other computing systems [13]. Strict enforcement of security techniques minimizes the misuse of cloud services [14].

The cloud computing paradigm must ensure following services to the users:

a) **Availability:** The cloud services must be available to legitimate cloud users irrespective of their location and time. Whenever the user wants to access the cloud services, cloud provider must ensure easy access to resources. Attackers or malicious users resort to numerous possible ways to make the resource unavailable to the users by performing attacks such as DOS (denial-of-service attack), DDOS (distributed denial-of-service attack), or flooding attacks. To make the resource available when requested, strategies such as distributed approach of operating system can be employed.

b) **Integrity:** As the communication between the entities occurs in a public environment, integrity of the message must be ensured. Integrity means that message was not manipulated or changed during the

transmission. It ensures that no illegitimate modification is made by any unauthorised user. Maintaining integrity of data is of vital importance. The receiver must receive the exact message sent by the sender.

c) Confidentiality: The secret data of the user stored on the cloud must be made available to legitimate users only. The data can be viewed, edited, and deleted by authorised entities only. For instance, malware such as Trojans, virus, spyware, etc. are used to perform attack on data. This is one of the basic prerequisites of cloud and cloud provider must promise the confidentiality of the secret data which is store or managed on the cloud.

d) Authentication: It is applied to validate the identity of the entities involved during communication. Enforcing integrity and authentication is vital as identity of sender and receiver must be verified as exchange of information/messages takes place in insecure environment.

e) Non-repudiation: It warrants that transmission cannot be denied between the sender and the receiver. The approved communication must be permitted to exchange messages.

5. Critical Evaluation

In this section, schemes proposed in literature have been critically evaluated. Several schemes have been proposed in the literature to ensure security of data. Table 1 shows the major contributions made by the authors over the years. These existing schemes have certain gaps in their study such as fails to ensure authenticity, confidentiality and many more.

Year	Author	Technique used	Result	Gaps
2012	Chen & Zhao [15]	Analysed the data security and protection of data	Survey of issues of data security	Employs special data structure on user side
2013	Liu et al [16]	Proposed “Multi-owner data sharing scheme” (Mona) in cloud using bilinear maps and group signature	Anonymous ly sharing of data	Computationally inefficient
2014	Wei et al. [17]	Proposed <i>SecCloud</i> employing probabilistic sampling and signature of verifier	Secure computation auditing protocol	Enhanced cost than existing
2015	Dong et al. [18]	Proposed SECO using hierarchical identity based encryption	Ensure cloud data access and updating	- Need of synchronization - Privacy issues
2016	Yao et al. [19]	Proposed framework for processing for big data using semantic association and inference methods.	Ensures security of data	No practical implementation of semantic analysis and HCI to meet data security analysis
2016	Aldossary& Allen [20]	Discussed schemes to attain confidentiality, integrity, and availability of data in the cloud	Issues pertaining to cloud platform are discussed	- Current solutions have flaws in achieving confidentiality of data integrity - Lacks availability
2019	Lin & Liu [21]	Deep learning based development of algorithm for big data analysis of industrial data	Analyse industrial big data	Lacks effectiveness
2019	Sun [22]	Proposed Attribute-based encryption framework for privacy protection	To achieve privacy	Flaws in privacy protection
2020	Yadav et al. [23]	Improving the security of the hospital website using Blockchain	To enhance security	Lacks effectiveness
2021	Tahir et al. [24]	Genetic algorithm based CryptoGA is proposed	To cope with data integrity and privacy issues.	Great memory overhead

6. Conclusion

Cloud paradigm is continually growing and being adopted in businesses all over the globe. Nevertheless, it also brings several security challenges with it. Security must be considered as vital for cloud. This paper discussed prominent security issues which have come up with the advent of technology and increase in computational power of attackers. The cloud service providers must enforce strong security constraints to ensure secrecy of data. Only authentic users should get access to data and the data must be made available to user as and when. Furthermore, this paper also suggests which can be adopted to avert threats.

References

- [1] Mell P, Grance T (2018) SP 800-145, The NIST Definition of cloud computing | CSRC (online) Csrc.nist.gov. <https://csrc.nist.gov/publications/detail/sp/800-145/fnal>.
- [2] Heiser J. What you need to know about cloud computing security and compliance, Gartner, Research, ID Number: G00168345, 2009.
- [3] Bhamare D, Samaka M, Erbad A, Jain R, Gupta L, Chan HA (2017) Optimal virtual network function placement in multi-cloud service function chaining architecture. *Comput Commun* 102:1–16
- [4] Yu S, Wang C, Ren K, Lou W (Mar 2010) Achieving secure, scalable, and fine-grained data access control in cloud computing. In: *Proceedings of the IEEE INFOCOM*
- [5] Akshaya MS, Padmavathi G (2019) Taxonomy of security attacks and risk assessment of cloud computing. In: Peter J, Alavi A, Javadi B (eds) *Advances in big data and cloud computing. Advances in intelligent systems and computing*, vol 750. Springer, Singapore
- [6] Subramanian N, Jeyaraj AJC, Engineering E (2018) Recent security challenges in cloud computing. *Comput Electr Eng* 71:28–42
- [7] Bernard Golden. Defining private clouds, 2009 /http://www.cio.com/article/492695/Defining_Private_Clouds_Part_OneS
- [8] Singh, S., & Sharma, A. (2020). Cryptosystems in Asymmetric Cryptography for Securing Data at various Level. *Delving*.
- [9] Sharma, E. R., & Singh, S. (2019, April). Reforming Cyber Security with Coding. In *2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)* (pp. 1-6). IEEE.
- [10] Sharma, A., Jha, P., & Singh, S. (2021). Data Control in Public Cloud Computing: Issues and Challenges. *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)*, 14(2), 564-579.
- [11] Jansen, W. A., & Grance, T. (2011). Guidelines on security and privacy in public cloud computing.
- [12] Stolfo, S. J., Salem, M. B., & Keromytis, A. D. (2012, May). Fog computing: Mitigating insider data theft attacks in the cloud. In *2012 IEEE symposium on security and privacy workshops* (pp. 125-128). IEEE.
- [13] Tripathi, A., & Mishra, A. (2011, September). Cloud Computing Security Considerations Interface. In *2011 IEEE International Conference on Signal Processing, Communications and Computing, Xi'an, China*.
- [14] Li, H. C., Liang, P. H., Yang, J. M., & Chen, S. J. (2010, November). Analysis on cloud-based security vulnerability assessment. In *2010 IEEE 7th International Conference on E-Business Engineering* (pp. 490-494). IEEE.
- [15] Chen, D., & Zhao, H. (2012, March). Data security and privacy protection issues in cloud computing. In *2012 International Conference on Computer Science and Electronics Engineering* (Vol. 1, pp. 647-651). IEEE.
- [16] Liu, X., Zhang, Y., Wang, B., & Yan, J. (2012). Mona: Secure multi-owner data sharing for dynamic groups in the cloud. *IEEE transactions on parallel and distributed systems*, 24(6), 1182-1191.
- [17] Wei, L., Zhu, H., Cao, Z., Dong, X., Jia, W., Chen, Y., & Vasilakos, A. V. (2014). Security and privacy for storage and computation in cloud computing. *Information sciences*, 258, 371-386.
- [18] Dong, X., Yu, J., Zhu, Y., Chen, Y., Luo, Y., & Li, M. (2015). SECO: Secure and scalable data collaboration services in cloud computing. *computers & security*, 50, 91-105.
- [19] Yao, Y., Zhang, L., Yi, J., Peng, Y., Hu, W., & Shi, L. (2016, September). A framework for big data security analysis and the semantic technology. In *2016 6th International Conference on IT Convergence and Security (ICITCS)* (pp. 1-4). IEEE.
- [20] Aldossary, S., & Allen, W. (2016). Data security, privacy, availability and integrity in cloud computing: issues and current solutions. *International Journal of Advanced Computer Science and Applications*, 7(4), 485-498.
- [21] Lin, J., & Liu, L. (2019, July). Research on security detection and data analysis for industrial internet. In *2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRS-C)* (pp. 466-470). IEEE.
- [22] Sun, P. J. (2019). Privacy protection and data security in cloud computing: a survey, challenges, and solutions. *IEEE Access*, 7, 147420-147452.
- [23] Yadav, D., Shinde, A., Nair, A., Patil, Y., & Kanchan, S. (2020, May). Enhancing data security in cloud using blockchain. In *2020 4th International Conference on Intelligent Computing and Control Systems (ICICCS)* (pp. 753-757). IEEE.
- [24] Tahir, M., Sardaraz, M., Mehmood, Z., & Muhammad, S. (2021). CryptoGA: a cryptosystem based on genetic algorithm for cloud data security. *Cluster Computing*, 24(2), 739-752.

Virtualization Attacks and Prevention Solutions in Cloud Environment

¹Vaneeta and ²Sandeep Kumar

¹ Assistant Professor, Department of Computer Science, Patel Memorial National College, Rajpura, Punjab.

² Assistant Professor, Department of Computer Science, Patel Institute of Management and Technology,
Rajpura,

: vaneeta.dhawan@gmail.com, sandip.kashyap17@gmail.com

Abstract: Cloud computing provides an efficient way to the users for accessing computing services and resource usage. To overcome the problems of data accessibility and security, it provides a mechanism of virtualization in which by creating a virtual image of resources and storage that can be accessed by more than one machine simultaneously. In this way, it helps in reducing the cost of hardware and it can pay as per the need of the user. The hypervisor controls the effectiveness and security of the virtual systems. Along with these benefits, cloud computing is still not secure due to local or remote attacks. From the security aspects, implementation of the secured virtualization infrastructure is necessary for overall security. This paper includes the types of virtualizations along with the risk associated with virtualization and solutions of how to prevent these attacks.

Keywords Cloud Computing, Virtualization, Prevention, Attacks

1. Introduction

Cloud computing is a platform that delivers IT services using diverse web-based applications and tools. It additionally provides a cloud-based storage system that makes it secure to store data in remote databases. Therefore, it can be easily retrieved and used. Though, there exist some attacks that affect the cloud resources and data available at remote locations [1]. Cloud provides different approaches to handle difficulties faced by an organization in monitoring and managing their data and that is achievable through Virtualization. Service models provide services at different levels such as IaaS (Infrastructure as a Service) is the bottom level in the model, PaaS (Platform as a Service) is the middle level and SaaS (Software as a Service) is the topmost level in service models. The lower level inherits the features of the higher level in the stack[2][3].

Virtualization is the fundamental concept in cloud computing. It provides a basis to deliver infrastructure services over the cloud environment. It administers the allocation of multiple Virtual machines on a single host machine i.e. a single CPU works as multiple CPUs virtually and same as in case of RAM and hard disks[4]. Virtual machine monitor (VMM) is the software used for virtualization that runs in between the physical server and the OS. It accesses the physical devices on a server such as CPU, disk, memory, etc. It also protects virtual machines[5]. Virtualization can be categorized based on software, memory, full and partial virtualization. There exist some tools such as open-source and commercial solutions that are used for cloud virtualization. Commercial tools such as Hyper-V, ESX, VMware, etc., and open-source tools such as CloudStack, OpenStack, Nimbus, Eucalyptus, etc [2].

Additionally, it also provides the environment over the internet and intranet that facilitate the user with these services, such as hardware resources, power saving, recovery and storage of data as well as flexibility through VM's and hypervisor (Virtual Machine Monitor). As a result, it overcomes the burden of the user as well as service providers, however, while at the time of designing and implementation of virtualization technology has faced a vital number of security issues and threats such as VM escape, Hyper jacking, and VM Image attack[2]. These threats are the major barriers in adoption of cloud environment. The paper is organized as follows: Section 2 includes the classification of virtualization and Section 3 covers the attacks or security issues in cloud and the virtualization security measures or prevention techniques and finally followed by the conclusion.

2. Types of Virtualizations

Virtualization creates a layer which is known as a Virtual machine monitor or the hypervisor between the guest and host. It Provides security but still, it is vulnerable to malicious attacks. Virtualization can be

categorized into different types as Client virtualization, Server virtualization, Storage virtualization, etc as represented in Figure 1.

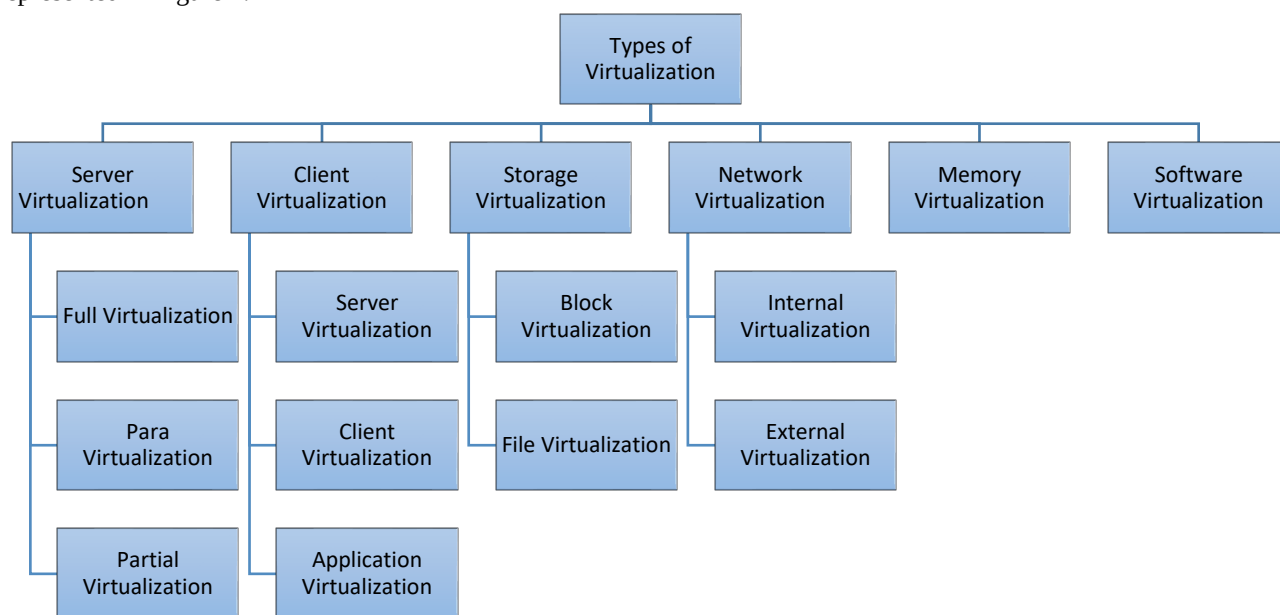


Figure 1: Types of Virtualization

2.1 Server Virtualization: Server Virtualization is achieved with the help of a single server that serves the functioning of multiple servers by dividing the resources. This virtualization is the most desirable approach to spend minimum resources and infrastructure [6]. The hypervisor layer hosts multiple applications and virtual machines on a local system or remote machines and controls the memory, processor, other components by running the different OS on the same machine. The basic purpose of virtualization is to merge small servers into one large server. It provides the service of efficient resource utilization, reduces cost, and availability of resources according to the need of the user [4]. It is divided into following types:

- i.**Full Virtualization:** The components of the physical server are directly accessed in full virtualization. Each virtual machine running on the same server is not aware of the other virtual machine. To run an application on the virtual server, the hypervisor asks the physical server for resource availability. In this way, the physical server provides the processing power to the hypervisor which in turn slows down the overall performance [6].
- ii.**Para Virtualization:** In this virtualization, all the virtual machines are aware of the existence of each other. The OS of each virtual machine running on the physical server is aware of the resource's needs. Minimum processing power is required by the hypervisor that in turn improves the overall server performance[6].
- iii.**OS level Virtualization:** The hypervisor is not required in OS-level virtualization. This technology divides the operating system in such a way to create a large number of virtual machines that are isolated from each other. It improves the security and availability of resources in a complex environment. This virtualization has the major drawbacks that the OS should remain the same for all virtual servers[6].

2.2. Client Virtualization: Client virtualization helps the system admin to monitor and modify the client system such as desktop, laptop, etc. virtually. It improves the management and security of the client machines[4]. It can be of three types as Server Virtualization, Client Virtualization, and Application virtualization.

- i.**Server Virtualization:** Server or remote virtualization is accessed by the client on a server.
- ii.**Client Virtualization:** It runs on the local machine with secured and virtual environment.

iii. **Application Virtualization:** An independent and virtual environment is required for the execution of an application[4].

2.3. Storage Virtualization: This virtualization is used to extract logical forms of storage from physical storage that can be scattered over the network [7]. This is easily implemented and is not expensive. This is used when there is a need to recover from a disaster where the data will be moved to some other virtual storage. This virtualization is used in the Storage area network [8]. Data storage can be of three types: Direct attached storage, Storage area network, Network-attached storage. It provides the following benefits as Utilization of storage effectively, Updates are provided easily, Availability of data, etc [4]. Virtual resources are managed easily as compared to physical storage. This storage virtualization can be of two types:

i. **Block Virtualization:** It creates storage devices that are virtually accessible from actual physical storage media which is fast, available and large according to the need of consumer. This virtualization uses the virtual disk for accessing operating systems and applications [4].

ii. **File Virtualization:** This virtualization is applied at the virtual storage for the operating system and applications using the software installed on the server[4]. This virtualization also helps in transferring unused data to the other inexpensive storage media that is available for both users and the applications[6]

2.4. Network Virtualization: This virtualization combines the hardware and software resources in such a way to create a virtual network[7]. It monitors and manages the network from a single administrator system. This virtualization is beneficial for networks with large, fast, and unpredictable traffic. It resulted in better utilization of resources and efficiency of the network with reduced cost [4]. Network Virtualization are categorized into two parts.

i. **Internal Virtualization:** It provides complete network functionality on a single system[4].

ii. **External Virtualization:** It merges more than one network into a virtual network to improve network efficiency[4]

2.5. Memory Virtualization: To Separate the memory from a server for distributed and shared working, Memory virtualization is required. It provides greater memory without the need to add more main memory. It can be implemented at the application level or the operating system level. At the application level, running applications can be directly connected to the memory using the file system or API. And at the operating system level, OS connects with memory first, and then the memory will be accessed by the applications[4].

2.6. Software Virtualization: This virtualization helps the organization to use the software on a large number of systems simultaneously. It can be accessed using the virtual layer where the software can be installed [9]. It also enables a system to run as a guest operating system such as on a Windows operating system Linux runs as a guest or vice versa[4]. It can be categorized into two types:

i. Application Virtualization

ii. Operating system Virtualization

3. Virtualization Security Issues and Prevention Techniques in Cloud Computing

Virtualization provides advanced features in the cloud for Memory, Applications, Operating system, etc. that helps in executing multiple OS on a single server. With the positive effects of virtualization, there exist some negative effects too. Various security issues are identified in cloud virtualization [7]. These are as follows:

3.1. VM Escape: The virtual environment provides influential isolation for the Virtual machines and the host. But an attacker can insert malicious code inside the virtual machine and whenever the code is executed, the virtual machine breaks the isolation and communicate directly with the host operating system and avoid the virtual machine or hypervisor layer as shown in Figure 2 [2]. In this way, the host machine can be fully accessed by an attacker that resulted in an entire failure of the system. This kind of attack is known as VM Escape[7].

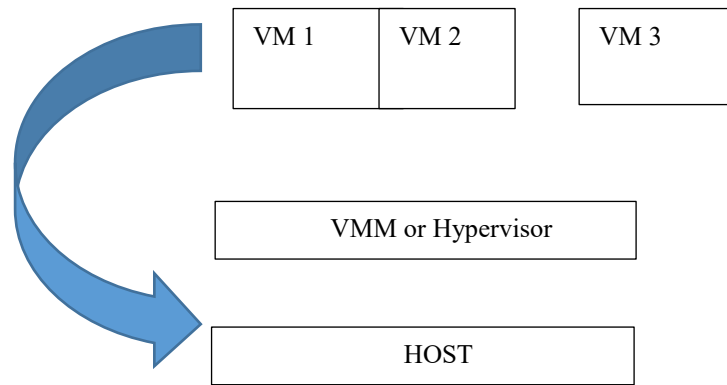


Figure 2: VM Escape

Prevention of VM Escape: This problem can be resolved by applying minimum adaptability in configuring host and guest. And other solution is to use HyperSafe, that is a lightweight approach that provides self-protection and flow integrity[10].

3.2. VM Migration: VM Migration constitutes transferring the virtual machine from one physical device to another physical device in an active mode i.e. without the need to shut down the virtual machine as shown in Figure 3 [10]. During migration, there is a need to protect the data from malicious attacks on the network that resulted in the loss of integrity and privacy. Migration can be performed due to the following reasons such as load balancing, fault tolerance [5].

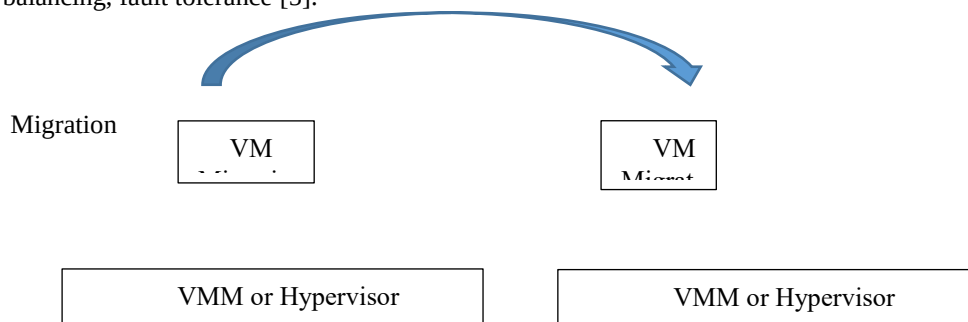


Figure 3: VM Migration

Prevention of VM Migration: PALM (Protection aegis for live migration) is used for secure migration that maintains privacy and integrity pre- and post-migration. Data must be encrypted to maintain privacy[10].

3.3. VM Image Sharing: VM image is a format that can be used to create virtual machines. It is a configured software that includes virtual machine configuration files. By sharing the vulnerable images of virtual machines compromises system security. It imposes risk on different systems such as on publisher and retriever. Publisher may provide sensitive information with the image that can be accessed by only limited users. Retriever may access the image that contains malicious code. Thus, in this way, VM image sharing becomes a serious attack in the cloud[10].

Prevention of VM Image Sharing: This problem can be resolved by applying limitations on the framework that can identify who accesses the image, cryptographic protection bits to check unauthorized modification in the VM images, follow the history from the origin of an image, etc[2][10].

3.4. VM Sprawl: In Virtualization, VM sprawl occurs when the resources are consumed unnecessarily i.e., the VM exists without proper management[2]. These VMs can retain the resources such as memory, CPU, network channel for a long period and are not assigned to the other VMs. This occurs due to VM was created easily without approval or after using the VM it was left idle for a long period [10]. Virtualization makes this possible in minutes to create VM that created this attack.

Prevention of VM Sprawl: This problem can be eliminated by employing policy with the help of an administrator that controls who may get which resource. A policy should be defined for the VM which are expired or not required because the VMs are not consuming CPU or memory but are archived in memory that may take memory. It should be cleaned after a particular period[10].

3.5. Hyperjacking: An attacker attacks the VM by controlling the hypervisor running on the host system and gains complete access to virtual machines and servers. This attack targets the operating system that is under the virtual machines to run the malicious programs[2]. This is a critical security issue in virtualized environment.

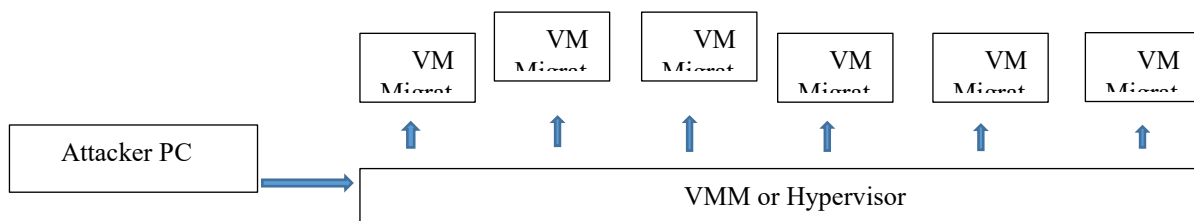


Figure 4: Hyperjacking

Prevention of Hyperjacking: This problem can be resolved by protecting the hypervisor control-flow integrity using HyperSafe and the other technique is to implement an intrusion detection system on the hypervisor[2].

3.6. Inside-VM attacks: Threats may affect the Virtual machines at run time using the OS rootkits. A rootkit is a malicious program that may get administrative control to the virtual machine. This attack takes the control of the virtual machine and accesses the system given in Figure 5[2].

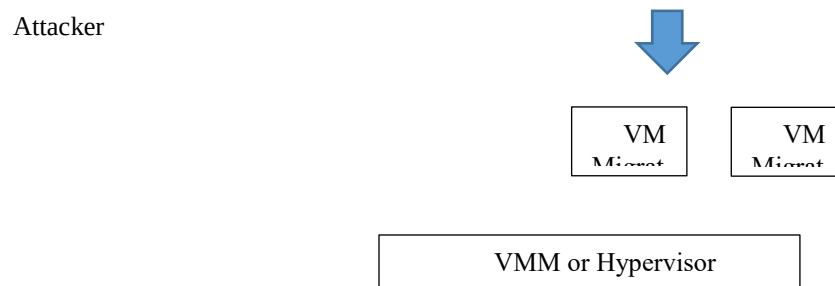


Figure 5: Inside VM attacks

Prevention of Inside-VM attacks: This problem can be resolved by introspecting virtual machines. CERT and RAND are used to classify insider attacks[11].

3.7. Host Machines attacks: A host virtual machine is administered from the server and is open for remote users over the internet. The host machine is the root of the VM that monitors and controls the applications running on the VMs. Host and virtual machines can interact with each other in the following ways[10].

- Host can start, pause, restart, shutdown the VM.
- It can modify the content on VM.
- Views the running applications on VMs.

All the traffic of virtual machines passes through the host that can monitor each virtual machine.

Prevention of Host machines attacks: The host controls all the virtual machines, thus it must be rigorously monitored and protected over the VM. Another solution of this attack prevention is isolation of host so that no attack affect the virtual machines and host[5][10].

4. Conclusion

Virtualization is an efficient way to reduces the costs and improving the availability and better utilization of resources. Cloud computing transformed the computing technology by using virtualization. It plays an important role in cloud as it provides a virtual memory, virtual server and services to the clients. This technology reduces cost by limiting the hardware resources. Due to the virtual cloud environment, security standards need to be implemented. This paper analysed the virtualization technology, types of virtualization, and the vulnerabilities that lead to attacks with their possible preventive solutions. Future work will include the implementation of these techniques for detection and prevention of attacks in virtual environment.

References

- [1] S. M. Alturfi, D. K. Muhsen, M. A. Mohammed, I. T. Aziz, and M. Aljshamee, "A Combination Techniques of Intrusion Prevention and Detection for Cloud Computing," *J. Phys. Conf. Ser.*, vol. 1804, no. 1, 2021, doi: 10.1088/1742-6596/1804/1/012121.
- [2] D. Tank, A. Aggarwal, and N. Chaubey, "Virtualization vulnerabilities, security issues, and solutions: a critical study and comparison," *Int. J. Inf. Technol.*, 2019, doi: 10.1007/s41870-019-00294-x.
- [3] F. Baroncelli, B. Martini, and P. Castoldi, "Network virtualization for cloud computing," *Ann. des Telecommun. Telecommun.*, vol. 65, no. 11–12, pp. 713–721, 2010, doi: 10.1007/s12243-010-0194-y.
- [4] A. Rashid and A. Chaturvedi, "Virtualization and its Role in Cloud Computing Environment," *Int. J. Comput. Sci. Eng.*, vol. 7, no. 4, pp. 1131–1136, 2019, doi: 10.26438/ijcse/v7i4.11311136.
- [5] L. Chen, M. Xian, J. Liu, and H. Wang, "Research on Virtualization Security in Cloud Computing," *IOP Conf. Ser. Mater. Sci. Eng.*, vol. 806, no. 1, 2020, doi: 10.1088/1757-899X/806/1/012027.
- [6] K. Basuki, "濟無No Title No Title," *ISSN 2502-3632 ISSN 2356-0304 J. Online Int. Nas. Vol. 7 No.1, Januari – Juni 2019 Univ. 17 Agustus 1945 Jakarta*, vol. 53, no. 9, pp. 1689–1699, 2019, [Online]. Available: www.journal.uta45jakarta.ac.id.
- [7] P. Kedia, R. Nagpal, and T. Pal Singh, "A Survey on Virtualization Service Providers, Security Issues, Tools and Future Trends," *Int. J. Comput. Appl.*, vol. 69, no. 24, pp. 36–42, 2013, doi: 10.5120/12123-8491.
- [8] K. R. Dinakar, "A Survey on Virtualization and Attacks on Virtual Machine Monitor (VMM)," pp. 6558–6563, 2019.
- [9] L. Malhotra, D. Agarwal, and A. Jaiswal, "Virtualization in Cloud Computing," *J. Inf. Technol. Softw. Eng.*, vol. 04, no. 02, pp. 2–4, 2014, doi: 10.4172/2165-7866.1000136.
- [10] J. M. Kizza, "Virtualization Security," pp. 473–490, 2015, doi: 10.1007/978-1-4471-6654-2_22.
- [11] A. J. Duncan, S. Creese, and M. Goldsmith, "Insider attacks in cloud computing," *Proc. 11th IEEE Int. Conf. Trust. Secur. Priv. Comput. Commun. Trust. - 11th IEEE Int. Conf. Ubiquitous Comput. Commun. IUCC-2012*, pp. 857–862, 2012, doi: 10.1109/TrustCom.2012.188.

Impact of Dataset Properties on Spam Email Classification Using Machine Learning

Pramod Prakash Ghogare¹ and Manoj P. Patil²

School of Computer Sciences, KBC North Maharashtra University, Jalgaon, India

¹pramod.ghogare@yahoo.com

²mpp145@gmail.com

Abstract: The use of e-mail increased in communication due to its efficiency and lower cost compared to other channels. Internet communication is regularly facing problems due to spam e-mails. The e-mail spam filtering methods were developed and used by various email service providers, but still, the spam e-mails are rising. The dataset has different attributes such as size, the period of e-mail collection, and the complexity of e-mails, the experiment is intended to extract the appropriate features from the e-mail header and classify spam e-mails using Support Vector Machine, Naïve Bayes, and Random Forest. The performance of the techniques measured in terms of precision, recall, accuracy, f-score, and receiver operator characteristics and the effect of the dataset properties were analyzed. The result is used to discern the performer algorithm in the classification of the spam e-mails of a different dataset.

Keywords: Spam E-mail, Classification, Machine Learning, Dataset, Random Forest, Naïve Bayes, Support Vector Machine

1. Introduction

Spam is an “Unsolicited, unwanted e-mail that was sent indiscriminately, directly or indirectly, by a sender having no current relationship with the user” [1]. “Real spam is that e-mail for advertising of product sent to list groups” [2]. These e-mails mostly have an advertisement, product details, and offers on the product, discounts, demanding donations, and requests for financial details.

The spam e-mail degrades work efficiency by troubling users constantly and irritating them [3]. Due to the frustration of spam e-mail, a few times genuine e-mails were removed. The e-mail server consumes a lot of time for processing and removing spam e-mails [4]. The internet service provider and network administrator are affected by the spam e-mail. The amount of money, time, and resources engaged by spam e-mails are the main reason to stop spamming. If spam e-mail is filtered before entering into a network, it can save many resources to use, engaged or spoiled.

2. Problem Definition

Spam e-mail must not spread to the inbox of users as it decreases the productivity of operations. Similarly, it is also necessary that the legitimate e-mail should not receive in the spam box, this can create a serious problem for the user [5]. No method claim to provide an ideal solution with 100% accuracy in spam e-mail classification [6]. Most of the researchers recognize the undesirable data and potentially dangerous activity detection by spam e-mails [7]. The orthodox methods of spam e-mail classification are not feasible to handle in a real-time environment. Several spam e-mail filters fail to decrease the false-positive rate [8]. The use of an e-mail system in unlawful activities makes filtering spam e-mails an active area for research. A wide scope is present to improve existing methods and the development of new classifiers [9].

3. Literature Review

The origin-based classifiers are based on source or network details to classify a spam e-mail. The details of the e-mail sender are compared with the history of the spam sender, if the incoming mail is matched with spamming properties then categorized as spam e-mail. The information present in the header of an e-mail has various evidence of the sender or source of an e-mail and these can be utilized for spam classification [10]. Majorly 60.3% of junk e-mails do not have valid sender addresses, so the e-mail address of the sender can be

an indication of spam [11]. Wang et al. used header details without screening the content of the e-mail and found the origin-based method can efficiently classify spam e-mails compared to the content-based [12, 13]. The origin-based spam e-mail filtration will reduce the cost of computation and do not consume too many system resources [13]. Hu et al. used the sender validity, recipients, IP address, and e-mail's subject for spam e-mail classification using machine learning [14]. Agrawal et al. used the origin-based technique of spam e-mail filtration as efficiently as spam e-mail blocked before entering the network which makes the communication network to be highly enhanced [10]. Liu et al. suggested the use of IP address and e-mail address, e-mail subject, recipients, information can be used for classifying spam e-mails [15]. Krause et al. use the header feature because these are language-independent and it saves the expensive language processing of e-mail content skipped, resulting in quicker classification [16]. Machine learning can be used for spam e-mail classification [17, 18, 19]. Machine learning is the process to learn machines to observe, unstated, and represent information about some numerical phenomenon. Machine learning detects irregularities in the given data. The artificial immune system, C4.5, Maximum entropy model, Support Vector Machine (SVM), Memory-based learning, Bayesian, genetic algorithm, clustering techniques, ANN, Perceptron, Naïve Bayes (NB), Random Forest (RF) are some of the machine learning algorithms.

Sao et. Al. used the Naive Bayes classifier and Support Vector Machine classifier used for testing the dataset. The Naive Bayes classifier produces a better result as compared to the Support Vector Machine and when the number of datasets is increased and it was seen that the Naive Bayes classifier has a low error rate than the Support Vector Machine [20]. The Naïve Bayes is simple, easy to implement, and fast for convergence. The Naïve Bayes classifier can give the finest precision to classify spam e-mails [21]. Hassan et al. used the Naïve Bayes classifier on the header field-based attributes dataset. The spam e-mail filtering using the e-mail body is time-consuming, so it was simplified by only using the header-based features without affecting the spam filter performance which improved spam filtering [22]. Chhabra et al. proved the Support Vector Machine is a good classifier because it's a sparse data format and has acceptable recall and precision [23]. Lakshmi et al. found that the Support Vector Machine has a minimum variance in accuracy [24] and the false positive rate is the lowest. The f-measure, which relates to the accuracy, was highest using the Support Vector Machine [5]. Mohammed et al. preferred the Support Vector Machine algorithm over the Naïve Bayes algorithm. The Support Vector Machine algorithm is recommended as per findings to classify the spam and not-spam emails as it outperforms the Naïve Bayes algorithm in precision [25]. Naksomboon et al. suggested the Random Forest algorithm suitable considering keywords and behavior features [26]. The datasets available for spam e-mail classification are huge and for these larger datasets, Random Forest produces better results with lower error rates and higher precision [7, 27]. Goswami et al. concluded that the random forest is slightly more accurate than a Support Vector Machine but expensive in terms of time and space [28, 29].

4. Experiment

Framework

The objective of the paper is to evaluate and compare classification results based on the origin of an e-mail using different machine learning algorithms. An e-mail contains pieces of evidence about the origin of an e-mail this evidences can be used for spam e-mail classification.

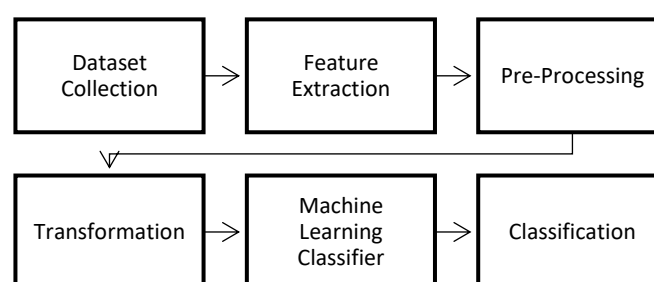


Fig. 28. Framework for Spam E-mail Classification

The classification framework is as shown in Fig 1, and the works of both are as follows.

Data Collection. - Dataset collected from the existing online repository and also e-mails from the personal yahoo mail was used to create an e-mail dataset. The details of the dataset with spam and Non-spam e-mails are shown in Table 1. The e-mails in each dataset are of .eml files and these e-mail files are used for feature extraction.

Feature Extraction. – The features are extracted from e-mail files from the dataset. The extracted features shown in Table 2 are used to build a feature vector that will be provided for a machine-learning-based classifier. An e-mail header has many fields but the focus is given to the dynamic features that are most likely to encircle spam e-mail activities. The features used for classification experiments are described as follows.

Length of Headers. - This feature contains the total number of headers in the e-mail.

Length of Subject. - The subject of an e-mail can be used for classifying an email as spam. The length of the subject is calculated after eliminating the stop words. The length of the subject is used as a feature; no spam probability of words in the subject is calculated or used because it will increase the burden of the classification process.

Message Content-Type. – The message content is of two types plain and multipart. The plain message type is represented by 0, whereas the multi-part is represented by 1 in the feature vector.

Delivered-To. – Value represents the mailbox in which this mail is to be delivered.

Reply-To. – The mail address to which a reply is to be sent.

TO, CC, and BCC. - The genuine emails enclosed the receivers' addresses in the “To” or “CC” field, while the spam emails left it blank [12]. One or more addresses can be used in these fields. The spammers duck the recipient’s addresses, as a large number of values in these fields are considered weird activities and it leads to detection by the anti-spam filters. The spammers use the “BCC” field to keep the list of recipients’ addresses, the reason is that the bcc field can have a higher number of recipients' addresses. A legitimate email has fewer recipients compared to a spam e-mail. A mail is sent to multiple users at the same time and the count of these multi-send mails is considered as the feature for classifying the spam e-mail.

Received IP Address and Email Address. – These fields contain IP addresses and email addresses of the receivers who received this mail before reaching to inbox. The server adds its information in this file automatically when it is received at the communication channel, ultimately it represents the path of the server path from where the mail is traveled.

From. - The “from” field of an e-mail has one or more email addresses of the authors of the email. In most spam cases the information in this field is not genuine.

Return Path. - Spammers habitually leave the real mail addresses to get the reply information from the receiver. The spammers use this information for future use.

X-Mailer. – Most of the programs for spamming provide invalid X-Mailer fields. The spammer keeps arbitrarily produced characters in the X-mailer field whereas when it is spam, then the information in the field is not legitimate [12].

No.	Header Feature	Value
1	Length of Headers	0...N
2	Length of Subject	0...N
3	Message Content-Type	0 Or 1
4	Count of Delivered To	0 Or 1
5	Count of Reply To	0...N
6	Count of To	0...N
7	Count of Cc	0...N
8	Count of Bcc	0...N

9	Count of Received IP Address	1...N
10	Count of Received Email Address	1...N
11	Count of From	0...N
12	Count of Return Path	0 Or 1
13	Count of X Mailer	0 Or 1

Table 17. Details of Feature from Dataset after Transformation

Preprocessing - Preprocessing is a vital step in classification as the e-mails files consist of special characters, stop words that needed to be removed before transforming them into the vector because these unessential data affect the speed of training and testing. Data cleaning, merger, alteration, and reducing are the main tasks in preprocessing. The pre-processing is done on the e-mails from corpora to alter and make it appropriate for the classification. This preprocessing phase involves the exclusion of punctuations and null characters, conversion to lower case. The special symbols like ~ [] ' < > () : { ` } \ / ! # \$ % ^ @ & * _ + = - ; , removed to avoid the unnecessary processing time.

Transformation – The values of features extracted from email files are used to transform the data required for the classification algorithm. In this case, the values of features after preprocessing are converted as shown in Table 1.

Classification. - Through the steps above, extract features transformed in the form of a vector of a fixed format. In this stage, the classification algorithm uses the vector form data for classification. The Naïve Bayes, Random Forest, and Support Vector Machine algorithms are used for spam e-mail classification.

Naïve Bayes -. The Naive Bayes classifier method is founded on the Bayesian theorem. The Naïve Bayes calculates the conditional possibilities of classes for a given instance and predicates the class with a higher probability. The basic concept of NB is considering at those tokens are present in the content; according to the tokens that are present in the e-mail classification is done as spam or non-spam.

$$P(c|\vec{x}) = \frac{P(c) \cdot P(\vec{x}|c)}{P(\vec{x})} \quad (1)$$

Given a token vector $\vec{x} = \{x_1, x_2, x_3, \dots, x_n\}$ of an e-mail, where are the values of attributes $x_1, x_2, x_3, \dots, x_n$ and n is the number of elements in the dataset. Every attribute is viewed as a particular token present or not. Let c indicate the category of e-mail to be predicted. The $P(c|\vec{x})$ is the chance that $\vec{x}$ belongs to c is as shown in (1) as to where, $P(\vec{x})$ denotes the probability of an arbitrarily chosen e-mail has vector $\vec{x}$ as its representation. $P(c)$ is a prior possibility of class c . $P(\vec{x}|c)$ signifies the probability of a randomly picked e-mail with class c has $\vec{x}$ as its representation.

Support Vector Machine. (SVM) - The Support Vector Machines is the supervised machine learning method used for classification. The SVM considers data as points, these points are placed in such a way the variance of the adjacent points is extreme. The testing instances are placed on the graph subject to which side of the margin they can be placed. The best separation has more remoteness between the nearby data points since a bigger margin specifies a lower generality error. Depending upon the margin testing points are classified as per class.

$$f(x) = \text{sgn}[(w^* \cdot x) + b^*] \quad (2)$$

where, w^* represents the orientation belonging to hyperplane and b^* represents the position of the hyperplane. x denote the input and $f(x)$ the maximum margin hyperplane.

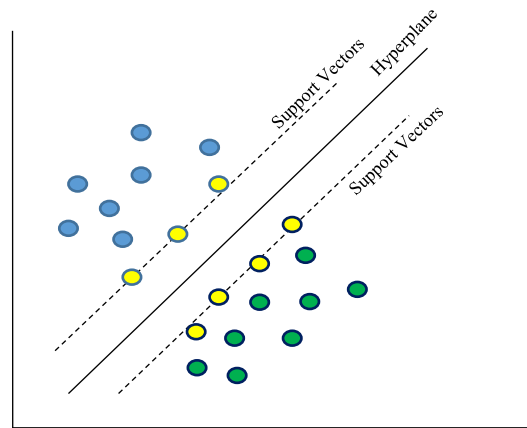


Fig. 29. Support Vector Machine with Linear Kernel

In the training phase, the SVM divides the labeled input data into two classes. A kernel function enables to map of the input space to a higher dimensional feature space. When liner kernel functions separate data linearly, that is, data of one class kept on a side and other class dots falls on the other side of the hyperplane shown in Figure 2. An optimal hyperplane is defined as one that maximizes the separation or margin between the support vectors and is calculated as Eq. (2).

Random Forest. (RF) - Random forest is an ensemble learning classification method suitable for problems involving the grouping of data into classes. During the training phase, several decision trees are constructed for the class prediction; by considering the voted classes of all the individual trees and the class with the maximum vote is considered to be the output of classification.

Dataset

Table 18. Dataset Details

Dataset	Period	Type of E-mail	Number of E-mails	Total E-mails
SpamAssassin ¹	29-06-2004 to 11-03-2005	Spam	2398	9349
		Non-spam	6951	
Enron ²	10-12-1999 to 06-09-2005	Spam	20988	46932
		Non-spam	31087	
Personal ³	22-06-2009 to 10-04-2020	Spam	1090	4416
		Non-spam	3326	

Numerous datasets are publicly available to be used by researchers for spam e-mail classification. Table 2 shows SpamAssassin, Enron, and personal e-mail account used in experimentation, which contains 60697 total e-mails with a spam rate of 40.32%. In the used dataset the spam rate is 34.49%, 44.74%, and 24.68% for SpamAssassin, Enron, and personal mail datasets respectively. Table 1 describes the total number of spam

¹ <https://spamassassin.apache.org/old/publiccorpus/>

² <https://www.cs.cmu.edu/~enron/>

³ E-mails from Author's Personal Yahoo Account

and non-spam e-mails, year span of e-mails circulated. The important aspect is the period of e-mail collection, Enron, SpamAssassin, and personal dataset in the duration of 2097, 255, and 3945 days. The Enron dataset has a higher number of e-mails, the longer duration of personal creates variation in content, size, spamming methods. Performance Measurement

Table 19. Performance Parameters

Classifier	Spam	Non-spam
Spam	Spam e-mails classified as Spam (True-Positive)	Spam e-mails classified as Non-spam (False-Negative)
Non-spam	Non-spam e-mails classified as Spam (False-Positive)	Non-spam e-mails classified as Non-spam (True-Negative)

The performance of the spam e-mail classification is measured in the form of precision, recall, accuracy, and f-score.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \times 100 \quad (3)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (5)$$

$$F - \text{measure} = 2 \times \left(\frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \right) \quad (6)$$

The accuracy is the ratio of true-positive and true-negative against all samples and is calculated by using the equation Eq. (3). The recall is the ratio of the true-positives found from the total of a true-positive and false-negative. The recall is calculated as per Eq. (4). Precision is the total true-positives out of the total predicted positive. To calculate precision, the equation Eq. (5) is used. The F-measure provides an accurate measure of the performance of the test by using both precision and recall and it is calculated as per Eq. (6).

5. Result and Discussion

The performance of Random Forest, Naive Bayes, and Support Vector Machine classification algorithms are compared. The experiment was conducted on three datasets namely SpamAssassin, Enron, and Personal yahoo mail dataset.

Dataset	Algorithm	Accuracy	Precision	Recall	F-Measure
Enron	NB	79.13	92.37	52.65	67.07
	SVM	95.29	94.51	93.77	94.14
	RF	99.08	99.14	98.57	98.86
SpamAssassin	NB	80.37	83.24	30.66	44.81
	SVM	82.99	83.87	42.80	56.68
	RF	95.94	92.71	91.56	92.13
Personal	NB	36.56	36.26	97.06	52.80
	SVM	72.04	65.38	50.00	56.67
	RF	69.89	60.00	52.94	56.25

Table 20. Classification Results

Table 4 shows that the Random Forest classifier showed the higher accuracy of classification for Enron and SpamAssassin datasets. The Random Forest classifier gave better accuracy on the larger dataset as compared to the smaller and the latest e-mail dataset. The Support Vector Machine classifier works better on the latest e-mail data. Similarly, the F-measure is the weighted average of precision and recall, if false negatives and false positives have similar rates, then accuracy gave better results. The f-measure obtained by RF is higher in all three datasets, representing the lower misclassification given by RF. The SVM algorithm is best suited to classify the better spam e-mails in this new era of spammers who develops regular new methods of spamming.

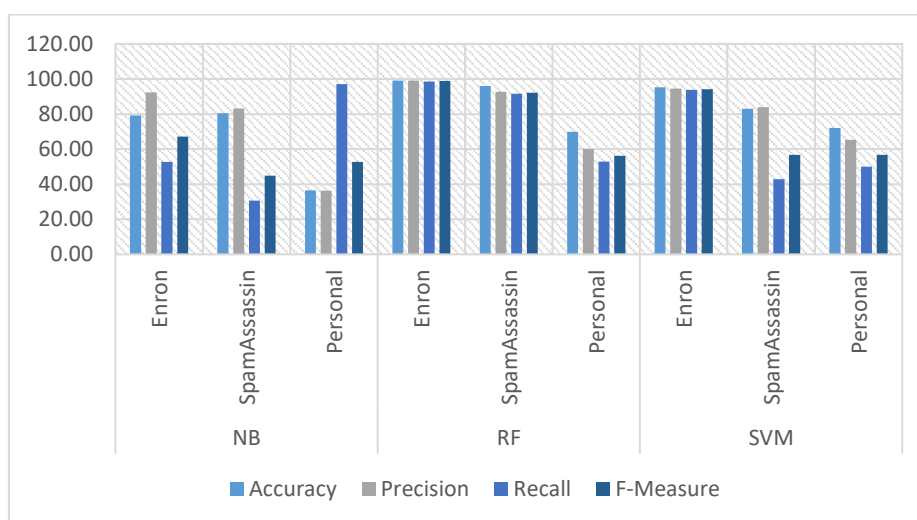


Fig. 30. Classification Results

Fig. 3 signifies that the Support Vector Machine is better than the Random Forest and Naïve Bayes classifier on the latest definition of e-mails which denotes minimum false positive and false negative values, this means the Support Vector Machine gave fewer false classification results. The higher recall rate by the Naïve Bayes classifier for the personal dataset signifies that the classifier has low false negative values in classification, whereas the other two classifiers are not up to the satisfactory performance. The higher recall rate by RF represents that the lower the false negative, it means fewer misclassification in spam e-mails are classified as legitimate e-mail. Support Vector Machine and Naïve Bayes higher recall on the SpamAssassin dataset has higher false negative. In classification, it is expected to achieve high recall and precision, the precision should ensure that less Non-spam e-mail should be classified as spam. Practically, the results are acceptable with high precision at a lower recall rate. The best precision is found by the Random Forest classifier on old and large datasets, the Support Vector Machine classifier has better precision on the latest e-mails dataset.

The results for the personal e-mail dataset denote that the Naïve Bayes and Support Vector Machine classifier works better than the Random Forest classifier on the latest e-mail dataset, Whereas the Random Forest classifier works better on older and larger datasets compared to the latest and smaller e-mail datasets. The Support Vector Machine on the unstructured data and comparatively fit high dimensional data. The Support Vector Machine and Random Forest both took a long training time on bulky datasets as compared to the Naïve Bayes classifier.

Receiver Operating Characteristic Curve.

The ROC curve displays the ability of the framework used to differentiate between classes. The ROC curve is planned as a true positive rate against the false-positive rate. The false-positive rate is plotted on the x-axis whereas the true positive rate is plotted on the y-axis. The true positive rate is calculated as the number of true positives divided by the sum of the number of true positives and the number of false negatives shown in Eq. (2) is called a recall rate or sensitivity. The ROC curve describes the method or framework is predicts the positive class when the actual result is positive.

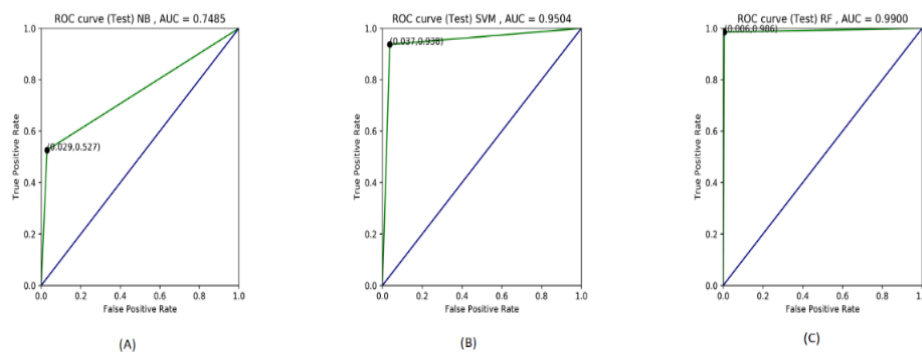


Fig. 31. ROC Curves for the Enron Dataset

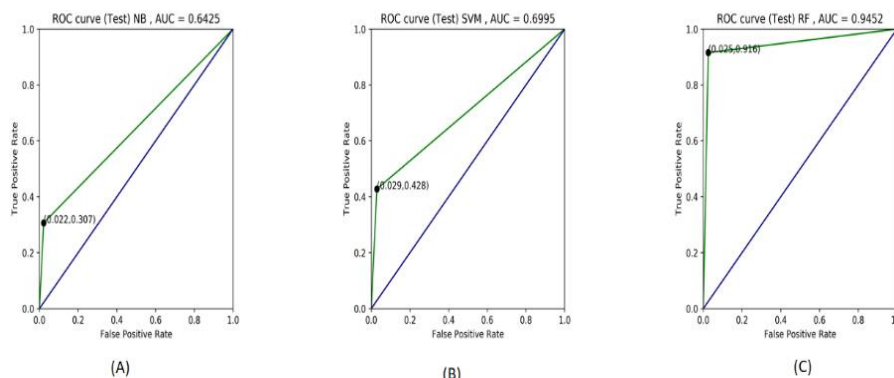


Fig. 32. ROC Curves for SpamAssassin Dataset

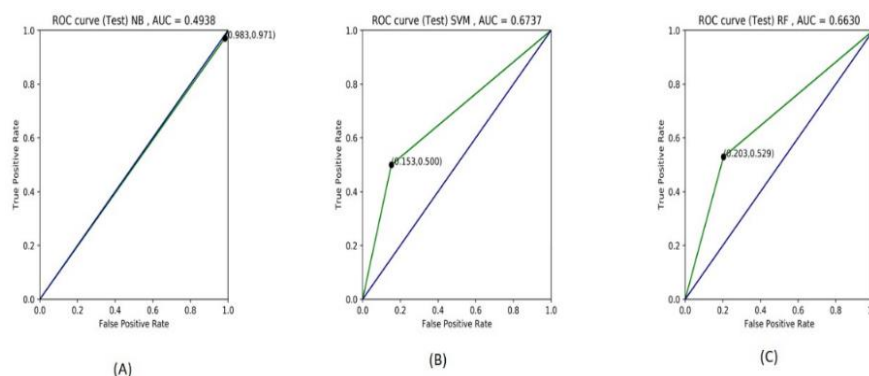


Fig. 33. ROC Curves for Personal Dataset

The ROC curve for the Enron dataset is reflected in the above Fig. 3. The AUC is 0.74%, 0.95%, and 0.99% for Naïve Bayes, Support Vector Machine, and Random Forest classifier respectively shown in Fig 4. (A), (B), and (C). The best ROC curve is for the Random Forest classifier shown in Fig. 4 (C), which has an AUC of 0.99% this represents that the Random Forest classifier has the highest true positives and lowest false positive rate better than Naïve Bayes and Support Vector Machine. Fig 4 (B) ROC shows that the false positive rate is lowest for the Support Vector Machine, this signifies that the Support Vector Machine has the lowest false positive rate compared to other classifiers. The AUC shown in Fig. 5 (A), (B), and (C) reflects that the TPR for the Naïve Bayes and Support Vector Machine on the SpamAssassin dataset is not that but promising whereas the Random Forest gives better TPR with the lowest FPR. The Random Forest classifier has the best classification amongst all three classifiers with minimum FPR. Fig. 6 shows the ROC and AUC for the personal dataset using the Random Forest, Support Vector Machine, and Naïve Bayes classifiers. The results for the personal e-mail dataset are not that encouraging but useful indications were found from the results. The Support Vector Machine gave the best AUC amongst all classifiers, which is different from the previous results for the Enron and SpamAssassin dataset where Random Forest gave the best results. The ROC and AUC in Fig 6 (B) reflect that the Support Vector Machine classifier works better than the Naïve

Bayes and Random Forest on the latest e-mail datasets. The false-positive rate is very discouraging for the Naïve Bayes classifier on the personal and the AUC for the Naïve Bayes is 0.49, which represents that the Naïve Bayes classifier is not suited for the latest definition of the e-mails when considering the origin-based features.

5.1 Conclusion

The evaluation of standard machine learning techniques and applicability to the problem of spam e-mail classification has been done. The origin-based classification ingests fewer resources to be efficient than content-based. The comparison of different classifier performances on the Enron, SpamAssassin, and personal corpus was presented. The Random Forest has a significant performance on older and larger datasets whereas the Support Vector Machine has a better performance shown on the latest, advanced and complex dataset. The classification results signify that the features, dataset size, and duration of e-mails have a reasonable role in machine learning spam e-mail classification.

References

- [1] G. V. Cormack and T. Lynam, "Spam Corpus Creation for TREC," in *Second Conference on Email and Anti-Spam*, California, USA, 2005.
- [2] N. Advilkar, P. Mane and D. Walunj, "Spam Mail Filtering," *International Journal of Advanced Research in Computer Engineering & Technology*, vol. 5, no. 1, pp. 99-104, 01 2016.
- [3] M. Siponen and C. Stucke, "Effective Antispam Strategies In Companies: An International Study," in *International Conference on System Sciences*, Kauia, HI, USA, 2006.
- [4] Namrata and Suman, "Review Paper on Spam Detection Antiphishing Techniques," *International Journal of Computer Sciences and Engineering*, vol. 6, no. 5, pp. 1156-1161, 2018.
- [5] S. K. Trivedi, "A Study Of Machine Learning Classifiers For Spam Detection," in *International Symposium on Computational and Business Intelligence (ISCBI)*, Olten, Switzerland, 2016.
- [6] V. Christina, S. Karpagavalli and G. Suganya, "A Study on Email Spam Filtering Techniques," *International Journal of Computer Applications*, vol. 12, no. 1, December 2010.
- [7] S. Jukić, J. Azemović, D. Kečo and J. Kevric, "Comparison of Machine Learning Techniques In Spam E-Mail Classification," *Southeast Europe Journal of Soft Computing*, vol. 4, no. 1, pp. 32-36, 2015.
- [8] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi and O. E. Ajibuwa, "Machine learning for email spam filtering: review, approaches and open research problems," *Heliyon*, vol. 5, pp. 1-23, 2019.
- [9] M. T. Banday, J. A. Qadri, T. R. Jan and N. A. Shah, "Detecting Threat E-mails using Bayesian Approach," *International Journal of Secure Digital Information Age*, vol. 1, no. 2, pp. 103-112, 2009.
- [10] N. Agrawal and S. Singh, "Origin (Dynamic Blacklisting) Based Spammer Detection and Spam Mail Filtering Approach," *International Conference on Digital Information Processing, Data Mining, and Wireless Communications*, pp. 99-104, 6-8 july 2016.
- [11] C.-C. Wang, "Sender and Receiver Addresses as Cues for Anti-spam Filtering," *Journal of Research and Practice in Information Technology*, vol. 36, no. 1, p. 3–7, 2004.
- [12] C.-C. Wang and S.-Y. Chen, "Using Header Session Messages to Anti-spamming," *computers & security*, p. 381–390, 2007.
- [13] J.-J. Sheu, "An Efficient Two-phase Spam Filtering Method Based on E-mails Categorization," *International Journal of Network Security*, vol. 9, no. 1, pp. 34-43, 2009.
- [14] Y. Hu, C. Guo, E. Ngai, M. Liu and S. Chen, "A scalable intelligent non-content-based spam-filtering framework," *Expert Systems with Applications*, vol. 37, p. 8557–8565, 2010.
- [15] P. Liu and T.-S. Moh, "Content Based Spam E-mail Filtering," in *International Conference on Collaboration Technologies and Systems*, 2016.
- [16] T. Krause, R. Uetz and T. Kretschmann, "Recognizing Email Spam from Meta Data Only," in *IEEE Conference on Communications and Network Security (CNS)*, 2019.
- [17] S. A. Saab, N. Mitri and M. Awad, "Ham or Spam? A comparative study for some Content-based Classification Algorithms for Email Filtering," in *17th IEEE Mediterranean Electrotechnical Conference*, Beirut, Lebanon, 2014.
- [18] L. Chih-Chin, "An Empirical Study of Three Machine Learning Methods for Spam Filtering," *Knowledge-Based Systems*, vol. 20, 2006.
- [19] H. Sohrab, A. Abtahee, I. Kashem, M. M. Hoque and I. H. Sarker, "Crime Prediction Using Spatio-Temporal Data," in *International Conference on Computing Science, Communication and Security*, Singapore, 2020.
- [20] P. Sao and K. Prashanthi, "E-mail Spam Classification Using Naïve Bayesian Classifier," *International Journal*

of Advanced Research in Computer Engineering & Technology (IJARCET), vol. 4, no. 6, pp. 2792-2796, 2015.

- [21] N. F. Rusland, N. Wahid, S. Kasim and H. Hafit, "Analysis of Naïve Bayes Algorithm for Email Spam Filtering across Multiple Datasets," in *International Research and Innovation Summit (IRIS2017)*, 2017.
- [22] M. M. Hassan and M. W. Hussain, "Header Based Spam Filtering Using Machine Learning Approach," *International Journal of Emerging Technologies in Engineering Research (IJETER)*, vol. 5, no. 10, OCT 2017.
- [23] P. Chhabra, R. Wadhvani and S. Shukla, "Spam Filtering Using Support Vector Machine," in *International Conference [ACCTA-2010]*, 2010.
- [24] R. D. Lakshmi and N. Radha, "Spam Classification Using Supervised Learning Techniques," in *the 1st Amrita ACM-W Celebration on Women in Computing*, 2010.
- [25] Z. Mohammed, F. J. A. Mohammed, I. M. P. Mohammed , M. Basthikodi and A. R. Faizabadi, "A Comparative Study for Spam Classifications in Email Using Naïve Bayes and SVM Algorithm," *Journal of Emerging Technologies and Innovative Research*, vol. 6, no. 5, pp. 391-393, 05 2019.
- [26] S. Naksomboon, C. Charnsripinyo and N. Wattanapongsakorn, "Considering Behavior Of Sender In Spam Mail Detection," in *6th International Conference on Networked Computing*, Gyeongju, South Korea, 2010.
- [27] P. Ozarkar and M. Patwardhan, "Efficient Spam Classification by Appropriate Feature Selection," *Global Journal of Computer Science and Technology Software & Data Engineering*, vol. 13, no. 5, 2013.
- [28] V. Goswami, V. Malviya and P. Sharma, "Detecting Spam Emails/SMS Using Naive Bayes, Support Vector Machine and Random Forest," *Innovative Data Communication Technologies and Application*, pp. 608-615, 2019.
- [29] Y. Tang, S. Krasser, Y. He, W. Yang and D. Alperovitch, "Support Vector Machines and Random Forests Modeling for Spam Senders Behavior Analysis," in *IEEE GLOBECOM 2008 - 2008 IEEE Global Telecommunications Conference*, New Orleans, LO, USA, 2008.

Load Forecasting Based Smart Grid Using Machine Learning

Pushpa¹, Sanjeev Indora², Mukhtiar Singh

^{1,2}Department of Computer Science and Engineering

¹Deenbandhu Chhotu Ram University of Science and Technology Sonapat, India.

³Department of Electrical Engineering

³Delhi Technological University, Delhi, India.

pushpaattwal03@gmail.com, sanjeev.cse@dcrustm.org

mukhtiorsingh@dce.ac.in

Abstract: With technological innovation, there may be an enormous increase in the usage of electrical power, which has become an issue not only for its development and but also for distribution. As a response, increased production raises the problem of the electricity grid by requiring higher accuracy, quality, stability, and global energy management concerns. These characteristics of an electric grid against smartness gradually became recognized as today's definition of "Smart Grid". A smart grid is defined as a consistent, convenient and flexible grid. Instant two-way supply of connectivity on energy, empowering power. Production to effectively maintain energy supply and transmission and give customers more power over energy outcomes. However, load forecasting has been among the most significant problems confronting the power industry since its inception in electricity generation. This review article analyzes the problem in load forecasting regarding forecasting techniques, adequate accuracy and error metrics. The existing smart grid-based load forecasting techniques are conferred and compared.

Keywords: Smart Grid, Load Forecasting, Energy Management, Machine Learning

I. Introduction

A smart grid is a form of electric network that will combine smart meters, smart appliances, renewable resources, and energy efficiency resources are just a few of the operation and energy measures available. In addition, the traditional grid is further used to transmit power from a small number of generators to a massive number of customers. The intelligent grid can be defined as a consistent, convenient, and flexible grid and instant two-way energy data supply, empowering electric power. Businesses to effectively control energy distribution, transmitting including empowering consumers to face more control over energy outcomes. Essentially, the term perception of a smart grid is designed to [1][2] give good clarity for low-voltage networks and allow consumers to be involved in smart meters and home appliances in the structure of a power grid. Using information technologies, a smart grid can provide energy efficiently and responsively to various conditions and actions. Generally, the aforementioned smart grid can relate to actions that take place somewhere on the grid, including power generation, transmission, distribution, and consumption, containing the adoption of appropriate strategies. For example, if an intermediate power failure action is performed on the system may drastically alter the power flow and restore the power supply service on a distribution system. Let us take a look at some other demand profile illustrations.

As the reduction in high demand and the smooth demand profile reduces the cost of plants and costs, in the long run, electricity could use actual rates to persuade other users to decrease their energy consumption; thereby, a full-fledged search profile is designed for high demand. More precisely, SG can be classified as two-way electronic connectivity, threat communication, also machine learning systems. Embedded fashion for every generation, delivery, channel installation, distribution to obtain a safe, convenient, and stable environment, durable, simple, along with reliable structure. With the growth of power grids, many things, such as cargo management, power trading, renewable energy, production [2][3] units, etc. It has contributed to the production of electricity and the use of energy. Predictability is among the most crucial efficiency factors and costs like a power system [3]. Predictability is required in the everyday activities of control centers to make strategic decisions on economic deployments, unit commitments, situational analysis, power connectivity, [3] power dealing along with security testing, etc. Usually, resource predictions develop this kind of weather prediction to interact with large quantities of energy loads at any given time. Load forecasting in a smart grid is a method of forecasting the energy required to achieve short-, medium-, and long-term demands. Forecasting correctly aids electric customers in managing their capacity, output, control, and energy use. Many modern technology prediction methods use conventional machine learning [4] [5] techniques to forecast power consumption. Large data platforms and a compatible distributed

computer need to harness the power of available solutions. Short-term (with an hour to seven days), medium-term (from seven days to a year and long-term (limits from 7 days to 1 year) load forecasting are the three types of load forecasting (ranges longer than one year) based on the forecasting time [6].

Machine learning is a method of data analysis where computers are asked to make the experience- based decisions. This training process uses 'learn' calculation methods. Techniques improve their performance dynamically as an amount of data becomes accessible. With the rise in the amount of large information, artificial intelligence has become a key problem-solving technique. It involves two methods: supervised learning, which develops a method with commonly identified input and output data to accurately [7] predict outputs, and unsupervised learning, which finds hidden patterns and inner structures within input data.

Table I. Comparison Of Existing Grid Vs Smart Grid

EXISTING GRID	SMART GRID
Few large power plants	Many small power producers
Centralized	Decentralized
Radial Network	Inter-connected Network
One-way communication	Two-way communication
Passive	Active
Electrical	Numerical
Artificial	Self-healing

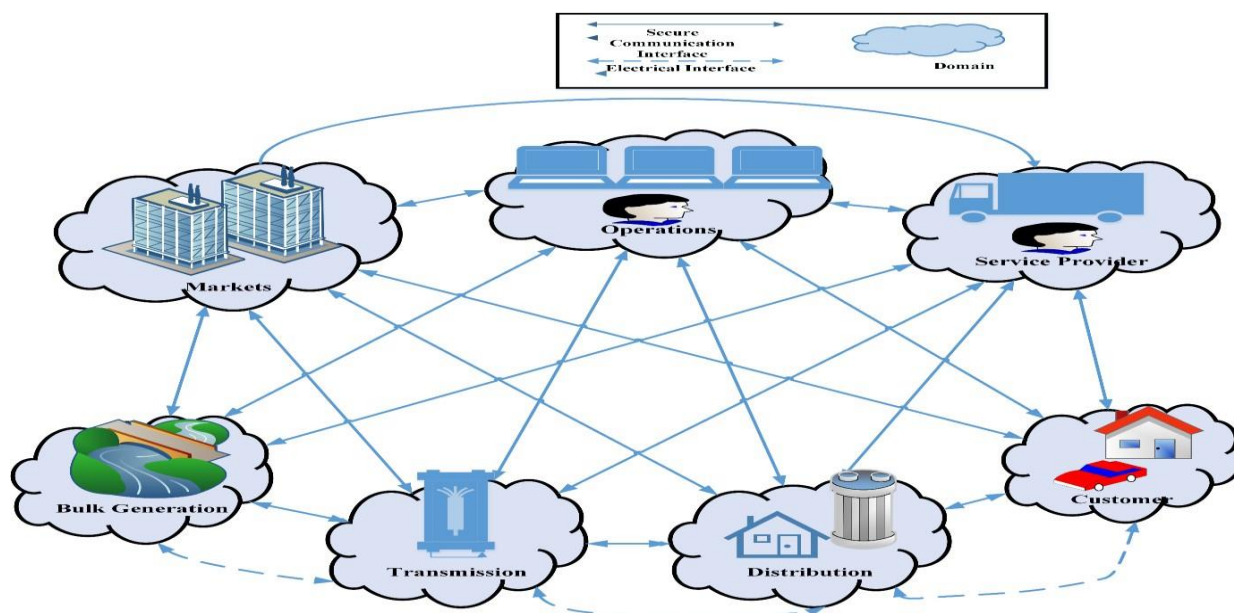


Fig.1. Working of smart grid

Information exchange between connectivity domains (displayed as a cloud) is being demonstrated in a smart grid infrastructure Fig.1[8]. Lower layer regions are linked to bulk generation, transmitter, production, and customer services. [9]. The higher supervisory level involves industries, businesses, and suppliers. Fig.1. shows an example of the bulk generating field classifies data to markets, operations, and service providers. Throughout the case of an electrical outage or system failure, operations and the market must take immediate

action [9]. The transmission field is primarily governed by operations that use knowledge exchange. To maintain power grid stability, its next transmission substation system should be outfitted with a Wide Area Situation Awareness (WASA) system that operates in tandem with an integrated Energy Management System (EMS). A distribution transformer framework may be used to set up future distribution domains. [8]. The customer domain has an interface with a service provider such as the internet [10].

2. Background

Many factors, such as system efficiency, electricity trading, renewable energy, energy resources, or other factors, also affect electric power generation and usage as power system networks have grown in size and complexity. Load forecasting is among the most crucial factors for ensuring such as [6] economic with reliable control device service. Load forecasting is the method of forecasting the quantity of energy required to maintain short-, medium-, and long-term requirements in a smart grid [11]. Accurate load forecasting assists electric utilities in managing production, process, control, and management of energy. Most communications systems method estimating techniques predict electrical load using traditional machine learning algorithms [4].

a. Features of Smart Grid

1. **Energy Efficiency:** - By optimizing existing networks, reducing waste streams, and maximizing the delivery of low-cost generation sources, smart grids would be able to generate more power from them [12]. Uniformity of local distribution with the inter-regional flow of energy will increase grid asset utilization, reduce grid congestion, and potentially reduce delay.

2. **Demand Response:** - Customers are raising their expectations of an energy provider and are looking for flexible energy, affordable alternatives, and a micro-generation option, all of which are currently not supported [13] by the electricity grid. Customers would have more direct control over the energy they use daily if the smart grid and associated smart technologies and devices were implemented.

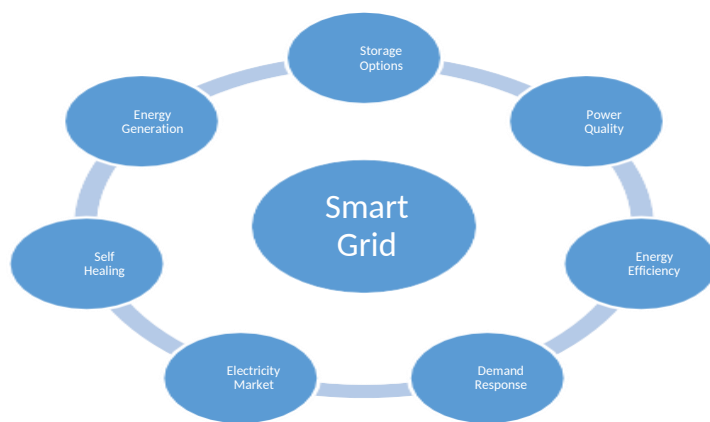


Fig. 2. Features of smart grid

3. **Energy Generation:** -As CO₂ levels rise and sources of energy deplete, it is critical that new types be discovered and integrated into the power supply. The existing electric grid cannot make it possible to draw energy directly from a range of energy sources using a variety of different technologies, but that is the power grid's goal [10].

4. **Electricity Market:** - Beginning a new market where renewable energy sources can be offered to consumers regardless of location [5]. Micro-energy manufacture would be a viable energy source, effectively providing a significant income source to local industries.

5. **Self-Healing:** - Automatically detects and responds to grid issues, ensuring quick recovery after disruptions [7]. The use of micro-grids also ensures that affected areas are isolated from the main network, resulting in minimal service interruption.

6. **Storage Options:** - All elements will be combined to provide consumers with access to both general and renewable energy resources. Energy storage will be useable to ensure that nothing is lost and that energy can be accessed later if needed [14].

7. **Power Quality:** - Power outages and poor quality cost businesses billions of dollars each year [1]. Smart grids can address issues by ensuring that everyone has access to electricity when solutions are needed at both the infrastructure and consumer levels.

b. Applications of Smart Grid

1. **Advanced Metering Infrastructure:** - A meter/utility contact view that is two-way. The AMI now includes two major components. First, through the utilization of actual meter reading information, as well as automatic meter reading (AMR) devices, is a first step for reducing costs [3]. Similarly, meter data management (MDM) offers an integrated platform of integration for various types of data sets. It allows data should be accessed to optimize market processes in natural environments and to exchange data with key business and operational applications to increase business productivity and selection [15].

2. **Distribution Management System:** - The application conceptually designs an electrical distribution channel as well as forecasts the effects of power failures, transmission, processing, frequency variation, or other variables [2]. It aids in the reduction of capital expenditure by demonstrating how to make better use of existing assets, allowing for maximum savings through demand response (DR), and improving the stability of the system [1].

3. **Geographic Data Network:** - The technique is primarily developed for such energy market's vital network modelling, design, and maintenance [5]. It gives a dynamic analysis of the framework can be created by integrating utility data and local mapping., allowing for cost savings through streamlined planning. Outage systems expedite outage recovery, allowing power to be rebuilt more quickly and keeping outage costs to a minimum [6].

4. **Compatible Equipment:** - Some earlier systems which are incompatible with energy-efficient technologies, must be upgraded. This could be a source of concern for utilities, regulators, and customers [9].

c. Characteristics of Smart Grid

1. **Reliability:** - The grid system's success is determined by customer demand, which is measured as reliability. This translates to a system with fewer flaws and errors, as well as a continuous supply of electricity. Smart Grid can detect any fault and allow the system to self-heal [16]. Traditional grids have problems with the interaction of renewable resources, micro grids, and demand response. As the size and complexity of these grids grow in response to need, this one becomes important to change assess their accuracy [17].

2. **Privacy:** -Protection is among the most challenging problems for smart grid development. Existing infrastructure automation, as well as remote monitoring and control, makes the grid more susceptible to cyber-attacks. [8].

3. **Energy Management System:** -Using two-way connectivity, the primary sector or consumer will communicate with the grid, a smart grid. It allows the consumer to use electricity more economically. It will not only help increase production performance, but will also help to develop system throughput. It helps the network minimize peak load and tension by lowering or moving power demands to external sources. It, therefore, offers consumers with a financial opportunity to use it.

4. **Metering:** - The smart automatic meter is associated with distribution system automation. Metering provides a platform for the power system model, there is two-way contact between both the customer and the distributor. They not only assist distributors in providing more accurate billing systems, but they also assist consumers in controlling with their utilization of electrical resources. Such Meters were outfitted to devices for monitoring, power quality control, and blackouts alerts. There are various drivers, such as rising prices following the deregulation of the periodical metering orders, the energy sector, and customer disappointment that encourage the use of smart meters [17].

5. **Self-Healing:** - A reliable smart grid must be capable of not only diagnosing but also remove faults in the grid to ensure a never-ending source of electricity. The self-healing network monitors the electrical characteristics of the grid using digital elements and actual contact placed across the network. Through this ability, the system becomes knowledgeable and effective in detecting potential problems that may arise naturally or as a result of human error [16]. Such expert systems immediately react to These anomalies are detected and identified until they become a major concern and cause a significant outage, and they naturally redirect energy transfer for service support unless an error is fixed.

d. Issues in Smart Grid

1. **Load Forecasting:** - It is a critical method in many areas of the energy sector. Several factors influence the load time series, such as climatic conditions, socioeconomic aspects, and random - effects. The load's non-linear and non-stationary attributes time series, as well as the industry's better performance requirements, take short load forecasting extremely difficult [14]. Every week load forecasts are essential for the power system's efficient and cost-effective operation.

2. **Storage Concerns:** - The whole group includes both bulk power renewables and distributed power generation. Because renewable energy generation cannot be uniform,i.e. intermittent and variable, it may necessitate storage [6]. The most common storage unit, the battery, has a really short life span of 4–5 years, to varying concerns. Pumped storage strategies in the United States, China, Japan, India, and Norway have efficiencies ranging from 70% to 85%. The issue with pumped storage strategies is that they require large areas as reservoirs, which are usually only available from the mountainside [14].

3. **Cyber Security:** - Identifying and closing security flaws before they cause a breach is crucial. The main goals of cybersecurity in the smart grid are being mentioned, namely availability, honesty, and confidentiality [16], [17]. The term "availability" refers to reliable and accurate access to databases and other information; "integrity" includes the protection of data from the improper interface, and "confidentiality" refers to the security of privacy from unauthorized access.

4. **Data Management:** - Smart Grid injections electric network with a massive number of meters, devices, and controllers Information from certain units or other sources, such as weather forecasting, surveillance cameras, and so on, helps operators improve their capabilities. Through accurate data analysis, a failure or harm can be avoided before it occurs. Consequently, this wide data set can be used for device action, alarms, supply, generation, and price forecasting, among another thing [10], [12], [13]. The volume of data collected is very large; for example, using a smart meter that allows readings every 15 minutes rather than once a month increases the data almost 3000 times.

5. **Communication:** - We include a variety of contact channels with smart grid implementation, each with its own set of drawbacks. The other technology has limited bandwidth and other works within a limited reach, a third has higher data loss, and one has limited capacity in underground installations. As a result, despite numerous benefits, smart grid interconnect technology already lacks a fully-tested solution [10]. On the smart grid network, communication protocols are not very well developed. GSM, GPRS, PLCC, 3G, ZigBee, Broadband over PLC, and other technologies fall into this category [3], [10]. GSM and GPRS have a range of up to 10 km but have slow data speeds.

3. Existing Load Forecasting Techniques Based On Smart Grid

Author [18] has discussed load forecasting in smart grid and proposed a technique ANN (Artificial Neural Network) and Back Propagation. This approach achieves power design optimization including energy savings [18]. These techniques are based on the effect of various learning algorithms of forecasting models using the backpropagation method throughout the neural network.

In this paper, the author has presented a methodology for SEDMS (smart energy distribution and management system) [19]. To design a novel approach efficient energy and greenhouse gas emissions can be minimized by improving power supply and maintenance. It is implemented in a MatLab simulator to examine the effectiveness of the existing methods which results in approximately 45.6% and 9-17% minimization of time

it takes for a service to react and power utilization [19].

The Author [20] has compared two techniques (ANN) artificial neural network and (SVM) support vector machine. This approach compares an application of SVM and ANN to mid-term forecasting models. The proposed mechanism continues examining in terms of different parameters energy consumption and accuracy [20]. And SVM has exceeded execution along with ANN when operating for quantities of 8760 tests or above, but performing lesser than ANN operating with quantities of 2160 tests or lower.

Author has designed clustering methods and (FCM) fuzzy c-means for the identification of loads [21]. This approach identification of loads along with classification and modification of inaccurate information, predictive modelling, or tariff setting is reviewed. The designed technique is a new approach that is most efficient, as well as modified conventional methods, should be established [21].

In this research article, author [22] has designed forecasting models with these techniques (PSO) Particle Swarm Optimization, (EMD) Empirical Mode Decomposition, (KELM) Extreme Learning Machine with Kernel and (EKF) Extended Kalman Filter. This approach EMD decomposes data from a dataset into many (IMF) Intrinsic Mode Function components. To forecast various types of IMF elements, two methods of forecasting algorithms (EKF or KELM) are used. Particle Swarm Optimization (PSO) is implemented in order to improve the model's variables [22]. Due to limited computation power, and execution design focused on off-line factor estimation, interval parametric development, or online predictive modelling was presented. The prediction model's reliability and utility are then increased through using four different micro-grids with varying users and capacities.

In this paper, author [23] has discussed two techniques short-term load forecasting and (ANN) artificial neural network. These methodologies are used to examine, recognize, measure, and review the output of intelligent machines focused predictive control systems, as well as to identify key factors. This approach obtained better results as it has increased forecasting accuracy.

Author has proposed AFC-STLF (accurate and fast converging short-term load forecasting) approaches for smart grid commercial processes [24]. Two techniques are used, (EDE) enhanced differential evolution and (MI) mutual information, for improvement in forecasting accuracy. This approach increases the forecasting accuracy and decreases the average implementation of the current bi-level forecasting policy.

The author [25] used the deep learning method in this paper to acknowledge that behavioral characteristics of FDI threats have used past information measured data, as well as to identify FDI threats and risks using collected attributes. And implemented recognition technique efficiently relieves predictions about future known attacks while obtaining high efficiency. In suggesting an optimizing method to clarify the characteristics of one form of FDI attack that influences the electrical theft power system's limited number of state parameters. We use simulation to show the performance of the new method approach using IEEE 118-bus test method [25]. By putting proposed method security mechanism to the test, able to put IEEE 300-bus test framework.

The author [26] presented a dynamic deep learning model that combines a (SWT) stationary wavelet transformation with the (LSTM) long short-term memory neural network methodology. Stationary wavelet transformation reduces uncertainty or optimizes information measures, that logically helps to increase the accuracy of LSTM predicting. Identification studies were carried out in support of the 'UK-DALE' project's data collection of real-world household energy usage data. The results show that the implemented system exceeds every comparable state-of-the-art process like the continuous process, (SVR) support vector regression, and (LSTM) long short-term memory, and CNN-LSTM (convolutional neural network with long-term memory) by distinct phase dimensions were tested at 5, 10, 20, and 30 minutes of three distinct error measurements [26].

Author [27] has examined a novel mechanism for predicting short-term load-dependent on its use of clustering techniques and a combination of various designs to improve overall system efficiency and performance. The methods incorporate various KF (Kalman filter), WNN and ANN (Wavelet, and Artificial Neural Network) structures. On the basis of clustering methods, six different approaches are designed. Simulations revealed that the proposed models performed better. Because the data used in this article is commercial, it has been scaled. The research project is put to the test using two different datasets from Egypt

and Canada. To validate the proposed work, the K-fold cross-validation approach was used. [27]. We suggest that 6-fold is the best size because when in comparison to 8-fold or 4-fold, it has the minimum mean (MAPE and nRMSE).

Author [28] proposed a design that would use the Intelligent IoT (Internet of Things) model, the capacity electronic devices may be used to develop a small solution for interconnecting electronic and electrical applications. This research is mostly concerned with the execution of prediction of smart grid loads, conventional correlation or machine-learning models are used. The state-of-the-art scenario for a scalable pattern classification scheme is designed, use limited, open-source resources has been developed takes to work with high accuracy. Moreover, the outcome of traditional and developed machine learning models supported on the computer is reviewed using various parameters such as percentage of errors, running time CPU core temperatures, and resource consumption [28]. A variety of unique machine learning algorithms that are thought can be appropriate for such presented system demonstrate overall impressive performance.

The author [29] proposed a quick and dependable hybrid electrical (FA-HELF) load forecasting system. The conventional FA-HELF structure is a three-module integrated structure. Second, to reduce redundancy, the hybrid classification approach is provided a combination of the random forest and Relief-F algorithm. Third, a kernel-based theory element design can be used to identify attributes in problem solving. dimensional reduction problem. At last, the support vector machine (SVM) forecaster is combined with a heuristic for timely and effective performance analysis [29] enhancer. The implemented FA-HELF system significantly improves forecast accuracy and convergence rate relative to other current forecasting models.

In this paper, author [30] has discussed the use of Apache Hadoop and Apache Spark are data analytics structures that are used for wireless computation to forecast loads use accessible big data. For distributed computing, MLib, the Apache Spark libraries is used for machine learning methods. MLib supports the testing of classical regression methods in contrast to survival and isotonic regression, these models include regression analysis, modified linear regression, decision trees, random forest, and gradient- boosted trees. Its research shows that Apache Spark generates greater precision when distributed. Prediction operations in extremely [30] competent planning and testing periods. In the load forecasting process, actual big data is used.

Table 2. Comparative Analysis Of Some Smart Gird Load Forecasting Techniques In Machine Learning

AUTHOR	TECHNIQUE	OBJECTIVE	PERFORMANCE IMPROVEMENT	SIMULATOR
Fang-Yuan XU, et.al.[18]	ANN(Artificial Neural Network), Backpropagation	The behavior of various learning algorithms for prediction models by back - propagation learning algorithm in Neural Network was studied.	Achieve optimization of power configuration and energy saving.	MATLAB
Byeongkwan Kang, Jinsung Byun[19]	SMS(Smart energy distribution and management system)	To improve Power efficiency or greenhouse gases emissions can be reduced by improving power distribution and maintenance.	The service throughput and energy consumption are reduced by approximately 45.6 % and 9-17 % respectively.	MATLAB

Xinxing Pan, Brian Lee[20]	Support vector machines (SVM) and Artificial neural networks (ANN)	To evaluate the results of SVM and ANN in mid-term load prediction.	SVM and ANN perform well when interacting with tests on a scale of 8760 or higher but underperforms when interacting with tests on a scale of 2160 tests or lower.	MATLAB
Chao Shen a,b ,et.al. [21]	A clustering technique for load classification	Load detection, along with analysis of unreliable data detection and correction, prediction, or tariff settings are review.	New method is more effective strategies, as well as improved conventional techniques, should be designed.	MATLAB
Jie Liu, Jianhua Zhang, et.al. [22]	PSO (Particle Swarm Optimization), KELM(Extreme Learning Machine with Kernel), EKF(Extended Kalman Filter), and EMD(Empirical Mode Decomposition	It is focused on off-line variable optimizing, period variable update, and on- line load prediction.	To achieve forecasting accuracy and time performance.	MATLAB
,Abbas Khosravi, et.al.[23]	Two methodology short- term load forecasting, and Artificial neural network are used.	Survey of predictive existing approaches based on Artificial Intelligence (AI)., as well as identify knowledge gap.	Increases forecasting efficiency.	MATLAB
Ashfaq Ahmad1,* et.al.[24]	Mutual Information (MI), EDE (Enhanced differential evolution)	To enhance prediction performance, a precise and (STLF) Short-Term Load Forecasting fast-converging method was used.	Increases the current bi-level forecast strategy improves forecast accuracy while decreasing average performance.	MATLAB
Youbiao He, Student Member, et.al. [25]	False Data Injection (FDI), (SCADA) Supervisory Control and Data Acquisition system.	To use the detected functions, identify FDI attacks in actual time	Achieve the high detection accuracy.	MATLAB
YANG DU 4 , WEI LI1 ,et.al.[26]	An ensemble SWT (stationary wavelet	The person household power utilization	Accuracy and energy consumption.	Python, Tensorflow and Keras

	transforms), LSTM (long short-term memory), and CNN-LSTM technique.	forecasting, to increase the accuracy of long short-term memory forecasting models		
H.H. Hamed Aly ,et.al. [27]	Wavelet and artificial neural network (WNN and ANN), KF (Kalman Filtering), techniques.	To enhance the total efficiency and quality of the system.	Improve the Performance of model.	R-studio and MATLAB
Amartya Mukherjee,*, et.al. [28]	Intelligent internet of things (IoT) and standard regression and artificial intelligence with smart grid demand planning and prediction.	To develop a Miniature load forecasting system that is sustainable for smart grid residential purposes.	Execution time, memory consumption, heating of the CPU core.	The Raspberry Pi ARM Microprocessor platform, Python
Khursheed Aurangzeb3, et.al. [29]	FA- HELF(fast and accurate hybrid electric load forecasting) structure is suggested.	A heuristic-based optimization approach is interconnected with the system to achieve fast and accurate Load forecasting (SVM) forecaster.	In contexts of estimate precision and rate of convergence,, FA-HELF outperforms other current forecasting models.	MATLAB, FA-HELF Framework
Haitham Abu-Rub† ,et.al. [30]	In Apache Spark and Hadoop are used as advanced analytics for distributed processing to estimate load based on current large datasets.	In addition to adaptive and isotonic regression, there is random forest, decision trees, linear regression, generalized linear regression,	Apache Spark achieved high accuracy when parallelizing the load operation.	JAVA and Hadoop ,Apache Spark
		and gradient- boosted trees.		

4. Conclusion

Load forecasting is a critical problem in smart grid. Smart grid load forecasting is a method of forecasting the accuracy needed to achieve demands in the short, medium, and long term. We also discussed the fundamentals and factors that influence forecast accuracy. And various load forecasting techniques used in smart grid are also discussed. This review article outlines the smart grid, its applications, features, characteristics and issues. Smart grid techniques and issues related to load forecasting and security approach. And comparative analysis of some smart grid load forecasting techniques in machine learning is discussed. The future work of the research article is that researchers can combine different load forecasting techniques for smart grid problem to get optimal results.

References

- [1] X. Fang, S. Misra, G. Xue, and D. Yang, "Smart grid - The new and improved power grid: A survey," *IEEE Commun. Surv. Tutor.*, vol. 14, no. 4, pp. 944–980, 2012, DOI: 10.1109/SURV.2011.101911.00087.
- [2] R. N. Anderson and H. Gharavi, "Smart Grid : The Future of the Electric Energy System," no. June, 2018.
- [3] R. Bayindir, I. Colak, G. Full, and K. Demirtas, "Smart grid technologies and applications," *Renew. Sustain. Energy Rev.*, vol. 66, pp. 499–516, 2016, DOI: 10.1016/j.rser.2016.08.002.
- [4] M. Ali, M. Adnan, M. Tariq, S. Member, H. V. Poor, and L. Fellow, "Load Forecasting Through Estimated Parametrized Based Fuzzy Inference System in Smart Grids," vol. 29, no. 1, pp. 156–165, 2021.
- [5] H. Daneshi, M. Shahidehpour, and A. L. Choobbari, "Long-term load forecasting in electricity market," *2008 IEEE Int. Conf. ElectroInformation Technol. IEEE EIT 2008 Conf.*, pp. 395–400, 2008, DOI: 10.1109/EIT.2008.4554335.
- [6] T. Anwar, B. Sharma, K. Chakraborty, and H. Sirohia, "Introduction to Load Forecasting," *Int. J. Pure Appl. Math.*, vol. 119, no. 15, pp. 1527–1538, 2018.
- [7] T. S. Bomfim, "Evolution of Machine Learning in Smart Grids," in *2020 IEEE 8th International Conference on Smart Energy Grid Engineering (SEGE)*, Oshawa, ON, Canada, Aug. 2020, pp. 82–87, DOI: 10.1109/SEGE49949.2020.9182023.
- [8] "NIST Framework and Roadmap for Smart Grid Interoperability Standards, Release 1.0," *NIST Spec. Publ.*, p. 145.
- [9] P. P. Parikh, M. G. Kanabar, and T. S. Sidhu, "Opportunities and challenges of wireless communication technologies for smart grid applications," *IEEE PES Gen. Meet. PES 2010*, no. Cc, 2010, DOI: 10.1109/PES.2010.5589988.
- [10] A. B. M. S. Ali, "Smart grids: Opportunities, developments, and trends," *Green Energy Technol.*, vol. 132, 2013, DOI: 10.1007/978-1-4471-5210-1.
- [11] S. Hosein and P. Hosein, "Load forecasting using deep neural networks," *2017 IEEE Power Energy Soc. Innov. Smart Grid Technol. Conf. ISGT 2017*, 2017, DOI: 10.1109/ISGT.2017.8085971.
- [12] F. Abdullah Asuhaimi, S. Bu, P. Valente Klaine, and M. A. Imran, "Channel Access and Power Control for Energy-Efficient Delay-Aware Heterogeneous Cellular Networks for Smart Grid Communications Using Deep Reinforcement Learning," *IEEE Access*, vol. 7, pp. 133474–133484, 2019, DOI: 10.1109/ACCESS.2019.2939827.
- [13] H. Mortaji, S. H. Ow, M. Moghavvemi, and H. A. F. Almurib, "Load Shedding and Smart-Direct Load Control Using Internet of Things in Smart Grid Demand Response Management," *IEEE Trans. Ind. Appl.*, vol. 53, no. 6, pp. 5155–5163, 2017, DOI: 10.1109/TIA.2017.2740832.
- [14] O. Flanagan, "Load Forecasting," *Undergr. Gas Storage Facil.*, pp. 9–24, 1995, DOI: 10.1016/b978-088415204-0/50002-9.
- [15] S. Fathi, R. Srinivasan, A. Fenner, and S. Fathi, "Machine learning applications in urban building energy performance forecasting: A systematic review," *Renew. Sustain. Energy Rev.*, vol. 133, p. 110287, Nov. 2020, DOI: 10.1016/j.rser.2020.110287.
- [16] S. Xia, X. Luo, and K. W. Chan, "A Framework for Self-healing Smart Grid with Incorporation of Multi-Agents," *Energy Procedia*, vol. 61, pp. 2123–2126, 2014, DOI: 10.1016/j.egypro.2014.12.090.
- [17] O. Majeed Butt, M. Zulqarnain, and T. Majeed Butt, "Recent advancement in smart grid technology: Future prospects in the electrical power network," *Ain Shams Eng. J.*, vol. 12, no. 1, pp. 687–695, Mar. 2021, DOI: 10.1016/j.asej.2020.05.004.
- [18] H. T. Zhang, F. Y. Xu, and L. Zhou, "Artificial Neural Network for load forecasting in smart grid," *2010 Int. Conf. Mach. Learn. Cybern. ICMLC 2010*, vol. 6, no. July, pp. 3200–3205, 2010, DOI: 10.1109/ICMLC.2010.5580713.
- [19] J. Byun, I. Hong, B. Kang, and S. Park, "A smart energy distribution and management system for renewable energy distribution and context-aware services based on user patterns and load forecasting," *IEEE Trans. Consum. Electron.*, vol. 57, no. 2, pp. 436–444, 2011, DOI: 10.1109/TCE.2011.5955177.
- [20] X. Pan and B. Lee, "A comparison of support vector machines and artificial neural networks for mid-term load forecasting," *2012 IEEE Int. Conf. Ind. Technol. ICIT 2012 Proc.*, pp. 95–101, 2012, doi: 10.1109/ICIT.2012.6209920.
- [21] K. L. Zhou, S. L. Yang, and C. Shen, "A review of electric load classification in smart grid environment," *Renew. Sustain. Energy Rev.*, vol. 24, pp. 103–110, 2013, DOI: 10.1016/j.rser.2013.03.023.
- [22] N. Liu, Q. Tang, J. Zhang, W. Fan, and J. Liu, "A hybrid forecasting model with parameter optimization for short-term load forecasting of micro-grids," *Appl. Energy*, vol. 129, pp. 336–345, 2014, DOI: 10.1016/j.apenergy.2014.05.023.

- [23] M. Q. Raza and A. Khosravi, "A review on artificial intelligence-based load demand forecasting techniques for smart grid and buildings," *Renew. Sustain. Energy Rev.*, vol. 50, pp. 1352–1372, 2015, DOI: 10.1016/j.rser.2015.04.065.
- [24] A. Ahmad, N. Javaid, M. Guizani, N. Alrajeh, and Z. A. Khan, "An Accurate and Fast Converging Short-Term Load Forecasting Model for Industrial Applications in a Smart Grid," *IEEE Trans. Ind. Inform.*, vol. 13, no. 5, pp. 2587–2596, 2017, doi: 10.1109/TII.2016.2638322.
- [25] Y. He, G. J. Mendis, and J. Wei, "Real-Time Detection of False Data Injection Attacks in Smart Grid: A Deep Learning-Based Intelligent Mechanism," *IEEE Trans. Smart Grid*, vol. 8, no. 5, pp. 2505–2516, 2017, DOI: 10.1109/TSG.2017.2703842.
- [26] K. Yan, W. Li, Z. Ji, M. Qi, and Y. Du, "A Hybrid LSTM Neural Network for Energy Consumption Forecasting of Individual Households," *IEEE Access*, vol. 7, pp. 157633– 157642, 2019, DOI: 10.1109/ACCESS.2019.2949065.
- [27] H. H. H. Aly, "A proposed intelligent short-term load forecasting hybrid models of ANN , WNN and KF based on clustering techniques for smart grid," *Electr. Power Syst. Res.*, vol. 182, no. February, pp. 106191–106191, 2020, DOI: 10.1016/j.epsr.2019.106191.
- [28] A. Mukherjee, P. Mukherjee, N. Dey, D. De, and B. K. Panigrahi, "Sustainable Computing : Informatics and Systems Lightweight sustainable intelligent load forecasting platform for smart grid applications," *Sustain. Comput. Inform. Syst.*, vol. 25, pp. 100356– 100356, 2020, doi: 10.1016/j.suscom.2019.100356.
- [29] G. Hafeez, I. Khan, M. Usman, K. Aurangzeb, and A. Ullah, "Fast and Accurate Hybrid Electric Load Forecasting with Novel Feature Engineering and Optimization Framework in Smart Grid," *Proc. - 2020 6th Conf. Data Sci. Mach. Learn. Appl. CDMA 2020*, pp. 31– 36, 2020, DOI: 10.1109/CDMA47397.2020.00011.
- [30] D. Syed, S. S. Refaat, and H. Abu-Rub, "Performance evaluation of distributed machine learning for load forecasting in smart grids," *Proc. 30th Int. Conf. Cybern. Inform. K 2020*, 2020, DOI: 10.1109/KI48306.2020.9039

Fertilizer Recommendation and Prediction using Machine learning

Sonu Sharma¹ and Poonam Mittal² and Payal Gulati³

¹ Research Scholar, J.C. Bose University of Science and Technology, YMCA

² J.C. Bose University of Science and Technology, YMCA

³ J.C. Bose University of Science and Technology, YMCA
sonusharma2095@gmail.com

Abstract. Agriculture plays a major part inside an economy's life. When more than 70 per cent of the population relies on agriculture, it is like a spine for developing countries like India. The growing issue among Indian farmers is that they pick and manually apply estimated quantities of fertilizers. The addition of excess or insufficient fertilizer can damage plant life and reduce yield. Agriculture is highly dependent on soil and environment character. Hence advancing in this field becomes important. Many factors like soil, temperature, irrigation, Fertilizers, and pesticides are responsible for increasing crop growth. We used soil parameters to increase crop production, since it is a key agricultural factor. In the case of a shortage, fertilizers or compost are mix in the soil to hold supplement levels in the dirt. This paper includes a study of numerous machine learning techniques for fertilizer suggestions used on the soil dataset for agriculture. I concentrated primarily on the specific soil parameters such as the importance Humidity, Temp, Moisture K, N, P etc. In this paper, we addressed the inappropriate use of fertilizer in agriculture and how it is very well addressed by Machine learning (ML) and other AI deep learning algorithms like random forest, recurrent neural network, Support vector machine and long short term memory (LSTM).

Keywords: Machine learning, recurrent neural network, Fertilizer prediction and recommendation, Support vector machine, Soil parameters, NPK, long short-term memory.

1. Introduction

The agriculture area is the fundamental donor in Indian economy and doing admirably in white, green and blue upheaval. Indian Agriculture has been the backbone of the Indian economy for thousands of years, and continues to be. It has played a key role in the nation's overall economic growth and affects the quality of living for a significant chunk of the population in the region. This sector has significant potential to drive growth as of today. Agriculture has been more than just a work over the years, For others it has become a way of life. With the constantly evolving situation of the emerging technology, it has become one of the nation fastest rising markets. Agriculture plays a very significant role in supplying food surplus to the growing population, leading to the development of resources, supplying industry with raw materials, selling consumer goods and making a big contributor to foreign trade. India is the largest producer of certain grains, even as it is the second-largest producer of pulses. If we equate India's yield per hectare with other nations, the world's lowest is 3 tones. The causes are lack of irrigation facilities, poor control of the pests, and soil degradation due to natural disasters, insufficient storage facilities, and lack of transport and deficiency of resources. However, most of Indian crop producers are poor and illiterate. They are unable to afford high quality seeds and modern fertilizers, and are unable to take decisions that help them improve crop production. Information technology should be used for the collection and analysis of such a vast volume of data. Specific machine learning techniques may be used to find guidance surrounding crops and fertilizers.

We can brief the serious issues looked by Indian Agriculture as given below

- i. Illiteracy: Some of the farmers are illiterate; they don't know how to use the fertilizer.
- ii. Cost of fertilizer and manures: Cost of fertilizer and pesticides are very high, farmers are not able to afford the fertilizer.
- iii. Scarcity and cost of farming information sources: lack of technology and modernization, farmers are still relies on the old techniques.

- iv. Lack of knowledge: They don't know how to use the modern equipment, lack of instruction and training.
- v. Uneconomic size of land holding: Land is unequally distributed among the farmers.

The motivation to carry out this work is to examine techniques of machine learning can be utilized to solve complicated agricultural problems. Agricultural soil feature high- light the soil properties, especially those identified with the organic matter such as water, Temp, K, N, P, Moisture [1] and fundamental plant nutrients, K, P, N. From crop reaction to treatment suggested micronutrient for plant development are water, Temp, K, N, P, Moisture [2]. The local profitable limit of soil is subject to fundamental properties like synthetic & Physical properties and prediction be conceivable [3]. So the principle purpose is to analysis, establishment as well as adjustment of crops development factors including parameters of soil properties. This paper focuses on applying machine learning algorithms which can help in predicting the fertilizer based on the soil parameters. It will help farmers increase their crop production. Recommendation of fertilizers is dependent on soil nutrients. This paper uses machine learning as well as deep learning algorithm to suggest crops because it operates well on large data-sets along with can oversee missing values. This Research paper explains however the long short term memory & recurrent neural network can be used to decide best suitable fertilizer for the crop depending on the soil nutrients (Temp, Humidity, moisture, N, P, K) material present in the soil.

The main organization of this paper is coordinated as follow. Section II presents detailed literature review of work or efforts done by the all researchers in this area and techniques used by them and their loopholes. Section III gives a brief introduction about the methodology and Proposed work. Section IV indicates Experimental outcomes. Section V provides the conclusion of the Paper.

1.1 Analysis of soil for fertilizer Recommendation

Between numerous trust areas of agricultural research such as crop production, demand analysis and price prediction, prediction of rainfall, soil classification plays an important role as soil analysis makes a significant contribution to helping the farmers. Machine learning is an interesting and new area of study in agricultural soil datasets. A soil's productive potential is dependent upon soil fertility. Soil fertility is the soil's capacity to sustain plant growth and is determined by the volume of macro and mi- crop nutrients, Temp, K, N, P, Moisture, and Humidity etc. Fertilizers are applied to soil to retain soil nutritional levels in case of deficiency. Better crop production relies heavily on fertilizer inputs. Not only does fertilizer increase crop production but it also creates organic soil content. This paper analyzes the different machine learning methods used to evaluate the fertilizer dataset for fertilizer recommendations and prediction.

2. Literature Review

Several studies have been done by the researchers to analyze different algorithm. Researchers made a comparable study of different machine learning methods for in- crease the accuracy in fertilizer prediction and recommendation.

Usman Ahmed, Jerry Chun-Wei Lin, Gautam Srivastava, Youcef Djenouri[4] Published in 2021 This paper proposed on improved genetic algorithm (IGA) to prescribe an ideal setting for supplements to various yields. The algorithm used for recommendation is optimization scheme, which includes a local investigation as well as exploitation methodology. The outcomes in reality datasets showed that the proposed IGA strategy could perform better against the standard suggestion. In this way, the calculation can improve the yield and keep up with supplement level. The main drawback of this paper is, the application is largely restricted to specific single crop on account of the worldwide changeability these fundamental supplements.

Jenskie Jerlin I. Haban, John Carlo V. Puno, Argel A. Bandala , Robert Kerwin Billones , Elmer P. Dadios , Edwin Sybingco[5] published in 2020 The main of this paper is to build a fuzzy logic system that helps in providing the exact quantity of fertilizer or supplements to soil. Parameter used in this study are season, phosphorus and potassium as well as Nitrogen. The specialists proposed four sorts of compost to use in this paper like Complete, Urea, Solophos and Muriate of Potash. The main conclusion of this study is that the work is done only on a single crop i.e Rice which is the only drawback of this paper.

Devdatta A.Bonde, Mr. Santosh Mahagaonkar[6] Published in 2019 The work proposed in this paper uses the Machine learning methods and techniques for example Support vector machine as well as random forest utilizes for fertilizer recommendation relied on soil nutrients. Soil classification act as a significant factor in recommendation. In case of soil classification accuracy delivered by the system is very high and efficient.

S. Bhanumathi, Mathoor Vineet, N.Rohit [7] published in 2019 The work proposed in this paper uses numerous data mining approaches, is a rising examination field in crop production analysis. Production expectation is significant issue in horticulture which separates the diverse related properties like area, pH. From which alkalinity of soil is settled and resolved. Alongside it, rates of enhancement or parameters like nitrogen phosphorus as well as potassium. All of these characteristics of data will be research, train the given information with various reasonable AI machine learning approaches for building the model. The structure goes with a model to be definite and exact in expecting crop yield and convey the end client with appropriate idea about require manure proportion subject to ecological as well as soil parameters of land which helps to improve the harvest yield as well as increment the farmers earning.

M.S. Sirsat, E. Cernadas, M. Fernández-Delgado, S. Barro [8] published in 2018 This paper proposes the fertilizer recommendation based on the soil nutrients by using soil fertility indices with the help of machine learning techniques, extremely randomized regression tree. The main advantage of this technique is variation utilizes 3 dataset training, testing and validation, rather than traditional cross-validation, which just utilize training as well as test dataset. By applying 2 dataset it achieve high exactness and best execution. The conclusion of this paper is that performance we will get from final complete model for other data-sets would be relied upon to be lower which is a main drawback of this paper.

Anna Chlingaryana, Salah Sukkarieha, Brett Whelanb [9] Published in 2018 This paper showed that the crop production is totally depends on the level of nitrogen present in soil. Nitrogen is the essential nutrient for the soil. These days remote detecting framework is the most part utilized in dynamics. The work proposed by the author is to for see or predict the crop production depend on nitrogen level substance present inside the soil with the help of remote detecting along with machine learning algorithm. The acquired outcomes are back-propagation neural system is performed better with high effectiveness.

S. Sirsat , E. Cernadas , M. Fernández-Delgado , R. Khan [10] Published in 2017 In this research paper, various ML algorithm are utilizes like Random forest, neural network and support vector machine. Extreme learning machine on soil parameters and soil nutrients to recommend or predict the appropriate amount for fertilizer and ideal crop. Random forest achieves the best performance for all the problems. The major conclusion of this research paper is that the performance of classifier is affected by using other classifiers.

E-Manjula[11] Published in 2017 In this paper author mainly focus on the soil nutrients and parameters for example potassium, calcium, zinc, magnesium etc. As we know that the soil fertility is completely based on the nutrient present in the soil because nitrogen is the vital element for the plant growth. The techniques used for prediction or suggestion of fertilizer are machine learning approaches for example Naïve bayes as well as Decision tree, as it is the part of random forest so we can anticipate the fertilizer with the help of voting. The main conclusion of this paper is that the classification algorithm performance is completely depend on the execution time.

Rohit kumar Rajak et al[12] Published in 2017 This methodology is explained with the help of soil dataset gathered from the field, The crop taken into consideration were delivered by horticulture specialists, accomplishment of parameters, for example, soil with the help of soil testing data-set. The information we get by the soil testing lab data-set given to proposal framework to utilize the whole information and data as well as train the model along the voting techniques (Random forest) utilizing SVM & ANN while trainer to suggest Crop for the site explicit parameter for high precision as well as effectiveness.

Vincent de Paul Obade , Rattan Lal[13] published in 2016 In this paper the author focus on the technique for soil assessment using partial least square regression method for predicting the fertilizer. The experimental result showed that the PLSR is very fast in prediction but the drawback of this paper is Soil attribute is critical for making informed decisions on appropriate soil quality management.

Niketa et al [14] Published in 2016 Season plays a significant role in overall crop production. Droughts can Detroit farmer's crop production. This issue could be re- solve by few data mining methods to assist the crop producer with suggesting the crop for the better production. Previous information can be used as a base for future information. They utilized LWL classifiers in which we used WEKA tool to arrange the outcomes. The primary figures take thought are least temperature, greatest and most Prominent temperature, normal temperature as well as prior year's harvest data information with yield information. Utilizing LWL classical. They arranged the past information divides in two different classes which are high return and low production. The acquired outcome as the harvest yield forecast utilizing classifier i.e. SMO offers less exactness whenever contrasted with LAD tree and j48.

S.S.Baskar,L.Arockiam ,S.Charles [15] published in 2013 In this study several AI techniques are used for recommended the fertilizer. Work classifies the soil dataset using various classification algorithms such as regression, J48, Naive Bayes. Accuracy level of j48 is 93.86% and this is the best result it can give. The fertilizer will be recommended depend on the soil sample.

Shivnath, Dr. Santanu koley[16] published in 2014 In this paper work is done on the soil parameters and plant nutrients for fertilizer prediction. Results will represent that the proposed revision strategy empowers improving fundamentally dirt properties expectation exactness and performs much better than generally utilized strategies comprising in programmed weight updating handling. The main disadvantage of BPN is this is more self-employed and so much training is needed to train BPN.

Table 1. Comparison table on various existing methodologies, their advantages, disadvantages.

S.No	Name of Author	Year of Publication	Methodology/ Technology	Proposed Work	Advantages	Disadvantages
1.	Usman Ahmed, Lin, Gautam Srivastava, Youcef Djenouri Jerry Chun-Wei [4]	2021	Uses the Improved Genetic algorithm	Fertilizer recommendation is based on the optimization scheme	Proposed IGA strategy could perform better against the standard suggestion.	The main drawback of this paper is, the application is largely restricted to specific single crop
2.	Jenskie Jerlin I. Haban, John Carlo V. Puno, Argel A. Bandala , Robert Kerwin Billones , Elmer P. Dadios , Edwin Sybingco[5]	2020	Uses the Fuzzy logic system	Fertilizer recommendation is based on the fuzzy season, nitrogen and potassium as well as phosphorus	Wet and dry season is considered in the Recommendation	Work done in only a single crop i.e Rice
3.	Devdatta A.Bonde, Mr. Santosh Mahagaonkar	2019	Use the Machine Learning techniques such as SVM, Random forest	Fertilizer Recommendation based on the soil nutrients. Soil classification is the main factor in the recommendation.	In soil classification Random forest the high accuracy and efficiency.	Support vector Machine is less accurate as compare to the Random forest.
4.	M.S Suchitra,Maya L.pai	2019	Use the Extreme Learning parameters for predicting the soil nutrients	Work on the classification of soil fertility indices.	Extreme machine learning is the best for finding the soil fertility.	Work done in Only kerala soil and single crop is used for Prediction.
5.	S. Bhanu- ma- thi,Mathoor Vineet, N.Rohit	2019	Use the various Data mining techniques	For training the model data mining techniques are used creating the model ML algorithm are used	Attributes such as location, Ph value used for finding the alkanity of soil and attributes of soil are Nitro- gen, potassium	Other parameters like calcium, zinc are not included.

					and Phosphorus.	
6.	M. Fernández-Delgado , E. Cernadas, S. Barro , M.S. Sirsat,	2018	Use the machine Learning, regression, extremely randomized regression tree.	Fertilizer Recommendation based on the soil Nutrients by using soil fertility indices.	This variant Uses 3 dataset i.e testing, Training & validation rather than classical cross- validation, problem with cross validation is it uses training & test datasets. By utilizing 2 da- tasets it accomplish best performance.	The performance of the final complete model for rest of the data-sets would be depended upon to be lower.
7.	Rafael hernandezmore- no, olmer garcia	2018	Use single Layer Multilayer perceptron.	Work on pre-Processing and training of Data.	ML shows Excel- lent performance in solving classification and prediction problems.	A small number of change in data affected the whole Result.
8.	Anna Chlingaryana, Salah Suk- karie- ha, Brett Whelanb	2018	Machine learning and Remote detection techniques	Predict the Crop production on the basis of nitrogen level Content in the soil.	Among the all machine learning techniques BPN get the best result.	Self-training of Back propagation neural net- work.
9.	R. Khan., S. Sirsat, M. Fer- nández Delga- do, E. Cerna- das	2017	Uses random forest, decision tree, support vector ma- chine, Extreme learning machine.	Use Soil Parameters and Soil nutrients in order to recommend suitable amount of fertilizer and preferable.	Random forest achieve get the best performance for all problems.	Performance for model influenced by utilizing different classifiers.
10.	HD Aparna,Dr. Kavitha KS,Dr. Ka- vitha C	2017	Simple linear regression, multiple regression ,Multilayer perceptron, Support vector machine	Use machine Learning techniques for predicting the fertilizer usage in agriculture production.	MLP is increasingly precise in anticipating or predicting the result.	Not work on Soil parameters, Temp and rainfall.
11.	E-Manjula	2017	Use the approaches of Machine learning techniques such as Decision tree and Naïve bayes.	Use soil nutrients and parameter such as P,K, Ca, Mg.	Classification algorithm performance is completely based on execution time.	Hybrid approach some- times leads to the slow performance.
12.	Rohit kumar Rajak et al	2017	Support vector ma- chine, random forest and ANN	Use crop parameter to recommend the fertilizer.	SVM, ANN and voting technique is the best for accuracy.	Multiple cropping pattern.
13.	Vincent de Paul Obade , Rattan Lal	2016	Partial least square regression method.	Partial least square regression , mean square method for predicting the fertilizer.	PLSR is quick in pre- diction.	Soil attribute is critical for making well decisions on appropriate soil quality management.
14.	Niketa et	2016	Data mining techniques and tool WEKA, j48 LAD tree, LWL classifier	Crop production forecasting of tropical wet and dry zone.	LAD tree show the high accuracy	LWL classifier shows low accuracy.

15.	Shivnath , Dr. Santanu	2014	Supervised learning as well as Back propagation neural Network.	Uses Soil fertility parameters and Plant nutrients for fertilizer prediction	Soil properties forecast precision and performs better than customarily utilized strategies comprising in programmed weight updation preparing	The main disadvantage of BPN is this is more self-employed and so much training is needed to train BPN.
16.	S.S.Baskar, S.Charles,,Aro ckiam	2K13	Data mining techniques	This paper focuses an analysis of soil dataset and parameters using data mining techniques for prediction. technique	Classical linear regression is a technique which is known for the best result but in this paper least median square give the high accuracy than Classical linear regression	Not focus on Multiple cropping pattern.

2.1 Problem Identification

- Existing models mainly focus on either prediction or recommendation but both of them are not taken in to consideration at the same time..
- Not focus on multiple cropping patterns.
- Soil parameters are taken as an important factor whereas other factors such as rainfall, temp are ignored.
- For the last two decades, continuing research has been conducted to solve this problem.
- However, this field of study is still a focus of intense research due to shortcoming of existing methods and there are still many directions to been explored.

3. Proposed Methodology

This section deals with an approach of fertilizer prediction and recommendation using Support vector machine (machine learning) along with LSTM and RNN (Deep learning). For fertilizer prediction LSTM and RNN are used and for recommendation SVM is used. The input dataset is set of soil features example soil, nitrogen, potassium, phosphorus, moisture, Temp etc. The climatic and soil factor are taken in to consideration. It is a complicated task to distinguish the best appropriate when there are more than one alternative accessible. Consequently, by deep learning, exact fertilizer can be Predict LSTM and RNN are used for Prediction of accurate fertilizer de- pending on the soil attributes or parameter and climatic factors. Proposed methodology is also compared with the existing techniques like Random forest, Support vector machine. The Proposed methodology consists of following steps:

- Collection of Data
- Pre-processing of Data
- Model Selection (LSTM-RNN, SVM)
- Train the model.
- Evaluation or Assessment of Model.
- Make Prediction.

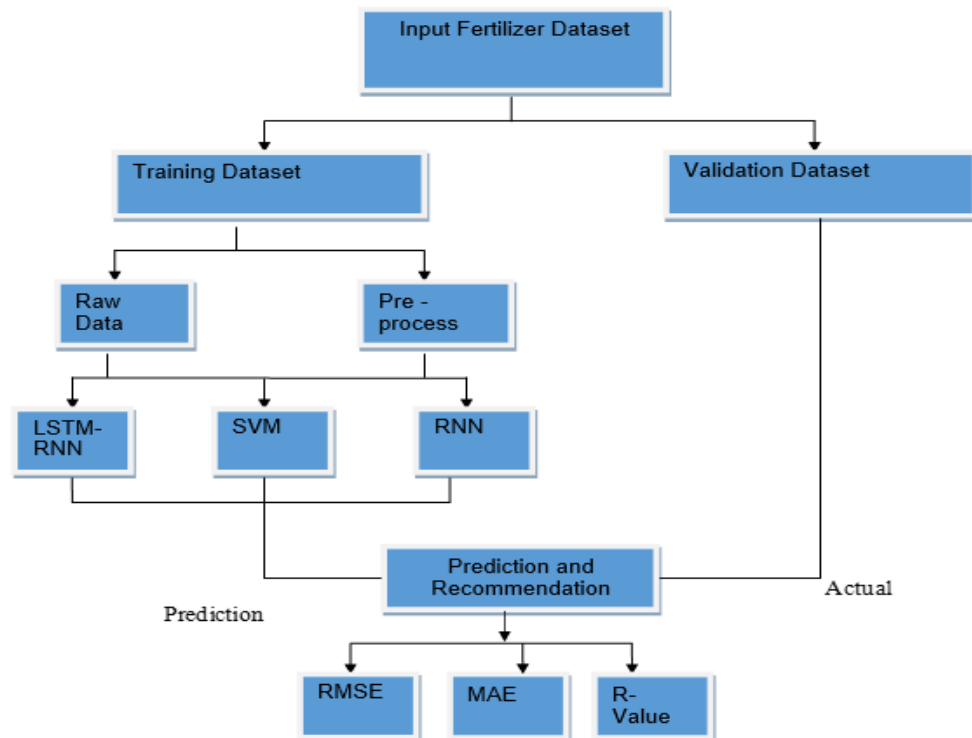


Fig. 34. Proposed Methodology

3.1 Material and Methods

1. Recurrent Neural network (RNN)

In RNN outputs from preceding step are taken care of which act as an input to present Step. In the neural network, every data is not depend on each other, yet in cases like when it is needed to foresee the following word from the sentence, the former words are needed. Therefore, here is a memorable requirement of the previous word. In this manner RNN arise into existence that give the solution to this problem with the assist of the Hidden Layer. The foremost feature of this network is the hidden state, after that it recalls few data about the present sequence.

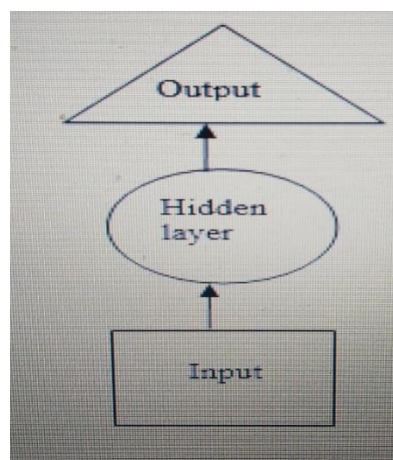


Fig. 35. Recurrent Neural Network

RNN have a “memory” which recollects all data about what has been calculated. It utilizes the indistinguishable for each input as it performs the same task on all the inputs or hidden layers to produce the

output. RNN have a problem of long interdependencies. It works fine when we are managing with short-term dependencies. This problem is overcome by LSTM i.e. Long short-term memory.

2. LSTM (Long short-term memory)

LSTM can be utilized for taking care of issue looked by the RNN. The RNN network feed the outcome of the block as an input which will be used for the next cycle or iteration utilize its feedback connection with store recent events inside the form of activation unlike the multilayer perceptron. LSTM utilized different processing elements and blocks to construct the improved recurrent network competent of learning continuing dependencies. The inputs to the LSTM cell are of past hidden as well as memory states along with the current information while the results are current hidden alongside memory state. In LSTM there are 3 gates.

- i. Input gates
- ii. Forget gates
- iii. Output gates

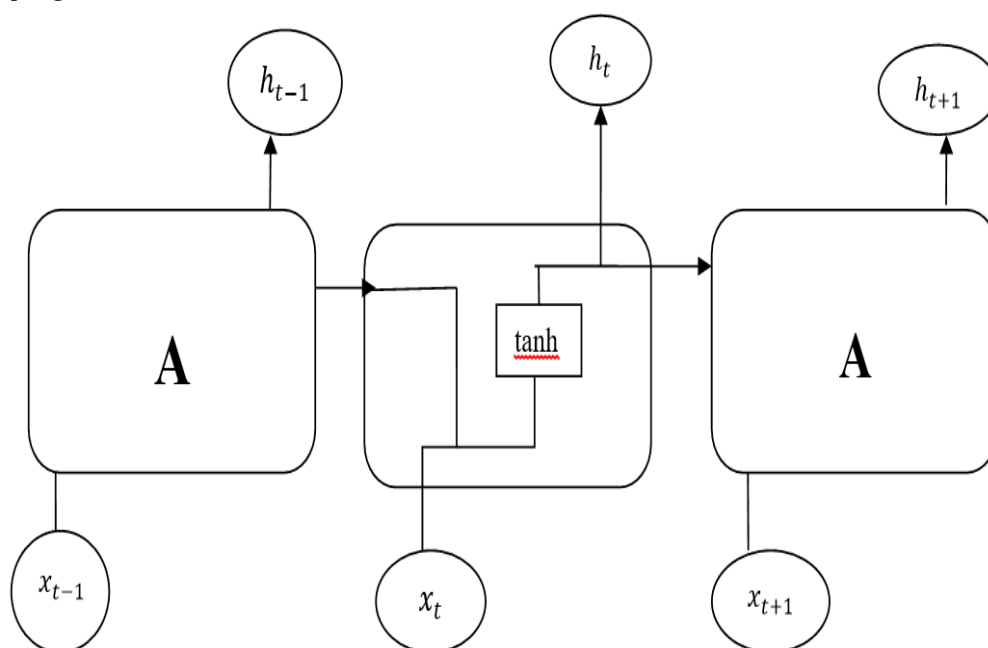


Fig. 36. LSTM Network

Equation of gates.

i_t stands for input gate.

f_t stands for forget gate.

o_t stands for output gate

σ stands for sigmoid function.

w_i stands for weight of the respective gate neurons.

h_{t-1} stands for output of the preceding lstm block.

x_t stands for input at the particular current timestamp.

$\tanh$ stands for the respective gate(x).

The memory present inside the LSTM act as a gated cell in which 3 gates control the ability to incorporate in other words expel data or let it sway the outcome at the latest current time step. It utilizes a sigmoid neural net layer and load to assign significance to data. It empowers the 3 gates to guarantee & it maintain the flow of all in- formation overcoming the two significant issues which are long term dependency problem and also gradient problem of Vanishing and present in conventional RNN. The process of LSTM can be utilized for predictions of time series which are consequently utilized inside the agriculture [17]

3. Support Vector Machine

SVM is one conventional machine learning algorithm utilized for classification and regression problem. SVM is a sort of supervised learning strategy that dependent on the possibility of hyper planes for choosing choice limits. Considering two dimensional spaces the hyper plane is a line. In light of the preparation tests, SVM characterizes them into one of the diverse participation classes.

While building the model, SVM training algorithm assigns new examples to one class or other class and assigns a hyper plane to output. This makes it a non-probabilistic binary linear classifier. An SVM model is a portrayal of the model as focuses in space, mapped such that the points in various classification are a greatest distance from the decision limit or boundary [18].

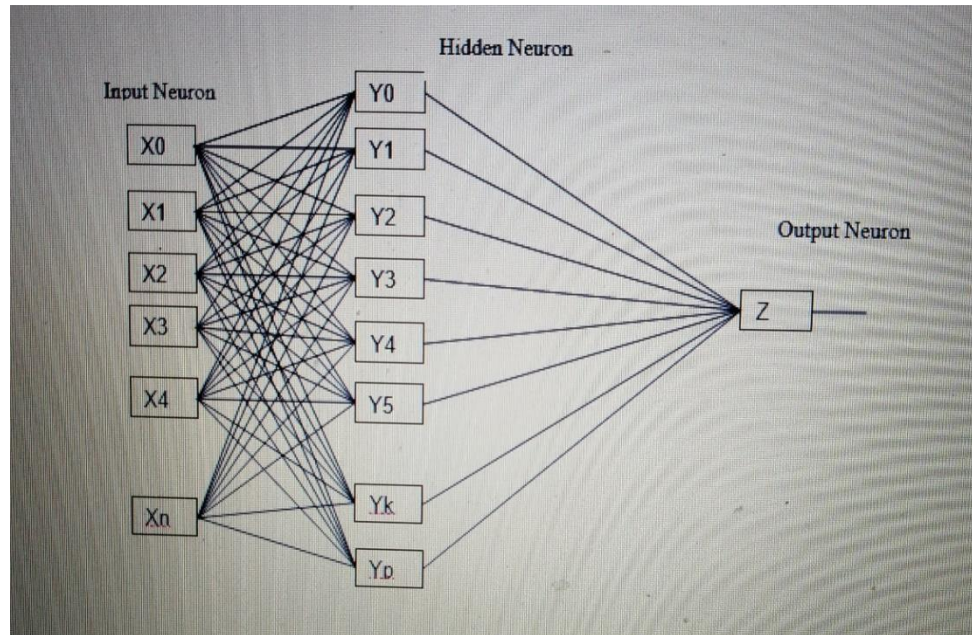


Fig. 37. Architecture of Support Vector Machine

4. Experimentation Result

The proposed work is implemented in various steps as follows:

Step 1: Result of the Main Proposed system and the accuracy achieved in the implementation.

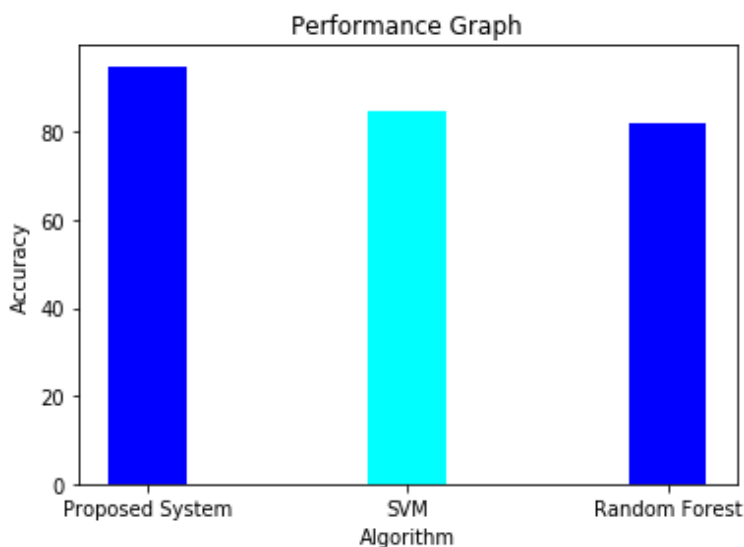
Accuracy					
95.0					
Confusion Matrix					
[[3 0 0 0 0 0 0]					
[0 4 0 0 0 0 0]					
[1 0 0 0 0 0 0]					
[0 0 0 1 0 0 0]					
[0 0 0 0 5 0 0]					
[0 0 0 0 0 2 0]					
[0 0 0 0 0 0 4]]					
	precision	recall	f1-score	support	
0	0.75	1.00	0.86	3	
1	1.00	1.00	1.00	4	
2	0.00	0.00	0.00	1	
3	1.00	1.00	1.00	1	
4	1.00	1.00	1.00	5	
5	1.00	1.00	1.00	2	
6	1.00	1.00	1.00	4	
accuracy			0.95	20	
macro avg	0.82	0.86	0.84	20	
weighted avg	0.91	0.95	0.93	20	
	Temperature	Humidity	Moisture	Nitrogen	Potassium
Phosphorous					
Temperature	1.000000	0.973164	0.091222	-0.033771	-0.023424
0.207545					
Humidity	0.973164	1.000000	0.091342	-0.060646	-0.003833
0.204044					
Moisture	0.091222	0.091342	1.000000	-0.095945	0.027727
0.009276					
Nitrogen	-0.033771	-0.060646	-0.095945	1.000000	-0.500087
-0.686971					
Potassium	-0.023424	-0.003833	0.027727	-0.500087	1.000000
0.089192					

Step 2: Simulation Final Result

Measurements of soil components such as Nitrogen, potassium and phosphorus are necessary constituents needed for recommendation and prediction of Fertilizers. Depending on the quantity of nitrogen the color of the leaves varies for example if the nitrogen quantity is very high or moderate in nature the leaves appear to be greener or if the quantity of nitrogen is acute the leaves appear to be slight yellowish. Reproductive system of the plant is dependent on the phosphorus content. The quantity of phosphorus helps in prediction of the growth of flowers and roots of the plants. Overall growth of the plant is affected by the potassium content. The strength of the plant root is predicted by its value.

Step 3: Performance Evaluation

Achieved algorithms are evaluated for accuracy and performance measurement based on the dataset. Standard exactness measure is utilizing for calculating accuracy. Accuracy is a small portion of the recovered data that is significant. It is set for assessment of accuracy. In the proposed calculation the exactness as well as accuracy of fertilizer and predicted fertilizer is looked at against real values given by experts for those arrangement of crops and fields. If the fertilizer recommended by the system belongs to the specialist recommendations sets. Considering these significant fertilizer accuracy of the proposed system is calculated.



5. Conclusion

As recommendation and prediction of fertilizers is most important for Indian farmers in Agriculture decision making, in this research concludes that the use of two algorithm to provide recommendation and prediction (machine learning and deep learning) Recommendation and suggestion of fertilizer is based on the soil parameters. Proposed system provides fertilizer recommendation on the basis of NPK (Nitrogen, Phosphorus, and Potassium) content of soil. Thus aim of this Research paper is to enhance the production of crops by suggesting correct fertilizer. The performance evaluation shows that the accuracy of achieved system is 95%.

References

- [1] Melissa A. Stine, Ray R.Weil, “The Relationship between soil quality and crop productivity across three tillage systems in south central Honduras”, American Journal of Alternative Agriculture, Vol.17, No. 1, 2002.
- [2] David H. Hardy, M. Ray Tucker, J. Kent Messick, Catherine Stokes, “Understanding the Soil Test Report”, published by N.C. Department of Agriculture & Consumer Services Agronomic Division, April 2012.
- [3] Sinkevičienė, D. Jodaugienė, R. Pupalienė and M. Urbonienė, “The influence of organic mulches on soil properties and crop yield”, Agronomy Research 7(Special issue I), 485–491, 2009.
- [4] Usman Ahmed, Jerry Chun-Wei Lin, Gautam Srivastava, Youcef Djenouri Computers and Electronics in Agriculture 189 (2021) 106407
- [5] Jenskie Jerlin I. Haban, John Carlo V. Puno, Argel A. Bandala , Robert Kerwin Billones , Elmer P. Dadios , Edwin Sybingco 2020 IEEE REGION 10 CONFERENCE (TENCON)
- [6] Devdatta A. Bondre, Mr. Santosh Mahagaonkar(2019), Prediction of Crop yield and Fertilizer Recommendation Using machine learning; Volume 4, issue 5,IJEAST
- [7] S.Bhanumathi, M.Vineeth and N.Rohit (2019). Crop Yield Prediction and Efficient use of Fertilizers, (pg.0769-0773), ICCSP.
- [8] Sirsta MS, Cernades E, Barro S. Automatic prediction of village wise soil fertility for many nutrients in India using wide range of regression method. Comput Electron Agric 2018, 154,120-33
- [9] Anna Chlingaryana, Salah Sukkarieha, Brett Whelanb — Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review , Computers and Electronics in Agriculture 151 (2018) 61– 69,Elsevier,2018.
- [10] Sirsat MS, Cernades E, Fernandez-Delgado M.Khan R Classification of agricultural soil parameters in India. Comput Electron Agric 2017, 135:269-79
- [11] E. Manjula, and S. Djodiltachoumy (2017). Data Mining Technique to Analyze Soil Nutrients based on Hybrid Classification. IJARCS.
- [12] Rohit Kumar Rajak, Ankit Pawar, Mitalee Pendke, Pooja Shinde, Suresh Rathod, and Avinash Devare (2017). Crop Recommendation System to maximize Crop yield using Machine Learning. Issue 12 IRJET.

- [13] de Paul Obade V, Lal R. Towards a standard technique for soil quality asses s- ment. *Geoderma* 2016;265:96–102.
- [14] Niketa Gandhi, Leisa J. Armstrong, Owaiz Petkar, and Amiya Kumar Tripathy, “Rice Crop Yield Prediction in India using Support Vector Machines”. In *International Joint Conference on Computer Science and Software Engineering (JCSSE)*. IEEE. 2016.
- [15] S. S. Baskar, L. Arockiam, S.Charles “Applying Data Mining Techniques on Soil Fertility Prediction”, *IJCAT*, 2013.
- [16] Mansi Shinde¹, Kimaya Ekbote, Sonali Ghorpade, Sanket Pawar, Shubhada Mone; Crop Recommendation and Fertilizer Purchase System; *International Journal of Computer Science and Information Technologies*, Vol. 7 (2) , 2016; pp. 665-667.
- [17] A.Moawad. (2018, Feb.) The magic of lstm neu-Ral networks. [Online]. Available:<https://medium.com/datathings/the-magicof-lstm-neural-networks-6775e8b540cd>.
- [18] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, “Support vector machines,” *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, July.
- [19] Suchithra MS, Improving the performance of sigmoidal kernels in multiclass SVM using OP for agricultural fertilizers recommendation system. *International conference on soft computing system, singapore*, Springer 2018 p.857.
- [20] Limin Chuana, Ping Hea; Establishing a scientific basis for fertilizer recommendations for wheat in China Yield response and agronomic efficiency; *Field Crops Research*; Volume 140; January 2013; pp. 1–8.

TCP/IP Benefits over ATM Architecture with Special Reference to 5G Network

*Deepthy Dubey¹, Thipendra Pal Singh², Arup Bhattacharjee³

^{1,2}School of Computer Science, University of Petroleum and Energy Studies, Dehradun

³Department of Computer Science and Engineering, National Institute of Technology, Silchar

¹deepthydubey@gmail.com, ²thipendra@gmail.com, ³arup@cse.nits.ac.in

Abstract. The rapid advances in ATM technology will trigger a lot of new opportunities for value-added services. An industry standard for linking computers and networks is known as Transmission Control Protocol/Internet Protocol (TCP/IP). As defined by the 3GPP, 5G is wireless technology. There are many Network Slices in each NSI, which were developed by the SA2 working group for 3GPP Service and System Aspects (SA2). 5G promotes the idea of "Network Slicing" as a way to decouple network resources and processes. Mobile services, direct broadcasting, private networks, and hybrid high-speed networks are all growing in popularity. Fibre-optic satellite networks would be used to construct these networks. New network architectures must be built in a 5G environment that interact seamlessly with existing standards, interfaces, and higher-layer protocols in order to properly implement these integrated systems using TCP/IP in its entirety. The main goal of this study is to compare and contrast TCP/IP with Asynchronous Transfer Mode and see which one offers the most advantages. New network architectures must be built in a 5G environment that interact seamlessly with existing standards, interfaces, and higher-layer protocols to fully implement these integrated systems utilizing TCP/IP.

Keywords: ATM, TCP/IP, 5G, IoT, VPC & VCC

1. Introduction

Networking protocols such as TCP and IP are the most widely used, and both are classed as Transport Layer protocols in the Open Systems Interconnection (OSI) paradigm. It is a model over networks for computer communication. Additionally, the TCP protocol establishes a secure connection amongst two hosts that are connected via an insecure network. IP, on the other hand, provides a service to the communication program at the other end of the communication chain. In the 1960s, the military, started working on implementation with the goal of connecting different locations that used different protocols and sharing information using a variety of different storage systems. This protocol suite was developed for sociological and technological purposes. Following its declaration as a reliable connection-oriented, we will examine some of the TCP protocol's basic features and how they function, as well as the TCP protocol's layer format. It will establish connections, terminate connections, and guarantee that data is sent correctly across the network. Many other layers and protocols in the networking will benefit from the services provided by this layer. [1].

Known collectively as the TCP/IP protocol suite, it is a bunch of many communication protocols that function over the Internet at different layers of the OSI model or the TCP/IP network layer model, depending on the protocol. Additionally, this protocol suite is adaptable to various private communication networks than the one described here. When computer networks first appeared, they presented us with a slew of advantages that were previously difficult or even impossible to get via conventional networking methods. However, this increased data transmission over the internet & use of networking has also resulted in numerous security breaches. The attacks that could occur as a result of this are discussed in the concluding section of this paper. Additionally, we provide some feasible solutions to the various kinds of attacks discussed here [2].

TCP/IP is a two-tiered communication protocol. TCP is the top layer that regulates this process. The Internet Protocol (IP) layer is the most fundamental layer and is in charge of processing the address part of each packet and ensuring that it reaches its intended destination. When a message is received, computer in the network checks the address from each gateway to decide how it should be delivered. However, even if certain packets in a message are distracted & routed differently than others, they will be reassembled at their final destination [3]. In computer network communication, the client/server paradigm governs how a computer user requests and receives services from another computer connected to the network.

As a general rule, communication flows in Transmission Control Protocol/Internet Protocol is point-to-point, which implies that each communication takes place between two points on a network or between two host computers. Due to the fact that each client request is handled independently of prior requests, TCP/IP and the higher-level applications that use it are generally referred to as "stateless." Being stateless frees up network routes, enabling them to be utilized continuously by anyone and everyone on the network. Any Internet user is familiar with the extra layer application protocols that connect to the Internet via TCP/IP, which are known as additional layer protocols.

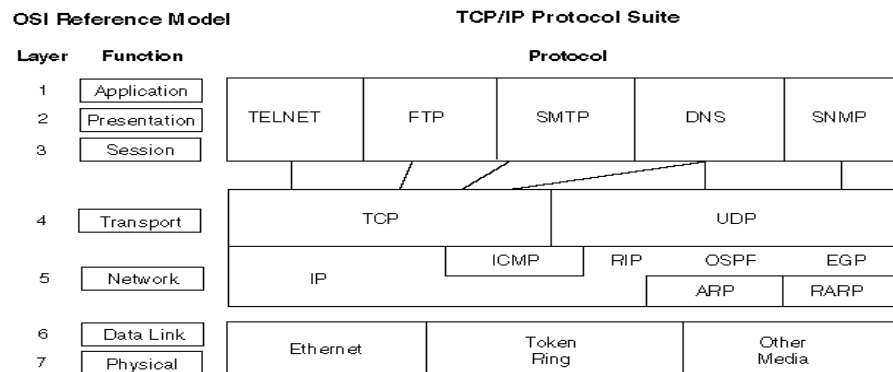


Fig.1. Protocol Suite of TCP/IP

The above fig 1 shows the functionality of layers between two different models OSI & TCP/IP.

TCP/IP, as well as these and other protocols, are often packaged together to create a "suite." The Serial Line Internet & the Point-to-Point Protocol (SLIP) or (PPP) Protocol are two protocols that are often used by personal computer users that connect to the Internet using an analog phone modem (PPP). In order to transmit IP packets via a dialup phone connection to the modem of an access provider [4]-[5], these techniques encapsulate IP packets.

TCP/IP protocol suite is an open standard, not a secret proprietary protocol owned by a single company. This enables anybody with adequate specialized knowledge to improve it. It is compatible with almost all sophisticated working frameworks, allowing any framework to communicate with another. It is also compatible with a broad variety of PC equipment and system configurations. It ensures the reliable transmission of information. For data transfer between two or more devices, a wide variety of high-speed networking technologies are available to choose from. ATM is one of the technologies that may be used. With its great speed, low latency, and flexibility to accommodate different types of traffic on a single network, ATM is well suited to handling large amounts of data in a number of bandwidth-intensive applications such as multimedia and medical imaging. It is one of the key differences between ATM and other networking systems that the concept of "virtual circuits" is used. When it comes to ATM circuit provisioning, the ability to negotiate a virtual circuit with defined bandwidth and quality of service criteria is included. Asynchronous transfer mode (ATM) networks benefit from virtual private networks (VPs) since they substantially speed up the connection setup phase at the expense of bandwidth loss caused by the usage of end-to-end network resource reservation. Accordingly, virtualization platforms (VPs) may be used to fine-tune the fundamental trade-off between network call performance and signaling system processing load. Aside from that, they may be used to provide customers with specific connection services, such as virtual networks (VNs). [6].

When it comes to allowing broadband multimedia communications, ATM has emerged as the most promising technology to date. When it comes to ATM networks operating efficiently and reliably, congestion control is essential. AT&T's Asynchronous Transfer Mode (ATM) protocol allows the transfer of a wide range of user traffic, including voice, data, and video signals. Its goal is to bridge the gap that now exists between telecommunications and computer networks. It uses asynchronous time division multiplexing to compress data into small, consistently sized cells, which are then sent across a network. [7]

Prior to the advent of ATM, data link layer communications were reliant on frame switching and frame networks to achieve their goals. Frames are used by a variety of protocols and may vary in both complexity and size. As networks get more obscure, the quantity of information needed in the header grows in proportion. A result of this is that the headers get larger in relation to the size of the data unit. Because of this, many

protocols have raised the size of the data unit in order to improve header use (sending more data with the same size header). Regrettably, large data fields result in a lot of waste. As a result, when there is little information to transmit, a significant part of the field stays unused. Several protocols offer users with the ability to adjust frame sizes in order to optimize use.

In order to link geographically distant ATM networks, ATM technology is becoming more popular. When it comes to global connection and statistical multiplexing, terrestrial ATM systems have progressed to the point where they allow end-to-end connectivity and QoS criterion compliance, but satellite-based ATM systems will be essential [8]. ATM's goal is to provide transportation and switching services that are both efficient and cost-effective for a wide variety of traffic sources. Because of the many advantages offered by satellite communications technology, satellite ATM networking is becoming more widespread [9] – [14]. Satellites' intrinsic broadcast capacity enables multipoint-to-multipoint communications, including ATM island connectivity, as well as bandwidth on-demand or Demand Assignment Multiple Access (DAMA) capabilities. Another advantage is disaster recovery as an alternative to fibre optic networks.

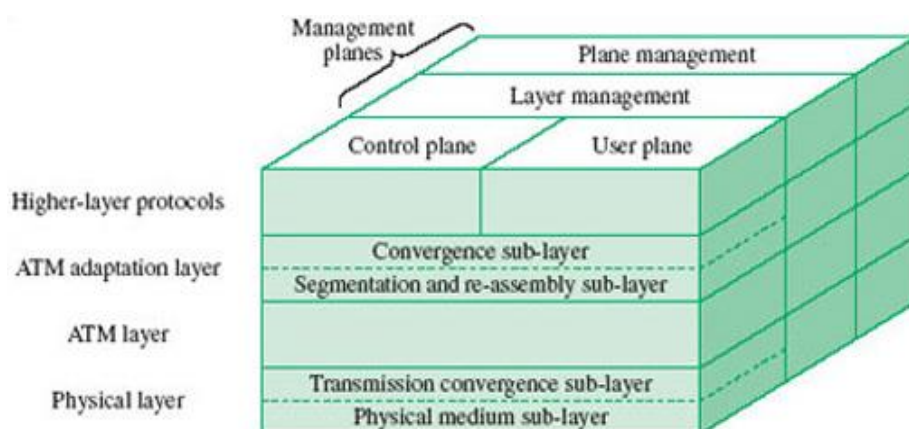


Fig.2. ATM Reference Model

The three layers of the reference model: the ATM adaptation layer (AAL), the ATM layer, and the physical layer as shown in above fig 2.

Network and end-system design may be used to improve performance in two ways. To optimize resource efficiency, fairness, and throughput, the network may use a number of layer-level optimization methods [15]. To name a few recent developments: feedback control for improving efficiency and fairness, resource-saving intelligent drop rules, per-VC buffer management, and even minimum throughput guarantees for higher layers. If the performance of the transport layer is to be improved while minimizing congestion collapse, several congestion avoidance and management methods may be used by the transport layer. Transport systems have seen a flurry of congestion management ideas and initiatives. When it comes to TCP, techniques like gradual start and congestion avoidance are used, along with quick retransmission and recovery, and selective acknowledgement [16]. In order to enhance data network quality, reliability, and "flexibility," the networking community is continuously developing and upgrading new technologies. A major business barrier stands in the way of internet service providers offering "flexibility" to their customers. As a result, DQoS (Differentiated Service Quality) will be used in this paper. Customers on a public network may establish customized connections thanks to the use of networking protocols, methods, or standards.

In 5G the "G" stands for "generation," while the number 5 represents the progress. Wireless phone technology began with 1G and was improved to 2G in the early 1990s when businesses allowed users to exchange text messages between two cellular handsets, a feat that captivated the globe. Finally, the globe transitioned to 3G, which enabled users to make phone calls, send text messages, and surf the internet at lightning speed. Many of the features that were previously only available with third-generation wireless have been improved with 4G. Individuals could surf the web at lightning speed, exchange text messages, conduct phone conversations, and even upload huge video files and download without encountering any difficulties or waiting periods. Then, businesses enhanced 4G connection with LTE, an acronym for "long term evolution". Both methods produced comparable results, but it was critical to establish a universal standard. LTE

accomplished this by speeding up 4G technology, laying the groundwork for 5G. People will be able to upload and download 3D and Ultra HD movies more easily with 5G. As a result, we may assert that the pace of life has increased. Consider updating your data connection from a garden hose to a flames hose. The distinction will be apparent and deserving of appreciation. [17]

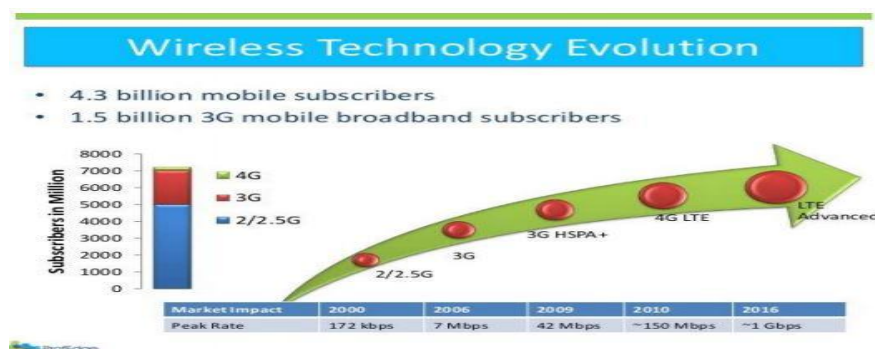


Fig.3. Wireless Technology Evolution

The evolution/growth of wireless technology around the globe is shown in fig 3 above.

5G is the first network to operate at telecommunications standards frequencies between 24 and 86 GHz. The development and significance of the 5G network are shown in Fig 1[18], which illustrates the relevance of cellular networks across the generations as technology advances at a quicker rate than prior generations

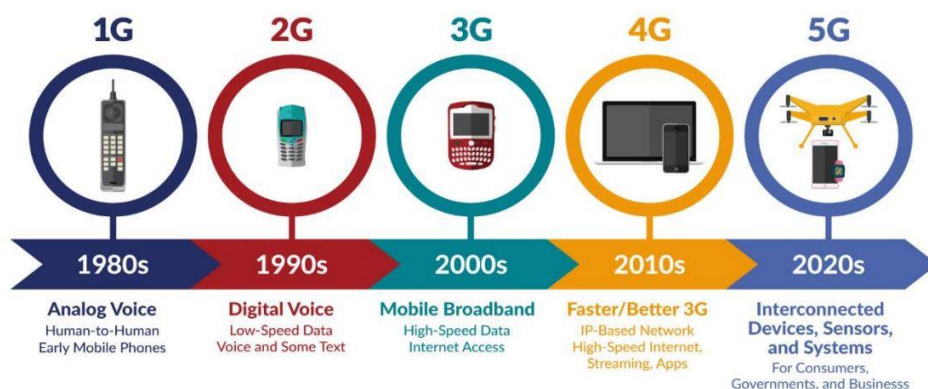


Fig.4. Imperative Growth of 5G Networks

The imperative Growth and importance of 5G network is shown in Fig 4 above.

As communication has become an integral part of our lives, and traffic is growing at a breakneck pace due to congestion, The 5G network significantly improves this kind of issue, making it more efficient, convenient and cost effective. [19]-[21]. If a new technology is created and updated, society's primary concern in every area involving data, information, and gadgets that enable safe digital marketplaces will be security. Security for 5G networks and associated communication services will be critical in the next years. Interoperability with terrestrial 5G networks is made possible by the suggested architecture for a 5G satellite network. This satellite-terrestrial integrated system is also looking into a new technique based on network code to allow the simultaneous usage of multiple routes. In comparison to the current terrestrial 5G network, the planned 5G satellite network is more interoperable. The simultaneous use of several routes in such a combined satellite-terrestrial system is also being explored using a unique method based on network coding. The optimum traffic split across terrestrial and satellite routes and the greatest degree of redundancy are determined using an analytical model for TCP-based applications.

2. Literature Review

An extensive literature overview on ATM networks may be found in [22]-[24]. In terms of routing, the ATM network is similar to a traditional packet switching network in that each message that arrives creates a virtual connection. As a result, ATM networks are more resource-efficient than traditional channel switching networks. Networks using ATM, on the other hand, vary fundamentally from those utilizing traditional packet switching. Remember that ATM networks have certain features that determine the specifics of the routing algorithms that are employed.

The substantial growth of the modern Internet, which is based on the Internet Protocol version 4 (IPv4), has resulted in an avalanche of problems relating to global network administration and operation. A plenty of research projects have been organized by many research groups with the goal of improving the current version of the Internet Protocol. The findings of the research [25] revealed the flaws in the current version of the Internet protocol. According to the findings of the research [26], the growth of the Internet is hindered by a lack of available public IP addresses. In order to connect to the Internet, the vast majority of mobile devices now need an IP address, resulting in a substantial growth in the number of IP addresses in use. An investigation by the Internet Engineering Task Force led to the recommendation of an updated version of the Internet Protocol, known as Internet Protocol version 6. (IPv6). Next is extensive material on the two versions of the Internet Protocol, IPv4 and IPv6. IPv6 networks are susceptible to some of the same issues that affect IPv4 networks [27]. Additionally, when the hacker community starts to actively target IPv6 networks, it is anticipated that a large number of previously undisclosed security flaws would be discovered and exploited.

During the course of the essay, In addition to these limitations, all algorithms were designed for specific purposes and had key or block sizes of 64, 128 or 256 bits, were only capable of processing text data and not voice, and had performance issues [28] - [30].

Several obstacles lie in the way of the development of 5G networks. Cellular communications are hindered by a shortage of radio frequency (RF) spectral bands, which is one of the most important barriers to their widespread adoption. In addition, the operation of advanced wireless technology is linked with a substantial increase in energy consumption, as previously stated. For environmental reasons, cellular operators have noticed and reported that the energy used by base stations accounts for more than 70% of their total electricity costs. When one looks in the market for 5G network the present status is clear that the network's different access methods are nearing a stalemate and need urgent upgrading to function properly. It is estimated that current technology, such as OFDMA, will be feasible for at least the next 50 years. Moreover, there is no need for technical advancement. In recent years, the wireless setup has progressed from 1G to 4G. Package providers are being pushed to transition to a 5G network when 4G is generally accessible because to the introduction of an application, or more specifically, the upgrading of the basic network to satisfy customer demands.

A worldwide 5G 3GP standard is already being implemented by the Chinese Communications Standards Association (C CSA). CCSA was established in 2012. It was established in September 2017 to try to make clear the country's 2020 plan for embracing newer technologies.

3. Methodology

Massive quantities of satellite data may now be sent using a new 5G architecture that is ground breaking in its application. As you can see in Figure 1, the overall structure is laid out. Satellite communication terminals include those that are directly or indirectly linked to IP or ATM or another network protocol. This design comprises terrestrial networks as well as ISL satellites and earth stations that are permanently located on the planet's surface.

Various TCP/IP data transfer methods for a satellite ATM network are enabled by existing and future ATM specifications. When it came to cost and performance, TCP/IP beat out ATM. This was true for all three service types – UBR, ABR and GFR – as well as the many implementation options associated with each. Regulating virtual path bandwidth follows the following strategy: An amount of the request bandwidth will be raised if the present bandwidth is not enough to connect the incoming call to the end node. Expand and create a virtual call circuit if there is extra bandwidth available; else, keep the current capacity and reject the

connection. Reduce virtual route utilization's bandwidth as much as feasible by a given amount. There is a TCP/IP network flowing via an ATM satellite, as shown in Figure 1, while Figure 2 shows the implementation's general topology. Depending on the method, the steps needed to alter the bandwidth or control range. The method's network performance is heavily influenced by the settings that have been specified. But only if it increases transmission efficiency without decreasing the advantages of VRS's decreased processing load will this approach be useful for wireless networks. This level is determined by looking at transmission efficiency and processing load.

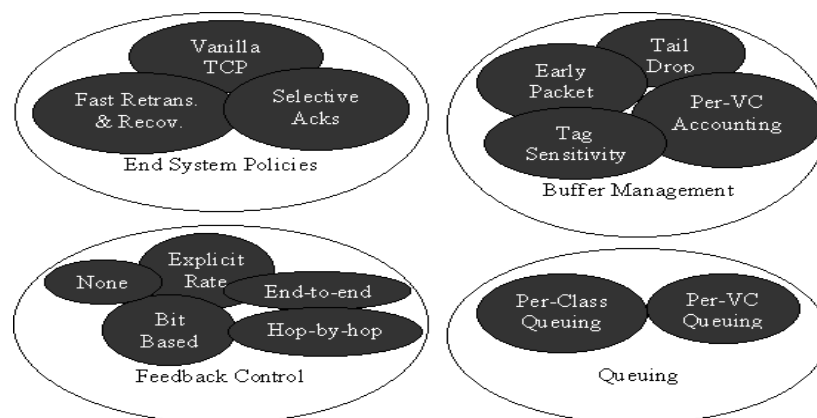


Fig.5. TCP/IP Data over a satellite ATM Network

The above fig 5 illustrates about the data of TCP/IP over a satellite ATM network.

Massive consortiums of big global telecommunications companies are already collaborating to generate global value around 5G. Although the majority of those standards have not been finalized, experts anticipate it to be more compatible (with 4G and 3G) and to have some global interoperability. Consider the scenario in which the base station communicates with the mobile stations in order to grasp the concept behind BDMA technologies. Using the BDMA technique, we can divide each mobile station's orthogonal beam according to its location, allowing for openhanded multiple access to the mobile stations while also increasing the system's capability. This is the primary process of communication in this system, and it is the primary process of the communication in this system.

The idea of moving to 5G is based on current trends; it is generally thought that 5G cellular networks will be able to overcome six obstacles that 4G networks will not be able to overcome, including, but not limited to, congestion:

- 1: Consistent Quality,
- 2: Connectivity to massive device,
- 3: Data rate higher,
- 4: Higher capacity,
- 5: Reduced cost,
- 6: End to End latency has been lowered.

To address environmental concerns, cellular operators have observed and reported that the energy used by base stations accounts for more than 70% of their power cost. When one looks at the present status of the 5G network in the market, it is clear that the network's different access methods are nearing a stalemate and need

urgent upgrading to function properly. It is estimated that current technology, such as OFDMA, will be feasible for at least the next 50 years. Furthermore, there is no need for technical advancement. The wireless setup has progressed from 1G to 4G in recent years. Package providers are being pushed to transition to a 5G network after 4G has been generally accessible due to the introduction of an application or, more specifically, the upgrading of the basic network to suit the demands of customers.

5G promotes the concept of Network Slicing to abstract network resources and network functions. It enables carriers to build a single physical network that can account for very different use cases: high bandwidth applications (e.g., streaming), low bandwidth applications (e.g., connecting Internet of Things (IoT) devices) with low latency requirements, enterprise extranets, etc. Specific verticals using private 5G networks or network slices of public 5G networks are just a special case of proprietary networks. Within private, proprietary networks, people are free to leverage specific non-IP based technologies without any impact on the global Internet.

In recent years, communication network engineering research has focused on service integration. The Asynchronous Transfer Mode has gotten a lot of attention for its effectiveness in creating a service grid that is linked. ATM works like a multiplex packet switch in that it splits the data into smaller, fixed cells that are identified by stream IDs and delivered across the network. There are two kinds of switching: circuit switching and packet switching. When it comes to switching, there are many different types to choose from. A circuit-switched network creates a physical communication channel only for the purpose of transporting a phone call across the network. Packet-switched networks, on the other hand, share their physical communication channels across many users. Switching may be categorized into two heads: circuit switching and packet switching. The more common kind of switching is Circuit switching. In a circuit-switched network, a physical communication channel is established only for the purpose of carrying a call across the network. The physical communication channels of a packet-switched network, on the other hand, are shared among many users.

Label switching is a technique of providing IP via an ATM network, and it is one of many. ATM network edge routers convert between forwarding equivalence classes and labels in order to provide better service to customers (i.e., important portions of IP network addresses). In ATM header fields known as virtual route identification and/or virtual channel identifier, the labels are transmitted together with the data

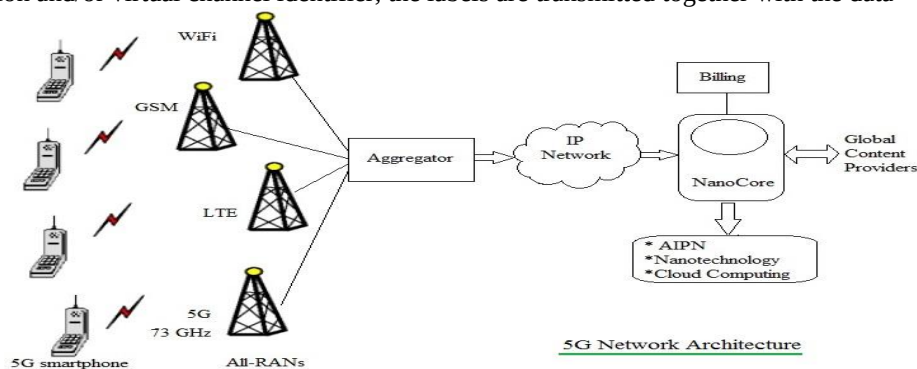


Fig.6. TCP/IP Data over a satellite ATM Network

The above fig 6 illustrate the architecture of 5G network.

The 5GC (5G Core) Network has been designed to be highly flexible, modular, and scalable in its architecture. The company offers a wide range of services, such as network slicing, to suit the varied requirements of its clients. Cloud computing on a distributed scale, network function virtualization, and Software-Defined Networking (SDN) are all provided.

Security Challenges of 5G & Evolving Architectures: Several problems connected with 5G technology and its security may be taken into consideration throughout the development process; for example, 1) Identity Management with a Diverse User Base, and 2) The provision of security as a service 3) Implement end-to-end encryption. 4) Create virtual network slices that are isolated from one another. 5) The protection of security and privacy, 6) Concerns about security as the number of IoT devices continues to increase.

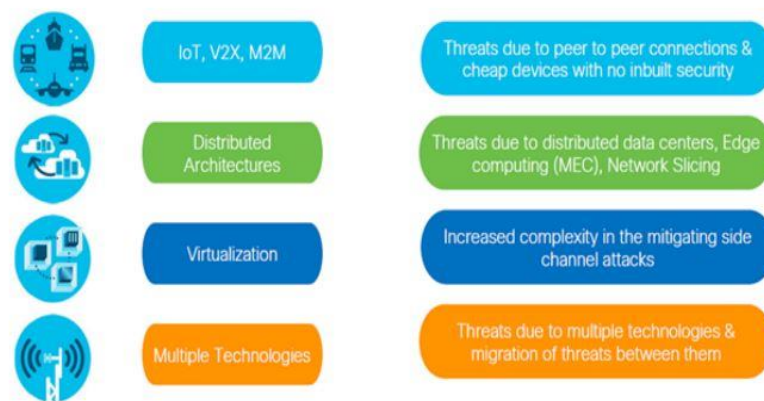


Fig.7. Security Challenges In 5G

Depending on the method, the steps needed to alter the bandwidth or control range. The method's network performance is heavily influenced by the settings that have been specified. Only through improving transmission efficiency without compromising the advantages of the VRS's decreased processing load can the algorithm stages be successful. This level is determined by looking at transmission efficiency and processing load. A large number of assumptions are made in order to make the analysis more straightforward. In other words, bandwidth is shared between all calls. In assigning bandwidth for a call or a virtual connection, no statistical multiplexing impact is taken into consideration; instead, the bandwidth is allocated deterministically.

However, in order to build an efficient 5G optical transport network, three criteria must be met: dense mobile access, high-capacity transit with low latency, and scalable networking

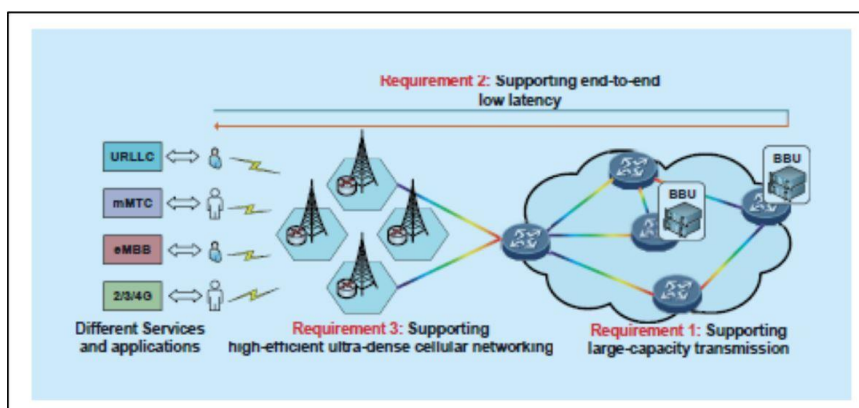


Fig.8. 5G Network Architecture Setup

In order to build an efficient 5G optical transport network, three criteria must be fulfilled, as shown in Figure 8: dense mobile access, high-capacity transit with low latency, and scalable networking

4. Results and Discussion

It is possible that the most complex networking technology is ATM technology is that we have ever met in our lives. The security of such a complex system will be more difficult to implement than the design of the system. And now, individuals are starting to address some of the issues related to ATM security that have arisen. It will take some time before we are able to achieve our complete security objectives. ATM security must be flexible and compatible with a variety of technologies in order to achieve its goal of providing a unified networking platform and communication infrastructure. This adds even another degree of difficulty to the security of ATMs and other financial institutions. Several draft security standards and framework papers have been produced by the ATM Forum's Security Working Group. Numerous other security-related issues have been discussed in the ATM Forum, in addition to those already mentioned. In addition, ATMs are capable of simultaneously providing a variety of other services over the same network. All of the

information is contained inside equally sized cells. The cell header contains information on how cells are routed via VCIs and VPIs, as well as information about the cells themselves. The interleaving of cells from various applications going to the same destination may allow them to share physical infrastructure and save costs. The ability to offer a range of services over the same physical infrastructure allows network operators to expand their business. This is beneficial to network providers because it allows for the most efficient use of available resources. The quantity of mobile and wireless traffic is expected to increase 1,000-fold over the next decade, with the majority of it being used by the predicted more than 50 billion connected devices connected to the cloud by 2020, according to estimates. To address the various challenges associated with the rapid growth of connected devices, measures have been taken to improve energy efficiency, increase capacity, reduce costs, and maximize spectrum utilization, as well as increase stability and scalability in order to handle the rapidly growing number of connected devices. Today's world is changing at the speed of light, and we have become increasingly dependent on technology that allows us to communicate more rapidly. The primary technical aim is to provide us with more information. a system idea that supports:

- Increase the volume of data per region by a factor of 1000
- The number of linked devices should be multiplied by ten to one hundred.
- enhanced average user data rate by 10 to 100 times

One part of the ATM 5G satellite network is made up of interconnected satellites, while the other is made up of different ATM networks on the ground. Various call and control processes are managed by a module that connects the ATM and satellite networks together. A network control centre manages and distributes resources throughout the network. TCP/IP over ATM architecture is used to evaluate a business network's data transfer efficiency since the performance evaluation is scalable and customizable. These higher-level standards, interfaces, and protocols have been taken into consideration while designing our 5G network topologies. It is possible to quantify the effectiveness of congestion management methods by utilizing performance metrics such as optimal resource usage, fairness, and increased layer throughput. A comparison of service assurance D-QoS standards across a company network shows that quality may be improved.

5. Conclusion

Service assurances are needed in intradomain networks, and D-QoS standards for data networks have traditionally proven very effective, especially when the target market is restricted. For small companies, it was an excellent predictor of future performance, but commercial attempts to reach a larger market were often unsuccessful. To counteract these improvements, IT management must re-examine whether or not expensive D-QoS services are still necessary on the corporate network. 5G Network Slicing is trying to launch in a market lacking premium D-QoS services, which is why this is happening. While efforts to finalize Network Slicing technical standards and operational details are ongoing, previous standards' experience suggests that greater clarity in the business model and the level of service security provided to users is critical for the standards to be as effective as the previous standards have shown. Analysing integration possibilities and evaluating the integrated system's virtualized features will need further study. The effectiveness of different network coding methods must also be evaluated.

References

1. Base of the Networking Protocol – TCP/IP Its Design and Security Aspects International Journal of Innovative Research in Computer and Communication Engineering (An ISO 3297: 2007 Certified Organization) Vol. 3, Issue 4, April 2015.
2. Abdullah H. Alqahtani, MohsinIftikhar, "TCP/IP Attacks, Defenses and Security Tools", International Journal of Science and Modern Engineering (IJISME) ISSN: 2319-6386, Volume-1, Issue-10, September 2013.
3. TCP-IP Model in Data Communication and Networking American Journal of Engineering Research (AJER) e-ISSN: 2320-0847 p-ISSN : 2320-0936 Volume-4, Issue-10, pp-102-107
4. S. Zaman S., F. Karray. Fuzzy ESVDF approach for Intrusion Detection System. The IEEE 23rd International Conference on Advanced Information Networking and Applications (AINA-09). May 26-29, 2009. [2]
5. I. Onut and A. Ghorbani. A Feature Classification Scheme for Network Intrusion Detection. International Journal of Network Security, Page(s): 1-15, July 2007

6. JAMES R. WEBSTER. Asynchronous Transfer Mode (ATM) Technology: An Overview, RL/NR/8143--93-7337, June 4, 1993.
7. An Overview of Network Architecture and security framework of Asynchronous Transfer Mode International Journal of Scientific and Research Publications, Volume 3, Issue 4, April 2013 1 ISSN 2250-3153
8. Framework of ATM Architecture Deliberation and TCP/IP Performance for 5G Network: An Overview International Journal of Future Generation Communication and Networking Vol. 14, No. 1, (2021), pp. 2786–2794
9. Ian F. Akyildiz, Seong-Ho Jeong, "Satellite ATM Networks: A Survey," IEEE Communications Magazine, July 1997, Vol 5.35. No. 7.
10. ITU-T Recommendation I-356, "B-ISDN ATM Layer Cell Transfer Performance," Geneva, 1993.
11. ITU-R 4B Preliminary Draft Recommendation, "Transmission of Asynchronous Transfer Mode (ATM) Traffic via Satellite," Geneva, January 1997.
12. M. Mathis, J. Madhavi, S. Floyd, A. Romanov, "TCP Selective Acknowledgement Options," Internet RFC 2018, October 1996.
13. Dr. M. Anand Kumar et al, Int. Journal of Computer Science and Mobile Computing, Vol.2 Issue. 11, November-2013, pg. 331-338
14. Rich Gobbi, Eduardo Elizondo, Anthony Modelino, Frank Gargione, "Evolution of the Astrolink™ System," Proceedings of the 3rd KaBand Utilization Conference, Sep 1997.
15. R. Goyal, R. Jain, S. Kalyanaraman, S. Fahmy and Seong-Cheol Kim, "UBR+: Improving Performance of TCP over ATM-UBR Service," Proc. ICC'97, June 1997.
16. V. Jacobson, "Congestion Avoidance and Control," Proceedings of the SIGCOMM'88 Symposium, pp. 314-32, August 1988.
17. A Review Paper on 5G Wireless Networks International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181 Published by, www.ijert.org V-IMPACT - 2016_2 Conference Proceedings
18. Prospects for Handling 5G Network Security: Challenges, Recommendations and Future Directions Journal of Physics: Conference Series 1714 (2021) 012052 IOP Publishing doi:10.1088/1742-6596/1714/1/012052
19. P. K. Agyapong, M. Iwamura, D. Staehle, W. Kiess, and A. Benjebbour, "Design considerations for a 5G network architecture," IEEE Communications Magazine, vol. 52, no. 11, pp. 65–75, 2014.
20. A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, M. Maternia, O. Queseth, M. Schellmann, H. Schotten, H. Taoka, H. Tullberg, M. A.
21. Uusitalo, B. Timus, and M. Fallgren, "Scenarios for the 5G mobile and wireless communications: the vision of the METIS project," IEEE Communications Magazine, vol. 52, no. 5, pp. 26–35, 2014.
22. ITU-R 4B Preliminary Draft Recommendation, "Transmission of Asynchronous Transfer Mode (ATM) Traffic via Satellite," Geneva, January 1997.
23. M. Mathis, J. Madhavi, S. Floyd, A. Romanov, "TCP Selective Acknowledgement Options," Internet RFC 2018, October 1996.
24. Raj Jain, "Congestion Control and Transaction Management in ATM Networks: Recent Advances and A Survey," Computer Networks and ISDN Systems, October 1996.
25. David, C. Lee., Daniel L. Lough, Scott F. Midkiff, Nathaniel J. Davis IV, and Phillip E. Benchoff, 1998. The Next Generation of the Internet: Aspects of the Internet Protocol Version 6, IEEE Network, 12(1): 28-33.
26. Saleh, A. M. , and J. M. Simmons, 2011. Technology and architecture to enable the explosive growth of the internet, IEEE Communications Magazine, 49(1): 126-132.
27. Caicedo, C. E. , J. B. Joshi, and D. Tuladhar, 2009. IPv6 Security Challenges, IEEE Computers, 42(2): 36-42.
28. Homin K. Lee, Tal Malkin, and Erich Nahum, 2007. Cryptographic strength of servers: current and recent practices, Proceedings of the ACM SIGCOMM conference on Internet measurement: 83-92.
29. Mingyan Wang, and Yanwen Que, 2009. The Design and Implementation of Passwords Management System Based on Blowfish Cryptographic Algorithm, Computer Science-Technology and Applications, 2(1): 24-28.
30. Ijaz Ali Shoukat, Kamalrulnizam Abu Bakar and Mohsin Iftikhar, 2011. A Survey about the Latest Trends and Research Issues of Cryptographic Elements, International Journal of Computer Science Issues, 8(3): 140-149

Design and Implementation of Augmented Reality for Smart Driver Assistance System

K. Baskar¹, *P. William², S. Yuvaraj³, K. Vengatesan⁴, Rahul Bibave⁵

¹Associate Professor, Department of Computer Science and Engineering,
Kongunadu College of Engineering and Technology, Tamil Nadu

²Research Scholar, Department of Computer Science and Engineering,
School of Engineering and Information Technology, MATS University, Raipur
³Assistant Professor, Department of Electronics and Communications Engineering,
SRM Institute of Science and Technology, Kattankulathur

⁴Professor, Department of Computer Science and Engineering,
Sanjivani College of Engineering, Kopargaon

⁵Assistant Professor, Department of Electrical Engineering,
Sanjivani College of Engineering, Kopargaon
william160891@gmail.com

Abstract. This research is to investigate momentum innovation for constantly detecting street signs from a moving vehicle. The most encouraging innovation for perceptive vehicle frameworks is vision sensors and image preparation, therefore this is examined the most thoroughly. Various handling calculations and study the world over concerned with sign acknowledgement are investigated. A functional framework has also been implemented using a regular web-camera installed in a testing car. This framework is limited to speed signs and achieves outstanding displays due to rapid but hearty computations. The division is based on shading data, and the recognition is based on a model coordinating computation. The human-computer interface is a voice that announces which sign has been discovered. Driver carelessness is a key element that commonly results in distortion of surroundings, for example, erroneous recognition or disregard of street and traffic signs. Although significant progress has been made in developing a vehicle capable of autonomous guided driving, development has been slow because to concerns of speed, safety, and the ever-changing complexity of the on-road situation. Street plays a key role for the motorist in gathering data from the sign and then acting in a similar fashion.

Keyword: Augmented Reality, Smart Driver, Road Accident, HCI, Street Sign

1 Introduction

With the improvement of the astute vehicle, safe driving help frameworks are turning out to be increasingly significant. In safe driving help framework, the traffic sign acknowledgment is a key innovation, and has been broadly utilized. The precision and short preparing time are significant for traffic sign acknowledgment. Be that as it may, in genuine driving circumstances, the different circumstances of traffic signs including the pivot, perspective, scale and enlightenment are unpredictable and bothersome. Accomplishing vigorous traffic sign acknowledgment with short handling occasions is a difficult undertaking. Traffic sign discovery incorporates traffic sign acknowledgment and traffic sign arrangement. So as to accomplish fast and hearty traffic sign identification, planning a processing effective and exceptionally discriminative element is fundamental. In the interim, so as to accomplish fast and hearty traffic sign arrangement, setting up a grouping procedure that can diminish the measure of highlights and continue order precision is likewise significant. Current deals with the traffic sign location and acknowledgment can be separated into three unique sorts. To begin with, pre-handling strategies are inquired about to find and perceive the traffic signs. Second, pre-handling strategies consolidating with groupings are received to accomplish fast and strong traffic signs acknowledgment. Third, explicit structure highlights brushing with the classifiers are utilized to accomplish the vigorous and registering proficient acknowledgment. This research is driven by the desire to devise a better system for driving where the smartness and efficacy of the driving may be improved to provide better outcomes. This new system that uses state-of-the-art technology such as Augmented Reality (AR) and Embedded Systems will allow drivers to get a better understanding of the findings, making it easier for them to act on it by providing relevant alerts.

1.1 Background

The Road Sign Recognition (RSR) is a field of applied PC vision look into worried about the programmed location and grouping of traffic signs in rush hour gridlock scene pictures procured from a moving vehicle. The aftereffect of the RSR look into exertion can be utilized as an emotionally supportive network for the driver. At the point when the neighbor condition is comprehended, PC backing can help the driver in pre-crash forecast and evasion. Driving is an assignment put together for the most part with respect to visual data and picture handling. The street signs and traffic signals characterize a visual language translated by drivers. Street signs gives numerous data essential to effective driving - they depict the present traffic circumstance, characterize option to proceed, disallow or license certain headings, caution about unsafe components and so on. Street signs additionally help drivers with route.



Fig. 1. Intelligent infrastructure and intelligent vehicles to warn about slippery road ahead.

1.2 Motivation

Out of the car, Smart Driver Assistance Systems are comprehensive frameworks that assist the driver in a variety of ways. These frameworks may be used to provide important data regarding agreements, conclusions, and clog clearance.

Frameworks may be used to determine whether the human driver has reached his or her limit, and if so, implement preliminary alerts or assess the driving. Larger frameworks may take away the human's ability to evaluate danger (like surpassing and stopping).

The primary benefits of using the assistance framework are that they enable communication across different cars, vehicle frameworks, and transportation, allowing the executives to concentrate on their respective areas of responsibility. By making this available, valuable data for carmakers will be exchanged, allowing for improved vision, limitation, arrangement, and fundamental leadership of cars.

Advanced Driver Assistance Systems and self-driving vehicles depend on locating and recognizing traffic signs to function properly. To make this announcement, we offer a system that can see and recognize traffic signs as a motorist is looking at them. This method relies on the driver's real-time 3D view gathered using a stereo imaging framework to see ahead and a non-contact 3D gaze tracker.

For identification, we applied a straightforward Support Vector Machine and made highlights via the Histogram of Situated Gradients. Scale Invariant Feature Transforms and coloring data are used to conduct acknowledgement. The approach we've developed both recognizes and detects warning indicators the driver is facing (and identifies signs which may have been ignored by the driver).

2 Literature Review

When applied to the issue statement, a variety of approaches may be used. We have referenced many scholarly publications. When we go over the papers for our final decision, we utilize different criteria to come up with the best solution approach. A summary of these review papers is presented below:

[Mehdi Mekni, Andre Lemieux, 2017] has introduced the paper on "Increased Reality", in which virtual setting is initially coordinated with the presentation of genuine environment visuals, a developing field of aggregate plan. The PC's continuous blending video produces content with live exhibitions. AR depends on advancements created in augmented reality and collaborates with the virtual world as well as has a level of reliance with this present reality. The expression "expanded the truth" was first begat at Boeing in 1990 by specialist Tom Coddell, who was approached to improve the costly graphs and checking apparatuses used to direct laborers on the manufacturing plant floor. Coming up next are its favorable circumstances in that it has an assortment of utilizations - medicinal, instruction, military, amusement and sports and so on. There are sure constraints with AR that must be survived. AR frameworks need to manage a lot of subtleties in all actuality. In this way little equipment ought to be utilized, effectively versatile and light and quick enough to speak to designs. Additionally the battery life utilized by these intricate AR gadgets is another constraint for AR utilization.

[K. Govindaraju, 2017] proposed a paper on Embedded Systems. An embedded system computer system that perform certain pre-determined program that are typically used on a large scale by an mechanical or electrical systems. Typically, it is introduced from small MP3 players to large scale hybrid vehicle systems. Some other examples of embedded systems often used in our daily life are - keyboards, mouse, ATM's, TV, PDA, cell phones, printers, elevators, smoke detectors, DVD player, refrigerator, cameras, GPS navigators, radios, TV remotes, Telephones, Game controllers, monitors, digital images processors, bar code readers, SD cards, washing machines, anti-lock braking systems, blender and more. We especially use embedded systems due to its dependent, efficient and it meets real-time constraints. Examples of embedded systems illustrate that we now take embedded systems for granted. Due to the implementation of smart embedded technologies in our house, we're quite acquainted with the phrase "smart home." Embedded systems are linked to the internet nearly all the time now. Embedded system applications may be roughly divided into two groups: household applications, such as dishwashers, washing machines, and microwaves, and commercial applications, such as gaming and movie playback devices. Embedded systems have a major draw-back since they are very cost-sensitive. Small adjustments in the price of a piece of construction equipment may have a significant impact on the case.

[Ronald Azuma, 2016] an augmented reality-based research proposal an AR system's ultimate aim is to increase the amount of information the user sees and interacts with by augmenting their perception and interaction with the actual world with virtual 3D objects that seem to be physically located alongside the real world. Augmented Reality work in synergy with other technology such as IOT (Internet of Thing) and many other to give people a rich view on their reality. Augmented reality is exactly what the name suggest, this is when a various reality dominates your already existing reality using certain device such as smartphone and smart-glass. Computer generated image of a different setting are superimposed on your device to change the view of reality. AR has limitations in that it may require a large amount of hardware. Despite many recent advances in AR, much remains to be done.

[Hassan Fleyeh, Mark Dougherty, 2016] a study describing an advanced way of road-side sign detection and recognition was given. This phrase explains the features of road signs, the procedures and obstacles that need to be overcome while trying to find road signs, and finally, road sign detection and recognition are essential in ITS. This is because traffic signs and road signs are vital in everyday life. These icons make it possible for motorist to comprehend the visual languages presented to them. They display current traffic conditions on the road, dangers and potential obstacles to drivers, and provide valuable information that helps drivers with their navigations. Color-based and shape-based road sign detection are among the features of road sign recognition. Because it needs more gear and is costly, it has certain restrictions.

[Lotfi Abdi, Meddeb, September 2016] The Application of Augmented Reality Inside a Vehicle" TSR, which incorporates In-Vehicle Augmented Reality (AR) For many years, the design of TSR has been a challenging issue. The road signage classifications and localizations provide as the groundwork for further methodology that's utilized for proper TSR. Weather conditions, angles of view, sign distortions, and different

backdrops all contribute to an obstruction of signals. It is critical to develop computing-efficient and highly discriminating features in order to produce a quick and robust TSR. In order to enhance driving safety and cut down on the effort, information that is easier to comprehend and process must be made available. The AR-TSR completes the vehicle's external view of traffic conditions with information about those conditions, which is shown on the driver's monitor.

[Felipe Jimenez, Aurelio Ponz, April, 2016] presents a paper on "Advanced Driver Assistance Systems for Road Environments to Improve Safety and Efficiency", which deals with 2 sides that were detected as key themes Is: "Safety and Efficiency". It facilitates the flow of details between vehicles and assists in the identification and data transmission process. Applications developed include: "Adaptive control with optimization, overtaking support systems and driver assistance". To give an early reply to risk conditions, and optimize safety measure for occupants of the vehicles and the characteristic of collisions, pre-collisions in system arise. Nevertheless it should be noted that these systems demand high amounts of reliable data about the vehicle and surrounding. It also requires support systems with motion control during man oeuvres, Collision avoidance system with the potential for developmental man oeuvres which includes pedestrians, cyclists and motorists and Cooperative optimization consideration of other vehicle and road verticals side and traffic signal detection characteristic.

[Christina olaverri Monreal, February 2017] proposed a paper on the subject: "Smartphone-based traffic sign detection" that represents the growth of TSD (Traffic Sign Detection) for traffic signal on an application using an image. It is red in colour. Traffic sign recognition (TSR) systems provide the driver with important information about road restriction and the surroundings world and identify. The traffic signal with computational powers and camera's quality provides a new platform to develop ADAS applications with lesser cost. To develop a smartphone application for the principle required for real-time detection of specific traffic signals and traffic sign detection. Key points involved in image processing and implementation for assistance included image asset, area of interest calculations, segmentations, colour recognition, size classifications, and signature recognition. In future work the number of false spies can be reduced by improving the application. High resolution requires more time to process the greater number of pixel.

[Mark Billing Hurst] had drafted a paper on the topic Augmented World (AR) is a study zone that wants to use PC-created information to improve our current reality. The PC-created image is superimposed on a PC client, as well as a point of view from this reality, thereby providing a composite vision. an enhanced reality (AR) is an experiential contact with verifiable circumstances where PC-produced perceptual data enhance the content that already exists in reality. The uses of AR military, medical, navigation, and gaming have all recently taken off. Even though reality has expanded, there are still certain obstacles and limitations that are required to stay alive. Certain defined reality limits include computer-generated data rendering and scaling for the viewpoint of the viewing environment. Constrained preparing power, small quantity of memory, and a low amount of storing are important considerations while creating a virtual world on a mobile phone.

[Mauricio Hincapié, Andrea Caponio, Horacio Rios, Eduardo González Mendívil] made a presentation on AR as a branch of VR (Virtual Reality Technology), which is sometimes referred to as "the use of unusual equipment and computer programming to create a replica of a replacement environment, which customers perceive as real or legitimate." It is innovation in virtual reality (VR) that enables a scenario in which the customer experiences himself inside a PC-generated virtual world while immersed in the world. Customers who are in the world feel and appear to be inside the PC-generated virtual world, as people do while immersed in the real world. While AR makes the client able to perceive this current world, it expands it with virtual information overlaid on top of it. On the other hand, since virtual reality supplants reality, augmented reality complements it, creating a realm in which genuine and virtual content may coexist. The positive aspects of AR are that it is extraordinarily flexible, particularly in industrial settings, and also may very well be realized in many applications. Despite its many positives, there are a few negatives that may jeopardize its authenticity in real-world applications. An additional important point to bear in mind is the weight of the equipment.

[R. Silva, G. A. Giraldi] displayed that Augmented Reality advancement has various potential applications in a wide extent of fields, including redirection, guidance, medication, planning and gathering. A fundamental sort of extended reality has been being utilized in the redirection and news business for quite a while. At whatever point you are viewing the night atmosphere figure, the speaker remains staying before changing atmosphere maps. In the studio the columnist is truly staying before a blue screen. This original photo is accompanied with computer-generated maps that are produced using a method called chromatic keying.

Another use of AR is seen in the game's level advancement. Military cockpits use displays in visors in front of the pilot or on wind-shields of cockpits to provide information to the pilot. This is an extended unscripted TV drama, which is referred to as an expanded unscripted TV drama.

[Filip Malawski, July 2018] exhibited that as of late, propelled picture preparing calculations have been utilized to investigate nature while driving and furnish the driver with helpful data. Driver help frameworks can identify people on foot, perceive street signs and give route directions. Then again, the methods for showing this data are somewhat unrefined and ordinarily incorporate a little show. Along these lines, the driver needs to isolate consideration between the presentation and this present reality. Tending to this problem may be done via Augmented Reality (AR). We propose a system in which a substantial amount of relevant data for the drivers is displayed on semi-straightforward glasses that are then inserted into the current world using AR technology. We conduct a condition inquiry on a cell phone, which is then connected to the headset. Thus, the headset is responsible for determining the location of each individual phone and changing the phone's viewpoint. Evidence of an idea in the context of a person just dozing off include a passer-by finding as well as someone nodding off, with the use of an implicit accelerometer.

[Paul George, Indira Thouvenin, Vincent Fr'emont and V'eronique Cherfaoui] By looking at how the drivers feel, we can assess how robust our planned new propelled driver assistance frameworks will be. A driving assistance architecture presented in this article is inherently linked to the client. Daria is useful for autonomous vehicles since it uses augmented reality (AR) to enhance the functionality of the sensor. Snags and their risk quantification are in the middle of the identification. Driver behavior is the other option. the driver is able to see the risky areas while keeping his eyes open by means of a suitable perceptual representation. At this point, the preliminary results indicate that our approach may be extended to automobiles, ships, or water routes. While the organizing assistance allows the client to get at its objective without earlier information out.

[Lucas Morillo Méndez, June 2018] proposed a paper on Augmented Reality. AR is progressively being created in the car space as an ADAS framework. This inexorably mainstream innovation can possibly decrease the fatalities out and about which include HF, anyway the psychological segments of AR are as yet being examined. This survey gives a brisk outline of the investigations related with the intellectual systems engaged with AR while heading to date. Related research is changed, a scientific categorization of the results is given. AR frameworks ought to pursue certain criteria to stay away from bothersome results, for example, subjective catch. Just data related with the fundamental driving errand ought to be appeared to the driver so as to stay away from impediment of the genuine street by non-driving related undertakings and high mental remaining task at hand. Notwithstanding, data ought not be appeared consistently so it doesn't influence the driving aptitudes of the clients and they don't create overreliance in the framework, which may prompt hazardous practices. Some famous employments of AR in the vehicle are route and as security framework. AR subjective results ought to be considered in these specific settings later on. This article is expected as a small scale control for makers and architects so as to improve the quality and the effectiveness of the frameworks that are presently being created. One of the potential disadvantages of AR inside the vehicle is that it can prompt intellectual catch if the AR territory is overloaded with data. This may make the driver react to the offered data instead of to the driving errand and an issue of overreliance, and subsequently, that clients take part in dangerous driving.

[Chin-Chen Chang, Shu-Chun Huag, June 2018] showed a presentation about where traffic signs are placed and how to use an assistance system. Driving along the street might provide several types of messages from street signs. To capture the attention of drivers, traffic signs have eye-catching colors and simple graphics. Drivers overlook signals from traffic signs when they drive in complicated situations or if their mental state is poor. In the case that there is a programming framework for reporting the discovery and acknowledgement of traffic signs, it can pass that information on to the driver and lower the burden of having to memorize the different types of traffic signs. When the vehicle misses a traffic sign, the platform is capable of showing a cautioning signal. This framework can help the driver software navigate the roadway, by figuring out the road state. So the driving experience is improved in huge ways, with the dangers of errors being minimized. In this research, we focus on the identification and finding of traffic signs to support vehicle assistance applications. We tried several different types of frameworks, although also checking to see if the framework execution was satisfactory. Street signs may convey numerous signals to the motorist, even while they're simply operating a car. To make sure the traffic signs stand out, they are made with striking colors and simple graphics. Despite this, a motorist can have an extremely poor understanding of traffic signage, especially if the driver is

perplexed about what they are supposed to do, or if they are in a poor mental state. Should there be a traffic sign recognition and reporting system in place, the motorist will get relevant notifications quicker and also find it easier to navigate the road. The framework has the ability to see whether the motorist misses a traffic sign. An encouraging warning This framework can help the driver software navigate the roadway, by figuring out the road state. It follows that a driver's comfort is much increased, and the risk of getting into an accident is reduced. This study focuses on how traffic indicators for driving assistance programs may help the driving population be found and distinguished. Additionally, we investigate several design blends of the framework in order to confirm the framework's performance.

[Ms.Yamini Gowardhan, Dr.Ramchand Hablani, Feb 2016] It was found that different papers on Advance Driver Assistance Systems (ADASs) have been written, with multiple objectives: include corresponding with known objects, discovering new street marks, and finding previously known street signs. This report provides information on how to spot and decipher traffic signs. Identifying and acknowledging people may be accomplished in many ways: using Hough circle changes, with changes in shade, or by converting the RGB to HSI model. When we consider how big the city is getting, and the increasing number of vehicles on the road, it is obvious that many problems might arise from congestion and traffic accidents. Implementing the PC-based vision-based driving support framework has been shown to be one of the most effective ways to address these difficulties. the recognition of traffic signs is critical to ITS (Intelligent Transport System). Numerous arrangement techniques, for example, basic graphic relation designs (for example, cross-correlation similarity), and modern programming approaches (for example, bolster vector machines, boosting, pseudorandom/normalized timberland, and so on) are effective when building an industry standard in commercial programming so the client gets a genuine end result. Another plus of expanding the inspection of traffic signs from single points to images is that it will allow for a significant reduction in error rates and better locating and identifying processes. Research teams have explored many topics connected to the progress of an automated pilot designed to identify street signs or road hazards such different cars or pedestrians. A scenario like this is possible because a driver missed a sign, or failed to plan properly. Drivers tend to ignore traffic warnings in bad weather, such a rainstorm, and focus on keeping themselves and their passengers safe. At night, drivers' ability to see people, signs, and obstacles on the road around them is reduced because lighting coming from approaching cars might momentarily dazzle them. The general method of assessing knowledge of traffic signs includes two modules: naming and describing. Shape and shading are regularly utilized to differentiate the locale, and sometimes the locale will have a traffic sign. Arrangement separates the meaning of traffic signs even further.

2.1 Comparative Study

Table: 1 Comparative analysis of Exiting Work

Title	Author & Year	Technique Used	Proposed Work	Limitation	Conclusion
DAARIA: Driver Assistance by Augmented Reality for Intelligent	Paul George, Indira Thouvenin, Vincent Fr'emont and V'eroniqueCherfao ui	Prototype of driver assistance system in augmented reality has been presented based on the weathervane metaphor adapted for a use in real conditions metaphor adapted for a use in real conditions	A lot of problem of integration (calibration, synchronization - not described here- and data exchange) have been resolved to conduct first experiment and prove the feasibility.	axe for future research is focused on the metaphor improvement	Prototype of driver assistance system in augmented reality has been presented.

Augmented Reality: Applications, Challenges and Future Trends	Andre Lemieux, Mehdi Mekni	Mobile Technology, Virtual Environment and Augmented Reality	Various Applications in military, medical, manufacturing and entertainment	AR systems have to deal with large amounts of details in reality. Therefore small hardware should be used	It describes work performed in different application domains
Driver Assistance System Using Augmented Reality Headset	Filip Malawski, 2018	Image Processing Algorithms, Augmented Reality	Proof-of-concept scenario includes pedestrian detection as well as falling asleep detection, with use of a built-in accelerometer.	driver needs to divide attention between the display and the real world	detect pedestrians, recognize road signs and provide navigation instructions.
Traffic Sign Detection and Analysis	Ms. Yamini Gowardhan, Dr. Ramchand Hablani, 2016	Color Segmentation, Shape Analysis, Conversion of Color Models	SVM Classifier and Edge Detection provides improved results.	Regions detected by Color detection cannot be determined to the exact sign region.	Using HIS color model to detect the traffic sign followed by circle detection.
Traffic Sign Detection and Recognition for Driving Assistance System	Chin-Chen Chang, Shu-Chun Huag, 2018	Convolutional Neural Networks, Color Features	Utilized two types of deep learning structures to train our system.	Hierarchical grouping methods make it little bit complex.	Selective Search in the Color space with the accuracy and recall rates, and the red filter screening results
Road And Traffic Sign Detection And Recognition	Hassan Fleyeh, Mark Dougherty,	Color Models, Neural Networks, Template Matching, Classical Classifiers.	Input does not need to be transformed into another representation state.	Training overhead, multilayer neural networks cannot be adapted due to its architecture	List of Candidate Objects that could be probable road signs

Augmented Reality as an Advanced Driver Assistance System: A cognitive Approach	Lucas Morillo Méndez, 2018	Cognitive and Perceptual Implications of AR inside the car	Provides safety, navigation and take over request assistant	Occlusions of the road elements may disturb the availability of AR information .	Cognitive outcomes of interacting with AR while driving
Smartphone-based traffic sign detection using OpenCV	Mohammad S. Aminian, Christina olaverri-Monreal, 2017	Image Processing Techniques	High object detection rate (84%) within few milliseconds of time	A high resolution device needs more time to process	Displays detected traffic signs and assists the driver in not missing important traffic information.
In-Vehicle Augmented Reality TSR to improve driving safety and enhance driver's experiences	Lotfi Abdi, ArefMeddeb, 2017	Haar cascade and the bag of visual words approach	Good compromise between resource efficiency and overall performance	Does not make use of augmented markers instead uses spatial information	A new AR-HUD approach to create real time interactive traffic animations is introduced

3 Design and Implementation

The main focus is to develop a rotating inductive road sign detection system by augmented reality and advanced posit algorithm. Glyph based augmented reality is used in our project in conjunction with the posit algorithm. The posit algorithm is extended by our novel approach. The objective is to use image processing with glyph recognition technology and better and efficient recognition and detection of road signs to aid the driver and reduce road accidents. Some glyph recognition markers have been created that show street signs in particular. (These markers should be placed on the road and made standard for real-life application). When these markers are placed in front of the camera they are identified and information is stored in the system. Information will be displayed in front of the driver in an automobile system using any device. Notify using an LCD, or mobile phone.

3.1 Posit Algorithm

The purpose of 3D pose estimation is to determine the change required to map an object from a 3D coordinate system to a 2D image. It plays a very important role since the beginning of computer vision. one specific application is to help mobile robots self- localize in unknown environments. Another application area that I am particularly interested in is augmented reality, it can help visual perception of 3D in areas such as hospitals, navigation and admissions. These techniques have already been used and will be widely used in the future. Therefore it is very important to study pose estimation. Many algorithms can perform currency estimation. One of the most popular and widely used algorithm is POSIT (from orthography to scaling with pose and repetition). Unlike some other algorithms, POSIT does not require preliminary estimation. It just requires four or more corresponding attribute points between the 3D object and the 2D image. In the POSIT algorithm, the

3D model should not have the feature points coplanar (in the same plane), otherwise it does not work. However, there is a low algorithm called Coplanar POSIT that can deal with this situation.

3.2 Details of Input/Data Used

Knowledge is available that can be used to deal with the difficulty of recognizing and detecting road sign in an effective manner. Every road signs are design, built and installed as per stringent set od rules set forth by federal committee.

Colour is set not only for the signal range (Red = stop, Yellow = hazard, etc.), but also according to the blot of colour or paint that shadows the signal, with a tolerance that is visible, at a specific wavelength should correspond to Spectrum. This is certainly an important detail, but caution will be exercised for getting use of it as the standards are set as per the controlled illuminations that prevails while performing experiments, whereas during practice, weather situations are subject to external illumination and of course Results will be affected by the colors captured by the camera. Colour over road signs also fades over time.

Text fonts as well as character heights are also regulated for shapes and dimensions as well as picture lines.

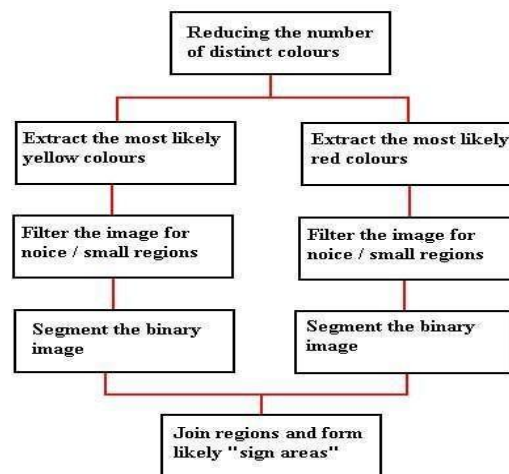


Fig. 2. Localizing Signs

Signs are mainly located at the right side of the road, in usual practice two to four meters from the roadside, which are not strictly followed and their lies expectation of overhead and clearance signs which are located on the central lane. This phenomenon is beneficial for sign recognition because a big portion of image of road can be neglected and thus speed up process.

Signs could appear in a variety of situations, including damaged, partially cloudy, and exposed to sunlight. Signs may also occur, for example, three or four signs appear above / beside each other.

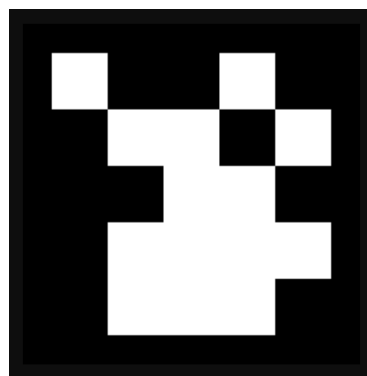


Fig. 3. - Augmented Marker

4 Experimental Results and Analysis

The Proteus Design Suite is a device suite, including one component in particular, the component that will robotize electrical structure. Schematic design and electronic print preparation are the common uses of the solution for electronic structure architects and experts. Only a few subnets of the roadway network can benefit from the concept of ad hoc architecture. A great illustration of this is a completely automated roadway in which only self-driving cars may be operated.

The fact that the first steps of intelligent vehicle research have yielded positive results, such as vehicles being able to operate themselves, means that it is technologically possible to have a complete vehicle and road automation that includes things like traffic lights and road signs (at least on highways or very structured roads). As a designer, you should be aware of and attentive about several factors outside of technical aspects, including any difficulties. Factors include laws dealing with responsibilities, failure to operate correctly, and the influence of automated driving (both human and passenger).

A good car should have an elegant, well-proportioned design and provide a comfortable environment. This is mostly down to how well-designed the passenger compartment is. The framework's interface will have a huge effect on how the framework's frame is seen and understood. There are many lengthy evaluations and reconfigurations that are required before these frameworks are open to the public. Additionally, the public has to endure more years of manually driven cars and markets before the street network can be transformed into one that uses intelligent and adaptable drivers and other means of data trade.

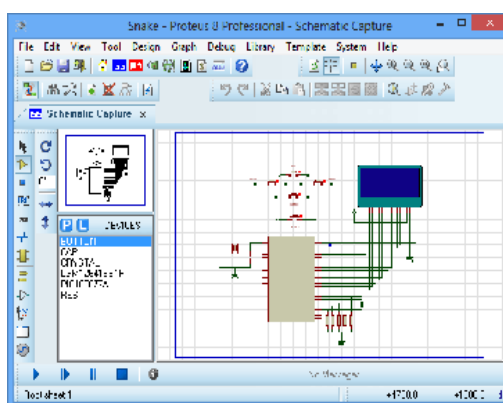


Fig. 5. – A screenshot of Proteus Design Suite

5 Results and Future Scope

The developed system quickly recognizes road signs and notifies to drive. The program also keeps track of the already passed public places such as petrol pumps and restaurants on the way and notifies the same to the driver.

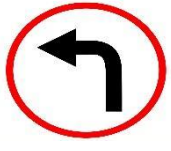
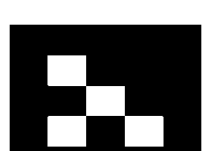
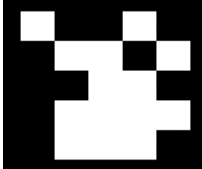
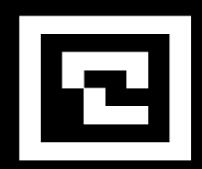


Fig. 6. – Screenshot 'Fifty'

5.1 Data Set Used Dataset

In the back end (embedded system) whenever the road sign will be recognized it will show a code on the display corresponding to that sign like for example for right turn it will display R, for left turn it will show L etc.

Table 2: Data Used for Train the Model

Name	Road Signs	Augmented Markers
No Parking		
Right Turn		
Left Turn		
U Turn		
No Left Turn		
No Right Turn		

AR is a colossal innovation at the same time, right now, it is as yet having significant issues by causing major issues that endanger its usage in mechanical world. In this report we have displayed the fundamental advantages that AR can offers to modern methods, with incredible consideration regarding look after tasks. AR could improve human exhibitions without a doubt, and this can give us extraordinary advantages not just from a practical perspective. A superior support of a car does not only mean lower price, but also great reliability and, thus, leads few failures and corresponding accidents. In our project Embedded Systems would be used at the back end and Microsoft Visual Studio would be used at the front end.

The developed system quickly recognizes road signs and notifies to drive. The program also keeps track of

the already passed public places such as petrol pumps and restaurants on the way and notifies the same to the driver.

In the back end (embedded system) whenever the road sign will be recognized it will show a code on the display corresponding to that sign like for example for right turn it will display R, for left turn it will show L etc.



Fig. 7. – Recognition of Augmented Marker

In-vehicle contextual Augmented Reality (AR) provides users with real-time visual and visceral input for a more efficient and impactful driving experience. To aid the driver in various driving situations, this technique is used to improve the traffic sign recognizer. It also makes drivers more comfortable, and also reduces the risk of vehicle accidents. Through this analysis, we have shown that AR can enhance TSR greatly. To improve driving safety and decrease driving stress, AR may be utilized. A visualization of the information presents it in a manner that's simple for the driver to comprehend, requiring minimal cognitive effort.

To control the traffic for road signs recognition using mobile phones or other smart devices, applications can be developed. Results based on calculations shows quite high object detection rate i.e. 84 % within a time of few milliseconds. The time used in detecting road signs is different from time used with the help of resolution of the device which is being used. More time is required to process more number of pixels in high resolution, but the detection comes out to be much accurate. By improving the application in future analysis one can reduce the amount of false and misleading detections of road signs.

A latest application using Augmented Reality approach is developed for creating real-time interactive traffic animations is introduced, in terms of rules for placement and visible to users, various traffic signs and movement of these is done to an in-built vehicle display of the device.

Important drawbacks that are majorly taking AR technology behind in the industrial environment were detected and solutions are proposed. Scientific solutions to solve these drawbacks are required to make AR more breakthrough technology. Much better substances, faster algorithms for solutions, small hardware elements are taken. The re- search committee takes responsibility of these needs and provides solutions in order to solve the problems.

5.2 Conclusion

Complete mechanization will be done distinctly to uncommon foundations, for example, in uses of businesses or open transportation for the present. At that point, mechanized innovation of vehicles will be gradually and bit by bit stretched out to numerous other significant transportation regions, for example, delivery of products. For instance for expensive trucks, where the expense of the vehicle itself and the ad- ministration it gives is a

lot higher than the expense of an autopilot which is practically oblivious before the previous one. At long last once innovation has been created and kept up and the most worthwhile arrangement and best calculations are utilized, huge. The developed model can recognize road signs but traffic lights had to be done in more detailed way that is the future scope of this project. The Road Sign Recognition systems will probably become an important element of future support systems of drivers in a very short period of time. Many experts and researchers have concluded that the initial generation of individual driver information and warning systems will be visible in the market in upcoming six to nine years. One of the first on the market would probably be Daimler Chrysler. The improvement of upcoming cars could be gained depending on both infrastructure and vehicles as well. Based on the particular app, user can get either some of the advantages or disadvantages. Improving the infrastructure and condition of road may give advantages to the architectures of transportation which are dependent on looping and already scheduled routes for driving, like transportation used in public and robotics in industries. On the contrary, a difficult and extensive organization and maintenance would be required for extended road networks for private vehicles which may become hard to understand and ambiguous as well as extremely costly.

References

- [1] Abdi, L., Meddeb, A., & Abdallah, F. B. (2015). Augmented Reality Based Traffic Signs Detection for better Driving experience. In *Communication Technologie for Vehicles* (pp. 94–102). Springer, Cham.
- [2] Bark, K., Tran, C., Fujimura, K., & Ng-Thow-Hing, V. (2014). Personal Navi: Advantages of an Augmented Reality. Navigational Aid Using a See-Thru 3D Volumetric HUD. In *proceeding of the 6th International Conference on Automotive User Interface and Interactive Vehicles Application* (pp. 1:1–1:8). New York, NY, USA: ACM.
- [3] Lin, J. H., Lin, C. M., Dow, C. R., & Wang, C. Q. (2011). Implementation and designing of Augmented Reality for better Driving Visual Guidance. In *2011 Second International Conference on Innovation in Bio-inspired Computing and Application*.
- [4] Müller Schneider, S. (2009). Augmented reality in a smart driver assistance system. In *workshop proceeding - 32nd Annual Conference on Artificial Intelligence, KI 2008 - Workshop on Planning, Scheduling, Design, and Configuration, PuK 2008*.
- [5] Pauzie, A. (2014). Head up display in automotive: A new reality for the user. In *design, users usability and experience: Interactive experience designing*. (pp. 506-517). Springer International Publishing.
- [6] Kim, S., & Dey, A. K. (2016). Augmenting human senses to enhance the user experience in car: augmented reality is applied and haptic approach to reduce cognitive distance. *Multimedia Tool and Application*.
- [7] De la Escalera, L.E. Moreno, M.A. Salich, and J.M. Armingol, "Road traffic sign recognition and detection," *IEEE Transaction on Industrial Electronic*.
- [8] J. Greenhalgh and M. Mirmehdi, "Real-time recognition and detection of roads traffic sign," *IEEE Transaction on Intelligent Transportation System*.
- [9] S.C. Huang, H.Y. Lin, C.C. Chang. An in car camera's system for traffic signs. *Proceeding of the Joint 16th World Congress of International Fuzzy System Association and 8th International Conference on Soft Computing and Intelligent Systems, recognition and detection, apan, June 2016*.
- [10] D.L. Kellmeyer and H.T. Zwahlen, "Determination of highway warning sign in usual video image using color image processing and neural network," in *Neural Network*, 1995.
- [11] Abdi, and L. Meddeb (2015). Augmented Reality Based Traffic Sign Recognition for Improved Driving Safety. In *Communication Technologies for Vehicles*.
- [12] Wang, J., & Söffker, D. (2016). Improving driving efficiency for hybrid electric vehicle with suitable interface. In *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 928–933).
- [13] C. Fang, S. Chen, and C. Fuh, "Road-sign detection and tracking," *IEEE Trans. on Vehicular Technology*, vol. 52, pp. 1329-1341, 2003.
- [14] A. de la Escalera, J. Armingol, and M. Mata, "Traffic sign recognition and analysis for intelligent vehicles," *Image and Vision Comput.*, vol. 21, pp. 247-258, 2003.
- [15] J. Miura, T. Kanda, and Y. Shirai, "An active vision system for real-time traffic sign recognition," presented at 2000 IEEE Intelligent Transportation Systems, Dearborn, MI, USA, 2000.
- [16] Houben, S., Stallkamp, J., Salmen, J., Schlipsing, M., Igel, C.: Detection of traffic signs in real-world images: the German traffic sign detection benchmark. In: *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE (2013)
- [17] Kun, Z., Wenpeng, W., Guangmin, S.: An effective recognition method for road information based on mobile terminal. *Math. Probl. Eng.* 2014:1–8 (2014)
- [18] Mogelmose, A., Trivedi, M.M., Moeslund, T.B.: Vision-based traffic sign detection and analysis for intelligent driver assistance systems: Perspectives and survey. *IEEE Transactions on Intelligent Transportation Systems* 13(4), 1484–1497 (2012).

Hybrid Machine Learning Technique for Personality Classification from Online Text using HEXACO Model

P. William¹, Dr. Abhishek Badholia², Dr. Anurag Sharma³, Dr. Vijayant Verma⁴

^{1,2,3,4}Department of Computer Science and Engineering,
School of Engineering and Information Technology, MATS University, Raipur (India)
¹william160891@gmail.com, ²abhibad@gmail.com, ³anusiraag@gmail.com,
⁴anil.vijayant@gmail.com

Abstract. Personality refers to a person's unique collection of traits that influence their habits, behaviours, attitudes, and thinking patterns. Text accessible on social networking sites may be used to automatically identify an individual's personality characteristics. The XGBoost classifier is used in this proposed work to predict personality traits based on the HEXACO model, namely Honesty – Humility, Emotionality, Extraversion, Openness to Experience, Agreeableness and Conscientiousness from input text using Machine Learning Technique. In the trials, a publicly accessible benchmark dataset from Kaggle is utilized. The major problem with the previous study was the skewness of the dataset, which was reduced by using the Re-sampling method, essentially random over-sampling, which resulted in improved performance. Pre-processing methods like as tokenization, word stemming, stop words removal, and feature selection using TF IDF are also used to further explore the personality from text. Although the results produced by all classifiers across all personality characteristics are acceptable, the performance of the XGBoost classifier is exceptional, reaching more than 99 percent precision and accuracy for various qualities. Individual preferences, talents, social and human values, personality variations, and human will power must all be addressed while analyzing entrepreneurial activities, intents, and performance, especially with entrepreneurship playing such an important part in the contemporary dynamic. Using the HEXACO Personality Characteristics Model, this study demonstrates a link between personality traits and entrepreneurial success, indicating that personality traits have a significant and direct effect on entrepreneurial performance.

Keywords: Personality recognition, Social Networks, HEXACO model, Individual Performance, Personality Traits.

1. Introduction

Every aspect of a person's life is surrounded by the person's personality in some manner. Individuals' behavior is predicted and characterized by a pattern of thinking, feeling, and characteristics, and by the way in which they respond to their environment. It also describes how emotions, preferences, motives, and health influence daily living activities. In recent years, the increasing popularity of social networking sites such as Twitter and Facebook have fueled the online community's need to share thoughts and sentiments with one another, thus displaying their attitude, behavior and personality (N. Majumder et al., 2017). The relationship between an individual's temperament and the behavior they exhibit on social media, whether in the form of comments or tweets, is undeniably strong. For academics, predicting personality traits from online discourse is becoming more popular in recent years. In terms of predicting personality from input text, there has already been a significant amount of study conducted. However, further work is required to improve the efficiency of the present personality identification method, which is hindered in the majority of cases by the presence of imbalanced personality characteristic classes. The personality recognition dataset is balanced using a dataset balancing method called re-sampling, which may result in better performance. Individuals' personality traits relate to their long-term, enduring qualities. The majority of studies in the late twentieth century believed that personality domains may be summarized in five dimensions, which became known as the Big Five Factors of personality. The HEXACO personality model is a six-dimensional framework that adds additional personality characteristic to the Big Five Factor Model's five qualities, making it a viable alternative to the Big Five Factors Framework.

Recently, academics have taken a keen interest in the economy's entrepreneurial arena, with a particular emphasis on modern corporate culture and entrepreneurship's contribution to the corporate sector (D. Xue et al., 2017). The goal of the research is to get a better understanding of the many elements and variables that influence entrepreneurial intentions, process, and performance. As a result, gaining a better knowledge of the

variables and processes that influence the development of entrepreneurial intents and the performance of entrepreneurial ventures would aid in encouraging more entrepreneurial success while wasting less financial and personal resources. Individuals' personality characteristics have a significant effect on their performance. As a result, the goal of this study is to determine the effect of HEXACO personality characteristics on entrepreneurial success and to get a better knowledge of how these personality categories affect and play a role in entrepreneurial performance.

Researchers are now interested in creating automated personality identification systems based on personality detection from social networking sites. Such apps' fundamental concept is based on many personality theories, such as the Big Five Factor Personality Model. This problem is mostly to blame for the decline in performance of the personality recognition system. In this paper, the benchmark personality recognition dataset is used to use a machine learning method called XGBoost to categorize the text into various personality characteristics like as Introversion-Extroversion(I-E), intuition-Sensing(N-S), Feeling-Thinking(F-T), and Judging-Perceiving(J-P) (J-P). Furthermore, to enhance the system's performance, the resampling method is used to reduce the dataset's skewness (L. R. Goldberg et al., 1990).

Personality-aware recommendation systems employ identical recommendation methods to traditional recommendation systems; the main difference is that they include the user's personality characteristics in the suggestion process. The major differences between traditional and personality-aware recommendation systems are shown in Figures 1 and 2. There are three phases in most traditional recommendation systems. The first step is the rating phase, in which the user communicates her preferences by ranking various things. The second step is filtering, which may be collaborative filtering, content filtering, or hybrid filtering, as previously stated. Finally, the system suggests the items found during the filtering step during the recommendation phase. Personality-aware recommendation systems, as illustrated in Figure 2, add two additional stages before the rating phase and alter the filtering stage as well. During the personality measurement phase, the system evaluates the users' personality types by administering a personality assessment questionnaire to them during registration, or by applying an automatic personality recognition scheme to the users' previously available data, such as data from online social networks. During the personality matching phase, the system tries to match the user's personality type with relevant items. This can be done using lexical matching, which links the textual descriptions of items to the associated personality types, or fine-grained rules, which can match items with personality types. It's worth mentioning that the system doesn't have any knowledge about the users' ratings during the personality matching phase, which helps to mitigate the impacts of the cold-start issue, which is one of the most difficult difficulties in the literature of recommendation systems. Personality-aware recommendation systems alter the filtering step by adding personality data into the similarity measurement used to identify each user's neighbors (L. B. Myers, 1962).

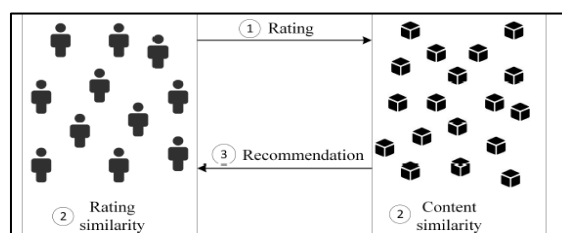


Fig. 1: Conventional recommendation systems

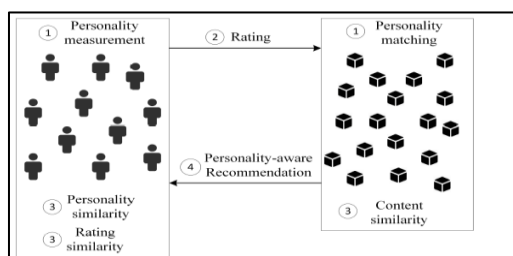


Fig. 2: Personality-Aware recommendation systems

2. Related Works

Harmonization regarding the structure of personality characteristics was a significant development in the twentieth century under the umbrella of personality psychology, i.e., most researchers agree that personality domains can be successfully summarized in five dimensions, which were dubbed the Big Five Factors. Later, there was substantial evidence in support of a substitutional demonstration of personality structure domains, and the structure was dubbed the HEXACO model of personality characteristics, which included six rather than five dimensions for personality trait evaluation. According to the HEXACO Model for Personality Traits, personality structure may be better classified into six main dimensions rather than five. In today's economy, entrepreneurship is very important. "A person undertaking innovations, resources, and business skills in an attempt to convert ideas into commercial products," according to entrepreneur, a French term. The motivations and behaviors that entrepreneurs use to make choices, grab lucrative possibilities, and maximize those opportunities are referred to as entrepreneurial behavior. Entrepreneurship may result in the formation of new organizations, such as new independent companies, or it can be used to revitalize existing organizations in response to any perceived opportunity. The most apparent type of entrepreneurship is the creation of new companies; but, in recent years, entrepreneurship has been expanded to include social and political forms of entrepreneurial activity inside established organizations (D. Shaffer et al., 1993).

Instead of providing broad "blanket" policies to all types of entrepreneurs, more emphasis is being placed on the various aspects of entrepreneurship as a contributing agent in economies, and policymakers must frame policies that support the varying needs and demands of various types of entrepreneurs. Personality traits are key predictors of an individual's ambitions to become an entrepreneur, as well as the entrepreneurial firm's future success, according to a number of significant theoretical and empirical studies. Several meta-analyses have also shown that entrepreneurs vary from other groups (i.e., managers) on a wide variety of personality characteristics, such as assertiveness. Entrepreneurs have a considerably greater risk propensity than regular managers, and they have a significantly stronger desire to succeed. Previous study has shown a link between personality characteristics and entrepreneurial success, and research has distinguished entrepreneurial and managerial populations based on these personality differences. Personality characteristics were shown to be linked to professional success, as well as the significance of personality traits in career and work happiness (S. Bharadwaj et al., 2018).

Individuals' emotional stability is defined as being emotionally steady, peaceful, and strong. Entrepreneurs are defined as those who remain calm, positive, and resolute in the face of adversity like as uncertainty, stress, or pressure. Entrepreneurs do not get sad as a result of failures and difficulties; instead, they accept and cope constructively with the physical, emotional, and mental hardships they face and continue on their way to their ultimate goal. That entrepreneurs with a high emotionality index deal with difficulties, pressures, and concerns by adopting a positive attitude, focusing on the goal, and taking direct and timely steps to address such issues. Entrepreneurs are more likely and expected to have extroverted characteristics, such as the desire and aptitude to interact effectively with stakeholders. Extroverts, according to previous study, are people who are pleasant, social, have a cheerful and optimistic outlook, and are aggressive. A proactive personality is defined by this personality characteristic. According to previous study, a proactive personality is necessary for fueling intuition and, as a consequence, creating a compelling image of an entrepreneur who is also socially acceptable. Entrepreneurial risk propensity and tolerance are likewise seen to be good constructs. Extroversion is also beneficial in establishing and producing a positive perceived locus of control among entrepreneurs, as they are motivated by their desire to succeed as well as their risk-taking propensity. Entrepreneurs are more extroverted than administrative employees, and their assertiveness is likely to have a favorable impact on entrepreneurial performance and success (M. Gjurkovic et al., 2018).

That agreeableness shows cooperative, trustworthy, tolerant, kind, and patient qualities, as well as a person's capacity to listen well, empathize, and induce harmony in their social relationships. Individuals that score low on the agreeability measure are seen to be manipulative, distrustful, and self-centered. The creation of a cooperative and trustworthy atmosphere will aid in the formation of excellent relationships among alliances, which will then enable technical exchanges and the raising of money for development and expansion. Individuals' agreeableness is also important for developing good relationships and internetworking. The capacity to establish a trusting connection among venture capitalists, the foundation of teams, and the creators of harmony and interlinking team members, according to researchers, is a key component of entrepreneurial

success. Acceptability should be favorably associated with and affecting all aspects of social entrepreneurship, resulting in entrepreneurial performance and success (B. Plank et al., 2015).

As personality characteristics, conscientiousness and openness to new experiences are inextricably connected to business success. According to research, individuals who are more conscientious are more drawn to entrepreneurship and are also more likely to succeed after becoming an entrepreneur. Conscientiousness has also been shown to be the most powerful predictor of performance. The main personality trait that distinguishes excellent performance in any management position from that of an entrepreneur is openness to new experiences. Entrepreneurs' success may be aided by thinking about fresh, creative, and unorthodox principles, concepts, and how they are implemented. As a result, we may deduce that openness to experience is a personality attribute that is strongly linked to invention, creativity, and, as a result, excellent performance in learning settings, all of which are regarded to be fundamental characteristics of entrepreneurs (O. Loyola-Gonzalez et al., 2016).

Honesty is defined as a trait that reduces an individual's chances of gaining personal benefit via the exploitation of others. Honesty is also linked to a lower chance of losses as a consequence of others' cooperation with drawl. Dishonesty, on the other hand, may be described as greedy, sneaky, arrogant, and pompous. Greed has been defined by researchers as exploitation or the desire to gain an advantage via subtle manipulation or even outright deception. To put it another way, dishonest people believe they are entitled or driven to benefit from the exploitation of others. At first sight, benefiting from others' exploitation and self-righteousness seems to be advantageous and a quick path to development and reward. An entrepreneurial process, on the other hand, is a continuous one, and profiting via deceit or manipulation will not be sustainable in the long term. Another issue that comes up here is that one may profit once in a blue moon by manipulating procedures and rules. It has two practical consequences. Other firms will follow suit, and the industry will collapse as a result of these unethical actions. Second, the practice will eventually be recognized, and counter laws and regulations will be enacted, resulting in the eradication of corrupt practices and organizations, and the organization will be eliminated. As a result, although the entrepreneurial company may prosper, develop, and become successful, it will eventually fail because its moral character foundations, particularly its potential for moral growth, will be severely harmed (I. Cantador et al., 2013).

On the other side, since an honest entrepreneur's activities are ethical and founded on solid foundations, he will be readily recognized and able to get recognition, resulting in his business being successful and thriving. Keeping the long-term development and performance of entrepreneurs in mind, it is possible to conclude that the personality characteristic of honesty is favorably and directly linked to entrepreneurial performance and growth. Entrepreneurs are also a powerful force in the development of new businesses. Individuals scoring higher in those psychological characteristics linked with entrepreneurial responsibilities and attributes are likely to be more successful entrepreneurs, if the findings of these studies are applied and included in entrepreneurial performance (B. Y. Pratama et al., 2013).

In this part, we provide an overview of the literature on personality identification from text. This work's literary studies are divided into four categories, as follows:

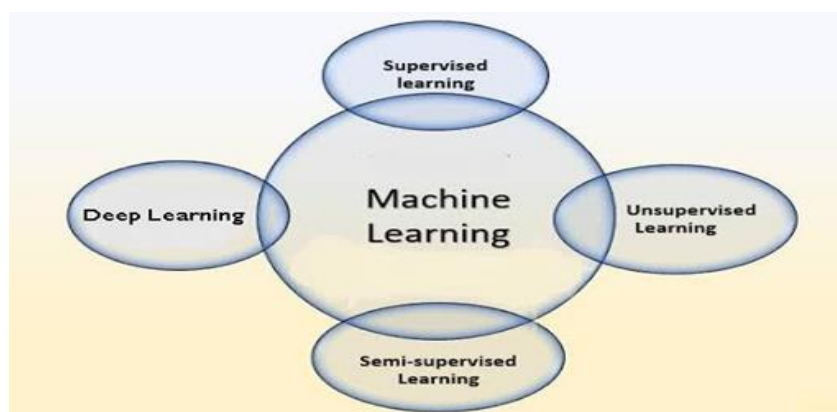


Fig. 3: Categorization Sketch of Literature Review

2.1 Supervised Learning Technique

Unlabeled data/ variables must be determined from labelled data, also known as independent variables, in these supervised learning methods. The following research are based on supervised learning techniques. Analyzing a person's social media posts/tweets and creating a personality profile as a result. The focus of the research is on data collecting, pre-processing techniques, and prediction using a machine learning algorithm. Different feature selection methods, such as Emolex, LIWC, and TF/IDF, are used to create the feature vectors. The feature vectors obtained are utilized in the training and testing of several machine learning algorithms, such as Neural Net, Nave Bayes, and SVM. SVM using all feature vectors, on the other hand, had the highest accuracy in all dimensions of the Myers-Briggs Type Indicator (MBTI) types. SVM, Logistic Regression, and other techniques are used to classify the data (MLP). Across all MBTI aspects, the classifier that included all language characteristics excelled. However, additional testing on more models is needed to get more reliable findings. The main drawback is that the postings include a huge amount of words, which makes it difficult to properly estimate personality. The four dimensions of the MBTI are predicted using a logistic regression model. As a feature selection, the binary word n-gram is utilized. This research revealed improvements in the I-E and T-F dimensions, but no gains in the S-N dimension and even a small decrease in the P-J dimension. Linguistic characteristics provide much better outcomes in terms of personality prediction. It's possible that including a better dataset will increase performance. From tweets sent in English and Indonesian, a method was built to identify user personality using the Big Five Factor personality model. On the My Personality dataset, several classifiers are used. Naive Bayes (NB) has a 60 percent accuracy, which is higher than KNN (58 percent) and SVM (50 percent) (59 percent). Although the accuracy of earlier study (61%) was not improved, the objective of determining personality from Twitter-based communications was accomplished. The findings may be improved by using a larger dataset and using a semantic approach.

For Bahasa Indonesian tweets, a personality, SVM and XGBoost, two machine learning classifiers, are used to implement various parameters such as the presence of, the elimination of stop words, and the use of LDA. Under the identical data and parameter settings, XGBoost outperformed the SVM. The major disadvantage of their approach is the limited dataset of just 359 cases for training and testing. Utilizing individual status text from Facebook to automatically identify the Big Five Factor Personality Model. For personality classification, several methods. MNB, on the other hand, outperformed the other techniques. The performance of the system may be improved by including feature selection and additional classifiers. Personality profiling was carried out using social media sites such as Twitter, Instagram, and Foursquare. For three distinct geographical areas, NUS-MSS, a multisource big dataset, is used. Various machine learning classifiers are used to assess the data for average accuracy. The classification performance improves by more than 17% when various data sources are concatenated into one feature vector.

The effectiveness of several machine learning classifiers is examined in order to evaluate the student's personality based on their Twitter accounts, with just the Extraversion characteristic of the Big 5 included. The OneR algorithm of the rules classifier performs the best of all, with an accuracy of 84 percent. In the future, all Big5 aspects may be evaluated to get additional information. To estimate a user's personality from online content, the performance of several classifiers is assessed using the MBTI model. Estimation is done using a variety of machine learning classifiers, including Nave Bayes, SVM, LR, and Random Forest. For all MBTI types, Logistic Regression has a 66.5 percent accuracy, which may be increased further with parameter tweaking. The XGBoost algorithm, which has won most Kaggle and other data science contests, may enhance the results even more. The methods of oversampling and undersampling skew the results. When applied to datasets with unbalanced classes, classification fails miserably. There are three commonly utilized methods to resolving the issue of class imbalance. In this research, the over-sampling approach (SMOTE) was shown to be more effective than the under-sampling strategy (RUS). In the future, further re-sampling methods will need to be explored (V. Ong et al., 2017).

The Big5 model is used to accurately classify Twitter users' personality characteristics based on their profiles. For prediction, a total of 50 topics with a total of 2000 tweets per user are evaluated. Two psycholinguistic techniques, LIWC and MRC, are used to analyze user content. The performance assessment is carried out with the help of two regression models: ZeroR and GP. The findings for the characteristics of "openness" and "agreeableness" are comparable to earlier research, while the results for other traits are less efficient. The findings may be improved if the dataset is expanded. TwiSTy, a new corpus of tweets for gender and personality prediction based on the MBTI type indicator, has been presented. Dutch, German, French, Italian, Portuguese, and Spanish are among the languages covered. The classifier is a Linear SVM, and the findings

are also evaluated using Logistic Regression. Character and word (n-gram) binary characteristics are used. In terms of gender prediction, it exceeded. It outperformed previous methods in personality prediction on two dimensions: I-E and T-F, but it did not improve on S-N and J-P. The summaries of the above-mentioned research for categorization and prediction of user personality using Supervised Machine Learning methods are shown in Table I.

2.2 Unsupervised Learning Approach

Unsupervised learning classifiers use just unlabeled training data (Dependent Variables) and do not predict or estimate any comparable output variables. The Twitter data was annotated for 12 distinct linguistic characteristics in order to find a link between a user's personality and writing style across various regions and devices. Users with several tweets are taken into account while evaluating them. More Twitter data for categorization may help the personality identification model work better. In the late 1990s, there was a consensus on personality structure and features, and the Big Five factor model of personality traits offered a framework for studying distinct dimensions and their roles in many areas. That the six-dimensional HEXACO model of personality traits is a valid and preferred model and alternative representation for assessing personality characteristics. Honesty-Humility (H), Emotionality (E), Extraversion (X), Agreeableness (A), Conscientiousness (C), and Openness to Experience (O) are the six distant personality characteristics in the HEXACO Model (O). The HEXACO personality assessment adds the first component, Honesty-Humility, to the Big Five Model. According to previous study, individuals with high scores on entrepreneurship-related personality characteristics are more likely to succeed in business and as entrepreneurs. Those individuals will be more likely to engage in the types of behaviors required in entrepreneurial processes and challenges, as well as exerting less conscious effort with less nervous strain, resulting in increased satisfaction, commitment, and motivation, and, as a result, higher performance and success as an entrepreneur.

Because of the crucial strategic role of the entrepreneur in the performance and success of a new endeavour, an entrepreneur's obsessive behaviour will have a significant impact on company performance. For example, show that entrepreneur personality characteristics have an indirect impact on new venture success through constructs including dedication, motivation, growth objectives, strategic decisions, and vision communication. As a result, substantial theoretical and empirical research indicate that an individual's personality characteristics are key predictors of his or her desire to become an entrepreneur, as well as his or her future performance and that of his or her business. Using unsupervised trait learning technique, investigate group-based personality identification. In the future, a larger and more diverse dataset will undoubtedly improve the performance of the suggested study. To illustrate the issue of how various personalities cooperate and behave on the social media site Twitter, an unsupervised personality categorization method was completed. This study employs linguistic and statistical features, which are subsequently evaluated on a data corpus supplemented using a personality model based on human assessment. According to system research, psychoneurotic users make more comments than secure users and have a longer chain of contact. The network visitors' characteristic and personality were recognised using an unsupervised machine learning technique called -Meas. This suggested project is based on the website's measurable content. The findings show that this approach may be used to better correctly forecast the personality characteristics of website and network users. For the extraction and categorization of user characteristics, several social media network sites are utilised. Linguistic characteristics are used to create a personality model. For an input sentence, the algorithm predicted personality and came up with an acceptable result (F. Alam et al., 2013).

2.3 Semi-Supervised Learning Approach

Semi-supervised ML methods are research that use a mix of language and lexical characteristics, supervised machine learning methodologies, and various feature selection algorithms. The semi-supervised and hybrid strategies were used in the following research. Based on their tweets, a multilingual prediction algorithm detected users' personality characteristics, age, and gender. For age and gender classification, the SGD classifier with n-gram features is employed, whereas for personality prediction, the LIWC with regressor model (ERCC) is used. In four distinct languages, an average accuracy of 68.5 percent was obtained for recognising user characteristics. Experiments in several languages, on the other hand, may improve author profiling. In the Bahasa Indonesian language, a method was developed to identify MBTI type personality characteristics using social media (Twitter). 97 individuals were chosen from 142 responses, with an average of 2500 tweets per user. WEKA is a software programme that is used to classify buildings and create a training set. For training set prediction, three methods are utilised. i) Machine Learning; ii) Lexicon-based; and iii)

Linguistic Rules-based In terms of accuracy and time, the Nave Bayes approach beat the other methods. It has an accuracy of 80% for the I/E trait and 60% for the S/N, T/F, and J/P traits. Due of the small corpus in Bhasha Indonesia, linguistic rule-driven and lexicon-based systems have lower accuracy. It has been found that expanding the training data set improves accuracy. Using word count, a method for predicting personality from social media-based writing has been suggested. It's compatible with both the MBTI and Big5 personality models, and it's available in eight different languages. The tests are conducted using four types of tagged corpus for both Big5 and BMTI. Each corpus contains 1000 of the most commonly used terms. Prediction accuracy for Big5's "openness" characteristic is higher across the board, whereas MBTI's S/N dimension has better prediction accuracy than the other dichotomies. The suggested system's major flaw is its reliance on word counts for prediction, which may be mitigated by using alternative feature selection and machine learning techniques (K. Buraya et al., 2017).

2.4 Deep Learning Strategy

Deep learning is a subset of machine learning (ML) in artificial intelligence (AI), in which computers can learn and gain experience without the need for human intervention to make choices. Deep learning executes tasks repeatedly, improving and modifying outcomes with each iteration. The papers listed below are summarised versions of previous Deep learning research. Using the MBTI dataset, a deep learning classifier was built that takes text/tweet as input and predicts the author's MBTI type. Following the use of several pre-processing methods, an embedding layer is employed, which maps all lemmatized words to create a dictionary. The model was tested using a real-life example of Donald Trump's 30,000 tweets, and it accurately predicted his MBTI personality type. As input, take a sample of text or a post and categorise it into various personality characteristics, like as (INFP, ENTP, and ISJF, etc.). For performance assessment, several classification techniques such as Softmax as a baseline, SVM, Nave Bayes, and deep learning are used. SVM beat NB and softmax in terms of train and test accuracy, with 34 percent train and 33 percent test accuracy, but the Deep Learning model surpassed them both with 40 percent train and 38 percent test accuracy. The accuracy, however, is still poor, as it does not even reach 50%. Using deep learning techniques, a personality categorization system recognises characteristics from online literature (N. R. Ngatirin et al., 2016).

Based on an essay dataset, a deep learning network classifies personality characteristics using the Big Five personality paradigm. To identify personality characteristics from an input essay, a Convolutional Neural Network (CNN) is employed. For personality trait categorization, several pre-processing methods such as word n-grams, sentence, word, and document level filtering, and extracting distinct characteristics are used. Using various configurations of characteristics and across all five traits, "OPN" traits obtained a better accuracy of 62.68 percent. More features will be added in the future, and the LSTM recurrent network may be used to get better results. The work on an automated personality identification system utilising Deep learning technique is summarised in Table IV. Extended annotated corpus, on the other hand, may improve the system's performance (S. Chaudhary et al., 2018).

3. Methodology

3.1 Data Level Re-Sampling Approach:

The goal of data manipulation sampling is to balance all class instances by rescaling the training datasets. Over-sampling and under-sampling are two common class resizing methods. The most well-known methods at the data level are oversampling and under sampling processes (A. More, 2016). Oversampling is a strategy for increasing the number of classes in the minority. Random oversampling is the easiest kind of oversampling since it just duplicates minority occurrences to increase the imbalance percentage. The effectiveness of the machine learning classifier for predicting personality characteristics was much enhanced by this doubling of minority class improvement. The indiscriminate removal or deletion of majority class instances is employed in the under sampling method to level class distribution (P. Kaur and A. Gosain, 2018). This procedure is repeated until the occurrences of the majority and minority classes are equal as shown in Figure 4.

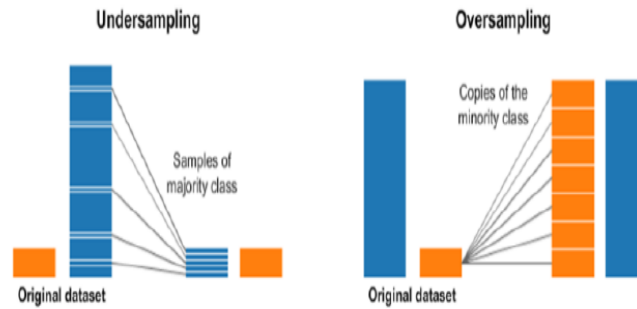


Fig. 4: Class Balancing using Under sampling and Oversampling.

3.2 K Fold Cross Validation:

This method involves dividing the data randomly into k subgroups of about equal size. The first fold of data is put aside for testing, while the remaining $k-1$ subsets are utilised to train the model. This procedure is repeated until the testing set for each Cross-validation fold (of k iterations) is complete. Because this process is performed k times,



Fig. 5: K-Fold Cross Validation Working Procedure.

The Mean Square Error is computed k times as well. As a result, the mean of the Mean Square Error across k folds is used to compute the k -fold Cross Validation error. The working method of K -Fold cross validation is shown in Figure 5.

4. Results and Discussion

Main aim of this contribution is to find a suitable machine learning technique for decision assistance for personality classification that delivers accurate and comprehensible results while remaining within budget constraints.

Table 1. Descriptions of decision/learning problems

Method	Problem				
	Ger-man	Gene Expr	Qual-ity	Spam	Direct
C45	72,8	63,4	58,6	93,0	89,6
C45Rules	72,4	57,1	60,0	94,2	89,7
LDA*	75,7	-	53,3	88,8	89,9
LibSVM	76,3	70,9	65,2	92,5	89,3
kNN	74,5	66,9	65,0	90,8	89,0
Random For-ests	76,3	66,3	69,8	95,4	89,8
Functional Trees	75,5	68,6	60,2	93,4	90,2
NBTree	75,3	-	56,7	93,2	89,6
LMT	75,9	-	60,6	93,7	90,2
Stacking	76,4	70,3	68,0	95,0	89,9
Hoeffding Trees	75,6	64,6	48,6	81,6	87,9

Only the prediction accuracy of chosen techniques is systematically evaluated in this study, as shown in Table 1. Random Forests and Stacking, the two ensemble techniques under consideration, are each indicated by a distinct table cell color. The size of the resultant information representation and the kind of learning technique used are used to subjectively evaluate understanding ability. As demonstrated in Figure 6, hybrid techniques (FT, LMT) only exhibit minor increases in prediction accuracy over conventional understandable methods. The prediction accuracy of hybrid ensemble techniques is similar to that of conventional ensemble methods (Figure 7).

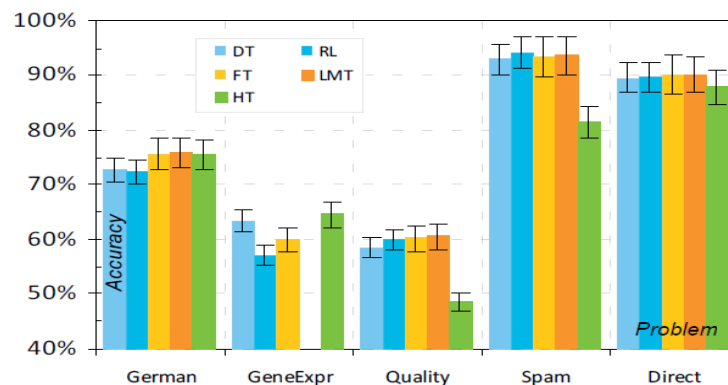


Fig. 6: For five distinct issues, the accuracy of basic understandable machine learning techniques was compared against hybrid methods.

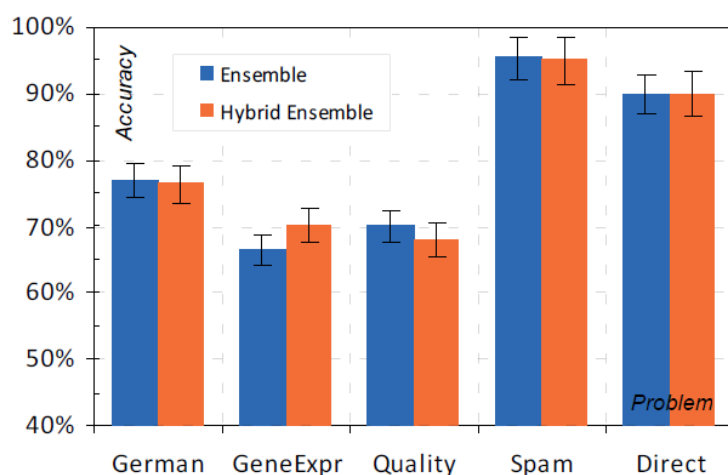


Fig. 7: For five distinct issues, the accuracy of ensemble and hybrid ensemble machine learning techniques was tested.

5. Constraints and Limitations

Other personality models, such as the Big Five Factor (BFF) and DISC personality Assessment models, have not been tested or researched for classification of personality traits. The proposed work's textual data for personality evaluation is completely comprised of English, with no material from other languages being evaluated. Simple over-sampling and under-sampling methods are employed to balance and level the skewness of the dataset. Because the dataset originates from just one site, the personality cafe forum, the findings may be skewed. All of the experiments in this proposal are based on conventional or classical machine learning methods. Only one social networking site, Twitter, is used to classify textual contents for personality trait identification; however, other social networking sites are ignored. In his proposed work, only textual data is analysed and investigated for the recognition of user personality traits. In text classification, feature extraction is given less weight, and just the TF-IDF method is used.

6. Future Research

Future study may look at the influence of organisational and industrial cultures' moderating and mediating impacts. The kind of company and industry needs may also have an effect on the impact of certain personality

characteristics. Entrepreneurs should also be distinguished as those who pursue this career on their own volition or those who were forced into it because they are a heterogeneous population. The effects of various personality characteristics may also be investigated at various phases of the entrepreneurial process, namely at the launch and post-launch stages. Future research should look into how different aspects of an entrepreneur's personality relate to specific entrepreneur behaviours or characteristics of a new venture, all of which are linked to the firm's success. For a more accurate evaluation of the characteristics, the personality modelS should be compared to the other model. Textual data in several languages, particularly Urdu and Pashto, may be analysed for personality categorization. For more stable and consistent performance, SMOTE (Synthetic Minority Over-sampling Technique) may be used as a class balancing technique. Multiple benchmark datasets may be required to gather labelled data from different sites such as "Reddit." Deep learning algorithms may be used in further personality identification studies. Other social networking sites, such as FACEBOOK posts and comments, must be investigated for automated personality characteristic inference. The job of identifying personality characteristics may be tested using data from social networking sites in the form of pictures and videos. For the suggested work to be improved, more sophisticated feature selection methods must be used.

7. Conclusion

The use of various machine learning techniques on the benchmark, Personality dataset used to classify text into different personality traits such as Honesty – Humility, Emotionality, Extraversion, Openness to Experience, Agreeableness and Conscientiousness is the central theme of this research. The HEXACO model is used for text classification and personality trait assessment. Following the application of class balancing techniques to the imbalanced classes, personality traits are detected using machine learning classifiers such as KNN, Decision Tree, Random Forest, MLP, Logistic Regression (LR), SVM, XGBoost, MNB, and Stochastic Gradient Descent (SGD). To evaluate and examine the overall effectiveness of the prediction model, evaluation measures such as accuracy, precision, recall, and - score are employed. The findings indicate that although all classifiers' scores across all personality characteristics are acceptable, the performance of the XGBoost classifier is exceptional. For I/E and S/N characteristics, we achieved more than 99 percent precision and accuracy, and for T/F and J/P dimensions, we received approximately 95 percent accuracy. The KNN classifier, on the other hand, had a lower overall performance.

In today's contemporary dynamic economy, entrepreneurship plays a critical and significant role. The development of a better understanding of the entrepreneurial process and metrics involved in those processes, as well as the study of the variables that facilitate and boost the entrepreneurial process and performance, has become necessary because failure of an individual in the capacity of an entrepreneur can be costly to society in terms of squandered resources and missed opportunities. Personality dimensions have a role to play in improving this important knowledge, according to one research.

References

- [1] N. Majumder, S. Poria, A. Gelbukh and E. Cambria, "Deep Learning-Based Document Modeling for Personality Detection from Text," in IEEE Intelligent Systems, vol. 32, no. 2, pp. 74-79, Mar.-Apr. 2017.
- [2] D. Xue et al., "Personality Recognition on Social Media with Label Distribution Learning," in IEEE Access, vol. 5, pp. 13478-13488, 2017.
- [3] L. R. Goldberg, L. R. , "An alternative" description of personality": the big-five factor structure," Journal of personality and social psychology, vol. 59, no. 6, p.1216, 1990
- [4] I. B. Myers, "The Myers-Briggs Type Indicator: Manual" ,1962
- [5] D. Shaffer, M. Schwab-Stone and P. Fisher, "Preparation, field testing, interrater reliability and acceptability of the DIS-C," J Am Acad Child Adolesc Psychiatry, vol. 32, pp. 643-648, 1993.
- [6] S. Bharadwaj, S. Sridhar, R. Choudhary and R. Srinath, "Persona Traits Identification based on Myers-Briggs Type Indicator(MBTI) - A Text Classification Approach," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, 2018, pp. 1076-1082.
- [7] M. Gjurić and J. Šnajder, "Reddit: A Gold Mine for Personality Prediction," In Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media , pp. 87-97, 2018.
- [8] B. Plank, and D. Hovy, "Personality traits on twitter—or—how to get 1,500 personality tests in a week." In Proceedings of the 6th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 92-98, 2015.

- [9] O. Loyola-González, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa and M. García-Borroto, "Study of the impact of resampling methods for contrast pattern-based classifiers in imbalanced databases," *Neurocomputing*, 175, pp. 935-947, 2016.
- [10] A. More, "Survey of resampling techniques for improving classification performance in unbalanced datasets," 2016, arXiv preprint arXiv:1608.06048
- [11] P. Kaur and A. Gosain, "Comparing the Behavior of Oversampling and Undersampling Approach of Class Imbalance Learning by Combining Class Imbalance Problem with Noise," In *ICT Based Innovations* , pp. 23-30, Springer, Singapore, 2018.
- [12] I. Cantador, I. Fernández-Tobías and A. Bellogín, "Relating personality types with user preferences in multiple entertainment domains," In *CEUR workshop proceedings*, Shlomo Berkovsky, 2013.
- [13] B. Y. Pratama and R. Sarno, "Personality classification based on Twitter text using Naive Bayes, KNN and SVM," 2015 International Conference on Data and Software Engineering (ICoDSE), Yogyakarta, 2015, pp. 170-174.
- [14] V. Ong et al., "Personality prediction based on Twitter information in Bahasa Indonesia," 2017 Federated Conference on Computer Science and Information Systems (FedCSIS), Prague, 2017, pp. 367-372.
- [15] F. Alam, E. A. Stepanov and G. Riccardi, "Personality traits recognition on social network-facebook," *WCPR (ICWSM-13)*, Cambridge, MA, USA, 2013.
- [16] K. Buraya, A. Farseev, A. Filchenkov and T. S. Chua, "Towards User Personality Profiling from Multiple Social Networks," In *AAAI*, pp. 4909-4910, 2017.
- [17] N. R. Ngatirin, Z. Zainol and T. L. C. Yoong, "A comparative study of different classifiers for automatic personality prediction," 2016 6th IEEE International Conference on Control System, Computing and Engineering (ICCSCE), Batu Ferringhi, 2016, pp. 435-440.
- [18] S. Chaudhary, R. Sing, S. T. Hasan and I. Kaur, "A comparative Study of Different Classifiers for Myers-Brigg Personality Prediction Model," *IRJET*, vol.05, pp.1410-1413, 2018.

Implementing Variability Techniques in Software Systems

Ashima Arya¹, Bhawna², Payal Chhabra³, Neha Bhatia⁴, Karuna

^{1,2,3,4,5} Assistant Professor, IT Department, PIET, Samalkha
ashiarya@gmail.com

Abstract—Variability is the ability of a system to be changed and adapted according to various situations and environments. The need of variability is to make the product more flexible and reuse. So far we have studied variability in software product line but now, it is required in software systems also. Variability is planned activity and implemented using variant features and variation points in each phase of software engineering. It has become a key property of many systems. We analyze existing research on variability in software systems. We investigate variability managing process that is implemented in each phase of software development life cycle. Variability is classified into eight dimensions that describe its types and mechanism. This classification provides a integrated approach of variability in a unifying manner in software systems. From the survey we have find out the implementing techniques on various system software such as service based systems, web based systems, adaptation systems and the effect of variability in each system.

Keywords—variability; feature model; dimension; service oriented system, web service oriented system.

1. Introduction

Variability is the ability of a software system to be changed so that it fits a specific context. These changes are adopted by enabling variation points and variants in software structure, behaviour or underlying process. The aim of variability is to ensure that systems successfully adapt changes and reusability of software system [1]. Variability has mainly been studied in the software product line [2], but now variability is built with software system such as open platforms, adaptive system, service based system and web based system [3]. Variability is needed and handled in various situations such as Configuration of system, runtime adaptation of a software service, customization of components [4].

2. Managing Variability Process

There are several activities from enabling of variations to the implement of variant points in software system during software development life cycle [5]. Variability is planned and anticipated change rather than change due to errors, maintenance or customer needs [6]. Different steps involved in Variability process is presented in fig1.

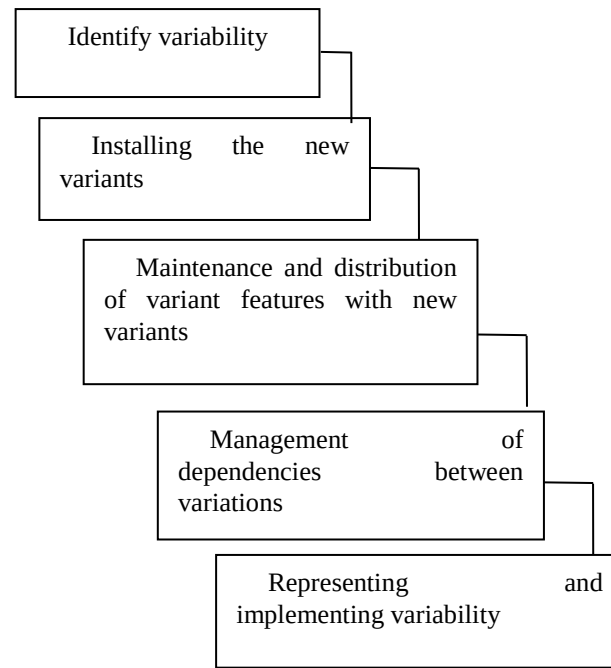


Fig1. Managing variability process

3. Variability Modelling

Variability modelling needed for user communities who have different goals and different product usage context. The different usage context may require different capabilities to handle quality attributes which have different characteristics. Variability models represent the common features of product in product line. Software engineering uses them to design product line architecture to maintain the features that depend on it [7].

Feature Model

Feature model is introduced as a part of the Feature Oriented Domain Analysis (FODA) methodology. Feature model is represented as tree like structure contain cross- tree constraints among all the features. From the survey different authors have proposed new approaches of FODA that extends the features of this modelling. different approaches are described in Table 1.

Table1: FODA Approaches

Reference no	Feature Modelling Approach	Description
[8]	FORM Feature Model	Introducing exclusive viewpoints which includes capability, operating environment, area technology, and implementation technology
[9]	FeatuRSEB Feature Model	Creating change by providing different notations for variant features, constraints, binding time for use and reuse time
[10]	Van Gorp et al. Feature Model	Introducing outside capabilities together with refining the generalization dating to OR-specialization and XOR-specialization relationships and giving crucial time notation for compile-time, link-time, and run-time binding.

[11]	PLUSS Feature Model	Doing significant changes for a cluster of alternative features as sole adaptors and a group of elective features as compound adaptors.
[12]	Hein et al. Feature Model	Delivering languages depending upon Unified Modelling Language.
[13][14]	GP-Extended Feature Model • Cardinality-Based Feature Model	Familiarizing the notion of feature cardinality, feature group, and group cardinality

Decision Modelling

[15] introduced the decision modelling technique for modelling variability. A decision model is comprised of domain-associated questions, set of possible solutions to the given problem, variation points and effect of decisions for each answer. It emphasizes on capturing decisions that are made in shaping products.

4. Dimensions of Variability

Variability is the capability to transform or amend a system. Here, classifying the variability in different dimensions is done. By surveying, it is found that a stable set of eight dimensions are formed. These dimensions are classified in two clusters. One is type of variability and the second is mechanism of variability [16,17,18]. A comprehensive brief about the dimensions is provided in table 2 and table 3.

Table1: dimension of variability-Type of variability

Dimension	Domain	Description
Requirement	Functionality	Required functionality in terms of input, output and behaviour of the system.
	Quality	How well the system deliver its functionality
Representation	Feature model	Graphical representation of all valid variant points
	Rules	Refer to the elements to realize or represent the system.
	Variant labels	Variability is represented by labels
	Profiles	Represents the descriptive summary of the users in the environment.
	Change scenario	Describe events, sequence and how they effect the system.
Artifact	Scenario	Related to any aspects of the system
	Business process	Gathering of correlated, structured activities or tasks that gives a precise service.
	Architect	Related to software architecture of the system
	Code Fragment	Variability in the source code during fragmentation in the module
Orthogonality	Separated	Variability is specified in distinct components
	Integrated	interwoven within other components of the system.

Table3: Dimension of Variability- Mechanism of Cluster

Dimension	Domain	Description
Trigger (binding of variants to variation points)	Stakeholder	Variant binding is initiated by the stake holder
	Business process	Binding is done with different business processes

	System	Binding is done with different processes on different systems
	Environment	Triggered on different component of a system in an environment
Realization technique(technical approach to variability)	Reorganization	Reorganize the structure and behaviour of the system
	Selection	Selection of variants of among all the features
	Value assignment	Assigning a value to the variation points
	Code generation	Generating the code of variant that is bound to a variation point.
Time of binding(time when variant are bound to variation points)	Software construction	Variants are bound to different points during system construction.
	Runtime	Variants are bound to different points during system operations.
Automation(degree of support to automate variant binding)	Manual	Binding is done manually
	Semi-automatic	Binding is done manually but with help of tools
	Automatic	Binding is done fully automatically.

Source: Variability In Software Systems—A Systematic Literature Review IEEE Transactions On Software Engineering/ VOL. 40,NO. 3,MARCH 2014

4.Implementing Variability on Software Systems

Service Oriented System

In these type of software systems new services are added, remove and replace with existing services by implementing new variant of services. A variability model represents variability in abstract level such as configuration of variation points during runtime .Different variability techniques are integrated with service oriented technologies to find the solution of variability in service oriented systems.

The Pattern Plugin (PP) approach [19] –this approach is also known as a generative approach. The service variants that enable the variation points are generated using the design model of this approach. In service oriented architecture, these points are chosen pattern plug-in describes an individual variant and transformed into code artifacts.

Component-based Service Implementation (CbSI) [20] integrates variability at the component level and refines services by adding a component layer.The approach a conceptual rather than generic.

Feature-Oriented Programming (FOP) is being used in the FOP-based Refinement (FOPbR) technique [21]. (FOP). A feature module in FOPbR encapsulates the code of a feature. A feature module is made out of a tree-like structure that connects different features to the main code.

FOP techniques also applied to SoC in the Class Wrapper (CW) approach [22]. CW, on the other hand, implements Java HotSwap to change bytecode in place while maintaining the same class identification. HotSwap is necessary to add features to the service implementation in the form of base classes and wrappers (plain Java code). The base class code simply changes internal algorithms and has no effect on the class schema.

Aspect-Oriented Programming (AOP) and message interception are used in the Aspect Service Weaver (ASW) technique [23]. Existing service communication chains (based on SOAP) between service consumer and provider are intercepted by ASW.

Web-Service Based System

Web –Service Based Systems coordinates the internetworked system to address the interoperability issues. Web Services' dynamic execution environment makes it possible to implement runtime. Variability management in these systems facilitates in the implementation of good services, increases system availability, and can be used to optimise system features by modifying configuration and rebinding services during runtime. Variability related issues in software industry are handled by a variability management framework called COVAMOF[24] which provides modelling capabilities at many variation points and dependencies across numerous abstraction layers.

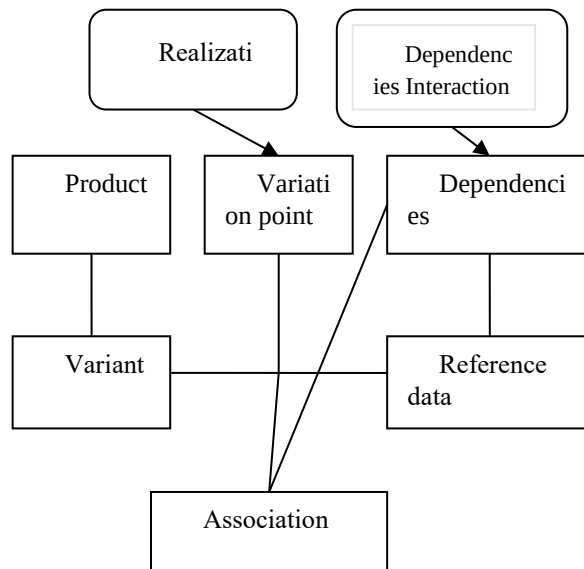


Fig 2: The COVAMOF model

This model works with following concepts:

- Variation points and variants: Location of variants is called variation point. Variants decide what effective action is executed when the variant is selected at particular location.
- Realization: It specifies the rules that identify variants at lower layer of abstraction.
- Dependency: It works as a bridge between various components. It's a system characteristic that determines how variation points are bound together.
- Association: The term "association" refers to a component of an entity. It is the value that has an impact on the system property.
- Reference Data: The value of the system attribute is stored here.

BPEL[25] is the defining language for web service-based systems that allows for runtime flexibility. It contains XML components that contain the necessary data.

The UML extension mechanism is defined by the meta-meta-modeling architectures which has four layer .each layer defines by this mechanism[26].

5. Conclusion

Variability refers to a software system's ability to be modified to meet a specific situation. In this paper we have studies variability modelling, dimensions of variability, managing variability process and implementation of variability techniques in system software such as service oriented architecture and web based service systems..From the survey we have concluded that variability ensures that systems successfully adapt changes and reusability of software system and how these dimensions are classified to solve the particular problems in system software. We have implemented dimensions such as profiles, scenarios, realizations in different techniques .Many More techniques will be implemented using these dimensions in future.

References

- [1] J. van Gurp, J. Bosch, and M. Svahnberg, "On the Notion of Variability in Software Product Lines," Proc. Working IEEE/IFIP Conf. Software Architecture, pp. 45-54, 2001.
- [2] L. Chen and M.A. Babar, "Variability Management in Software Product Lines: An Investigation of Contemporary Industrial Challenges," Proc. 14th Int'l Software Product Line Conf., pp. 1-15, 2010.
- [3] R. Hilliard, "On Representing Variation," Proc. Workshop Variability in Software Product Line Architectures, pp. 312-315, 2010.
- [4] M. Galster and P. Avgeriou, "Handling Variability in Software Architecture: Problems and Implications," Proc. Ninth IEEE/IFIP Working Conf. Software Architecture, pp. 171-180, 2011.
- [5] M. Svahnberg, J. van Gurp, and J. Bosch, "A Taxonomy of Variability Realization Techniques," Software—Practice and Experience, vol. 35, no. 8, pp. 705-754, Apr. 2005.
- [6] Williams and J. Carver, "Characterizing Software Architecture Changes: A Systematic Review," Information and Software Technology, vol. 52, no. 1, pp. 31-51, Jan. 2010.
- [7] Kang, K.C., Cohen, S.G., Hess, J.A., Novak, W.E., Peterson, A.S.: Feature-oriented domain analysis (FODA) Feasibility study. Technical report, CMU/SEI-90-TR-21, November 1990
- [8] Kang, K.C., Kim, S., Lee, J., Kim, K., Shin, E., Huh, M.: FORM: a feature-oriented reuse method with domain-specific reference architectures. Ann. Softw. Eng. 5, 143–168 (1998)
- [9] Griss, M.L., Favaro, J., d'Alessandro, M.: Integrating feature modeling with the RSEB. In: 5th International Conference on Software Reuse, pp. 76–85 (1998).
- [10] van Gurp, J., Bosch, J., Svahnberg, M.: On the notion of variability in software product lines. In: Working IEEE/IFIP Conference on Software Architecture, pp. 45–54 (2001)
- [11] Eriksson, M., Borstler, J., Borg, K.: The PLUSS approach – domain modeling with features, use cases and use case realizations. In: Obbink, H., Pohl, K. (eds.) SPLC 2005.
- [12] Hein, A., Schlick, M., Vinga-Martins, R.: Applying feature models in industrial settings. In: 1st International Software Product Line Conference, pp. 47–70 (2000).
- [13] Czarnecki, K., Bednarsch, T., Unger, P., Eisenecker, U.: Generative programming for embedded software: an industrial experience report. In: Batory, D., Consel, C., Taha, W. (eds.) GPCE 2002. LNCS, vol. 2487, pp. 156–172. Springer, Heidelberg (2002)
- [14] Czarnecki, K., Helsen, S., Eisenecker, U.: Staged configuration using feature models. In: Nord, R.L. (ed.) SPLC 2004. LNCS, vol. 3154, pp. 266–283. Springer, Heidelberg (2004)
- [15] Atkinson, C., Bayer, J., Bunse, C., Kamsties, E., Laitenberger, O., Laqua, R., Muthig, D., Peach, B., Wust, J., Zettl, J.: Component-Based Product Line Engineering with UML. Addison-Wesley, London (2002).
- [16] M. Svahnberg, J. van Gurp, and J. Bosch, "A Taxonomy of Variability Realization Techniques," Software—Practice and Experience, vol. 35, no. 8, pp. 705-754, Apr. 2005
- [17] R. Kazhamiakin et al., Adaptation of Service-Based Systems, M.P Papazoglou et al., eds., pp. 117-156, Springer Verlag, 2010.
- [18] M. Ivarsson and T. Gorschek, "A Method for Evaluating Rigor and Industrial Relevance of Technology Evaluations," Empirical Software Eng., vol. 16, no. 3, pp. 365-395, 2011.
- [19] N. C. Narendra, K. Ponnalagu, B. Srivastava, and G. S. Banavar. Variation-Oriented Engineering (VOE): Enhancing Reusability of SOA-Based Solutions. In: 5th IEEE Intern. Conference on Services Computing, pages 257–264, 2008.
- [20] M. Medeiros, E. S. de Almeida, and S. R. L. Meira. Towards an Approach for Service Oriented Product Line Architectures. In: 3rd Workshop on Service-Oriented Architectures and Software Product Lines, 2009.
- [21] S. Apel, C. Kaestner, and C. Lengauer. Research Challenges in the Tension Between Features and Services. In: 2nd Intern. Workshop on System Development in SOA Environments, pages 53–58, 2008.
- [22] N. Siegmund, M. Pukall, M. Soffner, V. Köppen, and G. Saake. Using Software Product Lines for Runtime Interoperability. In: Workshop on Reflection, AOP and Meta-Data for Software Evolution, pages 1–7, 2009
- [23] V. Monfort and S. Hammoudi. Towards Adaptable SOA: Model Driven Development, Context and Aspect. In: 7th Intern. Joint Conference on Service-Oriented Computing, pages 175–189, 2009.
- [24] Sybren Deelstra, Marco Sinnema, Jan Bosch, Product derivation in software product families: a case study, Journal of Systems and Software, v.74 n.2, p.173-194, 15 January 2005
- [25] Michiel Koning, Chang-ai Sun, Marco Sinnema, Paris
- [26] Avgeriou, VxBPEL: Supporting variability for Web services in BPEL, Information and Software Technology, v.51 n.2, p.258-269, February, 2009.
- [27] Nenad Medvidovic, David S. Rosenblum, David F. Redmiles, Jason E. Robbins, Modeling software architectures in the Unified Modeling Language, ACM Transactions on Software Engineering and Methodology (TOSEM), v.11 n.1, p.2-57, January 2002

A Modified approach for Accuracy Enhancement in Intruder Detection with optimally Certain Features

Shivani Gaba¹, Shally Nagpal², Alankrita Aggarwal³ and Dr. Deepak Kumar⁴

^{1,2,3}Assistant Professor, ⁴Associate Professor, Department of Computer Science and Engineering, Panipat
Institute of Engineering and Technology, Panipat, Haryana

¹sgsgknl@gmail.com, ²shally.ngpl@gmail.com, ³alankrita.agg@gmail.com,

⁴wadhwa123deepak@gmail.com

Abstract: As mobile ad-hoc networks are vibrant and incessantly varying ad-hoc network, so having centralized checking on the same is impossible. Vehicular ad-hoc network is as like mobile ad-hoc network where vehicles keep on interconnecting with adjacent cars as well as roadside units. Here in this paper, we use mobile ad-hoc networks which are the combination of various mobile nodes which reduces interruption however could not remove them. Intrusion detection in MANETs is an assignment which can be associated to the machine learning area. In this paper our main focus is mainly on reduction of features for achieving extreme precision and to reduce the overhead time of machine learning algorithms. The interruption dataset is reserved from the customary NSL-"KDD" dataset. Here in this paper the genetic algorithm (GA) is substituted by the Gravitational search algorithm (GSA) which delivers the maximum precision and consumes very less time for training the model.

Keywords: MANETS, IDS, VANET, "KDD"

I.INTRODUCTION

MANET is a vibrant and incessantly varying ad-hoc network, so having centralized monitoring on it is impossible. VANET is highly vibrant in nature. Identifying the interloper in it is a challenging undertaking. An interruption detection system (IDS) screens network traffic and alarms the framework or organization executive. IDS can likewise react to abnormal traffic by impeding the client or source IP address from getting to the organization [1].

A few environments (like the military-strategic activities) have exceptionally severe prerequisites on security, which make the sending of safety-related innovations fundamental. Like encryption and verification, interruption avoidance measures can be utilized in MANETs to lessen interruptions yet can't dispense with them. The ancient framework of safety research has shown that irrespective of the quantity of disruption avoidance procedures are exploited; there are some weak spots in the framework [1][4]. In an organization with high-security prerequisites, it is essential to convey interruption recognition procedures. The stunning contrasts between MANETs and wired organizations make it unimportant to apply for customary wired ID advances straightforwardly to mobile ad-hoc networks. MANET doesn't have a decent framework. While the more significant part of the present wired IDSs, which depend on ongoing traffic parse, channel, organization, and investigation, generally screen the traffic at switches, switches, and entryways.

Every node may utilize the fractional and limited correspondence exercises as the accessible review follows. MANET also has a few qualities, like disengaged activities, which only sometimes exist in wired organizations. It is recommended that IDS of an alternate design be created to be relevant on the MANET stage.

Interruption Detection in ad-hoc networks is the errand which can be identified with AI field. The informational collection is accessible with NSL-"KDD" information for interruption discovery. The dataset comprises of past history of gatecrashers which are field names and their mathematical qualities. This dataset is exceptionally huge so dimensionality decrease must be performed to choose the best appropriate elements which gives most noteworthy exactness and furthermore consumes less time in preparing the model.

II.DATASET DEPICTION

The NSL-”KDD” informational collection is an enhancement above the “KDD”99” informational group. The “NSS-”KDD” informative collection has 42 properties that are utilized in a detailed review. In “NSS-”KDD”, informational index copy occurrences were eliminated to dispose of one-sided grouping results. Many quantities of adaptations are accessible of this dataset, out of which 20% of the preparation information is utilized. Preparing information is recognized as “KDD” Train+ 20Percent with an absolute number of 125973 examples. The test information is recognized as “KDD” Test+, and it has an aggregate of 22543 samples. The quantity of characteristics in each is 42. With varieties in the number of occasions, various sorts of setups of the informational collection are accessible. Marked property 42 is the 'class' property which shows that the occurrence is an ordinary association example or an assault. In table 1, the portrayal of “KDD” dataset credits with class names is displayed. Out of 42, 41 characteristics can be arranged into four distinct classes as talked about underneath:

- Basic (B) Structures are the properties of discrete TCP associations.
- Content (C) highlights are the properties inside an association proposed by the space information
- Traffic (T) highlights are the traits processed utilizing a two-second time window
- Host (H) highlights are the traits intended to evaluate assaults that continue for over two seconds.

Table 1: The details of “KDD” data set attributes (class-wise)

Serial no.	Attribute name	Serial no.	Attribute name	Serial no.	Attribute name	Serial no.	Attribute name
1	Duration	11	Num_failed_logins	21	Is_hot_login	31	Srv_diff_host_rate
2	Protocol_type	12	Logged_in	22	Is_guest_login	32	Dst_host_count
3	Service	13	Num_compromised	23	Count	33	Dst_host_srv_count
4	Src_bytes	14	Root_shell	24	Error_rate	34	Dst_host_same_srv_rate
5	Dst_bytes	15	Su_attempted	25	Rerror_rate	35	Dst_host_diff_srv_rate
6	Flag	16	Num_root	26	Same_srv_rate	36	Dst_host_same_srv_port_rate
7	Land	17	Num-file_creations	27	Diff_srv_rate	37	Dst_host_diff_srv_host_rate
8	Wrong_fragment	18	Num_shells	28	Srv_count	38	Dst_host_serror_rate
9	Urgent	19	Num_access_files	29	Srv_serror_rate	39	Dst_host_srv_serror_rate
10	Hot	20	Num_outbound_cmd	30	Srv_rerror_rate	40	Dst_host_rerror_rate

III.THE SELECTION OF OPTIMIZED FEATURES FROM DATASET

The dataset devises 42 characteristics for each example and not all examples add to precision due to either every one of the zeros in the property or not characterized values in that. Likewise certain non-applicable properties are additionally furnished with the information which might diminish the accuracy. So we really want to discover them and eliminate those to work on the accuracy. For this reason we depend upon gravitational search algorithm (GSA) which is a worldwide meta heuristic improvement technique dependent on developments of planets in a circle. Every planet draws in other with its gravitational power and power of fascination will be more for which mass is higher. Here the mass addresses the accuracy in interloper location

for a specific arrangement of qualities. As the time expands, the circle gets longer and gravitational steady which is to be viewed as a consistent so far will rot with time as in condition 1.

$$(t) = G0e^{-\alpha t/T} \quad (1)$$

"G0 and α are introduced toward the start and will be decreased with time to control the search accuracy". T is the absolute number of emphases. These cycles are characterized by client and most extreme emphases changes from one application to another. Inside these emphases, the most extreme accuracy for a picked set of qualities should be achieved and no more expansion in accuracy ought to be there. At the end of the day an immersion condition of accuracy in these cycles ought to be acquired which implies the framework has accomplished and found the ideal arrangement of traits for which accuracy is greatest. GSA strategy does this assignment iteratively. The places of specialists in GSA are addressed by double digits. The 1 address the characteristic is picked and 0 addresses that trait is disposed of. A sum of 41 0's and 1's series address the specialist's positions. In an emphasis, there are 20 specialists which have 20 arrangement of 0's and 1's series. We compute the accuracy for that multitude of 20 unique ascribes mixes utilizing SVM (Support vector machine) characterization strategy and a matrix is created for these 20 qualities. Every one of specialist's positions are refreshed by following conditions:

$$x^d_{i \ d(t+1)} = v_i \ d(t+1) + x_i \ d(t) \dots (2)$$

$$v_i \ d(t+1) = rand \cdot v_i \ d(t) + a_i \ d(t) \dots (3)$$

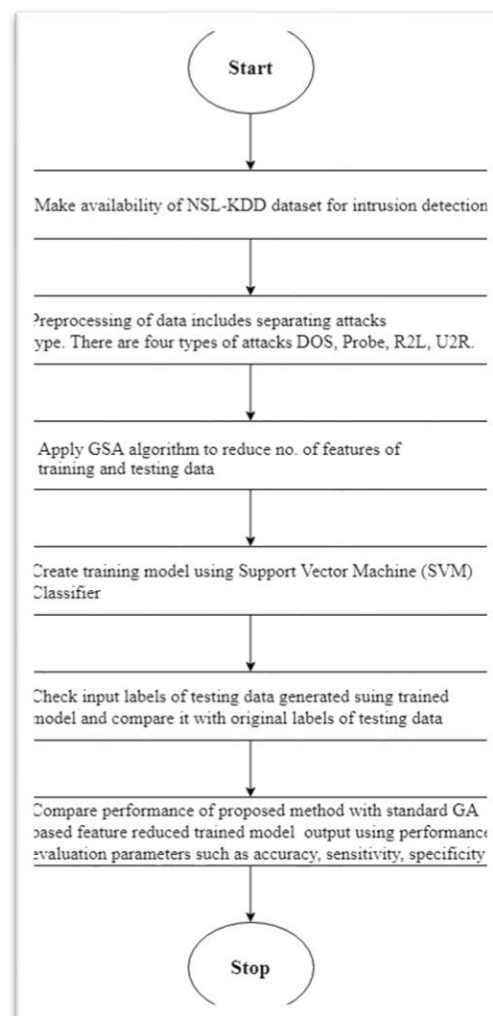


Figure 1: Complete Layout of work

“Where $xi d(t+1)$ is the new agent's position for the next iteration and $xi d(t)$ is the present position. $vi d(t+1)$ is the new velocity of movement and $ai d(t)$ is the present acceleration. This can be further calculated as”:

$$ai d(t) = Fi d(t) \dots (4)$$

$Fi d(t)$ is the total force acting on it agent calculated as:

$$Fi d(t) = \sum randjFij d(t) j \in kbest \dots (5)$$

$Fij d(t)$ can be computed as:

$$Fij d(t) = G(t) \cdot (Mpi(t) \times Mai(t) Rij(t) + \epsilon) \cdot (xj d(t) - xi d(t)) \dots (6)$$

(t) is the mass of an agent which is normalized accuracy value for each agent. It is formulated as:

$$mi(t) = (t) - wor(t) \text{ best}(t) - worst(t) (7)$$

Where (t) is the fitness value of each agent, $wors(t)$ is the minimum accuracy value among all present agents and $best(t)$ is the maximum accuracy value.

The new refreshed situation of each of the 20 specialists as in condition 3.2 is utilized to really take a look at the accuracy for them. One more matrix is created which stores the accuracy for that multitude of new arrangement of traits. This interaction proceeds till all cycles are not wrapped up. At last the most extreme accuracy ordered specialist's position is considered as the ideal arrangement of characteristics. For these characteristics just accuracy got considered as greatest accuracy. This way it lessens the overhead in characterization and works on the time and test accuracy. Complete pipeline for the work is displayed in figure 1. Complete pipeline for the work progress is displayed in figure 1. Check yield names of testing information produced utilizing prepared model and contrast it and unique names of testing information.

IV.RESULTS

In this work, we have proposed a thorough report on GSA (Gravitational Search Algorithm) advancement calculation for highlight decrease for making a superior interruption identification framework (IDS). A ton of inbuilt capacities in MATLAB makes the utilization simpler and saves our chance to fabricate our code without any preparation to utilize that time in the issue arrangement of examination. The dataset contains 125973 preparings informational collection and 22543 informational testing index with absolute 41 components having three emblematic provisions and yield is ordered as sorts of assaults. We applied GSA to decrease the number of conditions from including the table of preparing and testing dataset. The SVM multi-class classifier preparing information is used to make a preparation model, and testing information is tried for anticipated yield names utilizing this trained model. We have separated our outcomes into four cases relying on kind of assault. There are four significant sorts of assaults, for example, Disavowal of administration (DoS), User to Root (U2R), Probing, and Remote to User (R2L) assaults.

Case-1 Denial of Service (DOS) Attack In the NSL-”KDD” dataset, DOS assault is additionally classified into six subtypes. We have analyzed the execution of GSA-based element decrease with GA-based element decrease.

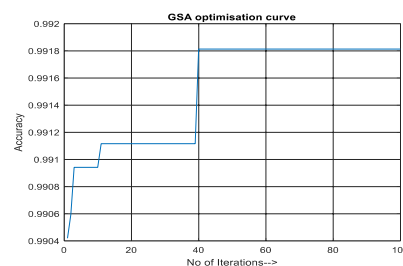


Figure 4.1: Iteration curve for DOS attack

It is seen that proposed strategy has more accuracy, accuracy, review, F measure, affectability and particularity for DOS kind of assault. The accuracy and Fmeasure bend are displayed in figure 4.2 and 4.3.

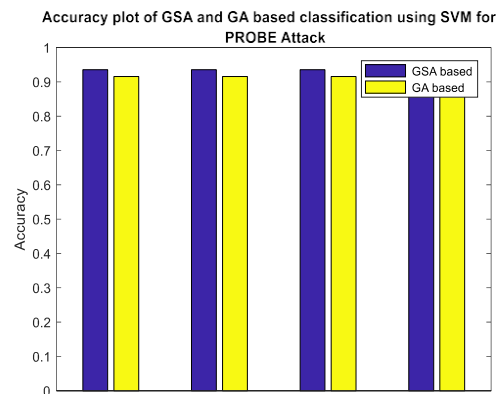


Figure 4.2: Comparison of precision between genetic algorithm and Gravitational search algorithm selected features for PROBE attack

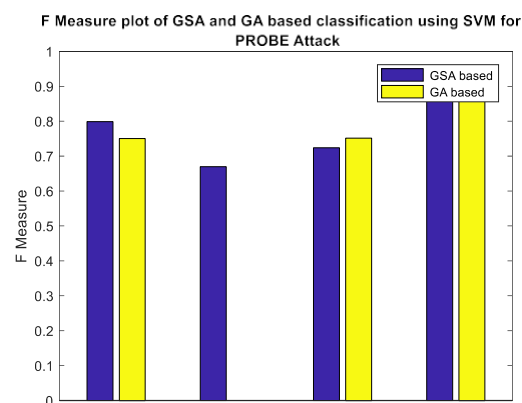


Figure 4.3: F-measure comparison genetic algorithm and Gravitational search algorithm selected features for PROBE attack

Case-3 User to Root (U2R) Attack

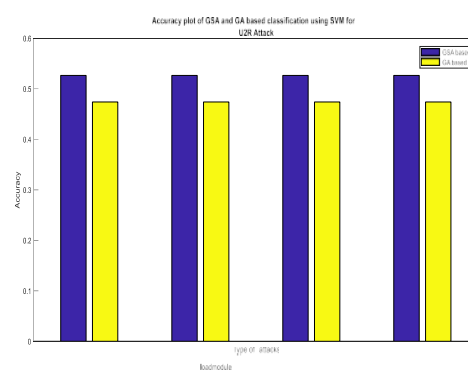


Figure 4.4: Precision Comparison for U2R attack

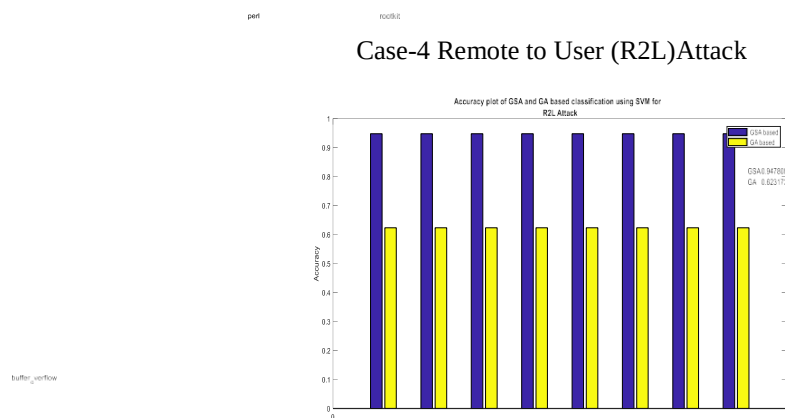


Figure 4.5 Precision comparison for R2L attack

V.CONCLUSION

MANET give wellbeing from the assaults. AI field give the arrangement of this issue. This work utilized public accessible NSL-"KDD" dataset which is changed and sifted adaptation of "KDD" cup dataset. The work gave extraordinary precision and devour less time when contrasted with past one. GSA is utilized distinctly for those elements which really participates in assault identification .The more seasoned one utilize the GA (Genetic Algorithm) however in this we use GSA which results 98% precision in these assaults. GSA gives 54% more precision than the GA.

REFERENCES

- [1] Z. Ullah, M. S. Khan, I. Ahmed, N. Javaid and M. I. Khan, "Fuzzy-Based Trust Model for Detection of Selfish Nodes in MANETs," 2016 IEEE 30th International Conference on Advanced Information Networking and Applications (AINA), Crans-Montana, 2016, pp. 965-972.
- [2] M. A. Abdelshafy and P. J. B. King, "Dynamic source routing under attacks," 2015 7th International Workshop on Reliable Networks Design and Modeling (RNDM), Munich, 2015, pp. 174-180.
- [3] C. Alocious, H. Xiao and B. Christianson, "Analysis of DoS attacks at MAC Layer in mobile adhoc networks," 2015 International Wireless Communications and Mobile Computing Conference (IWCMC), Dubrovnik, 2015, pp. 811-816.
- [4] A. Quyoom, R. Ali, D. N. Gouttam and H. Sharma, "A novel mechanism of detection of denial of service attack (DoS) in VANET using Malicious and Irrelevant Packet Detection Algorithm (MIPDA)," International Conference on Computing, Communication & Automation, Noida, 2015, pp. 414-419.
- [5] A. M. Shabut, K. P. Dahal, S. K. Bista and I. U. Awan, "Recommendation Based Trust Model with an Effective Defence Scheme for MANETs," in IEEE Transactions on Mobile Computing, vol. 14, no. 10, pp. 2101-2115, Oct. 1 2015.
- [6] A. Menaka Pushpa and K. Kathiravan, "Resilient PUMA (Protocol for Unified Multicasting through Announcement) against internal attacks in Mobile Ad hoc Networks," 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Mysore, 2013, pp. 1906-1912.
- [7] M. A. Abdelshafy and P. J. B. King, "Analysis of security attacks on AODV routing," 8th International Conference for Internet Technology and Secured Transactions (ICITST-2013), London, 2013, pp. 290-295.
- [8] A. M. Kurkure and B. Chaudhari, "Analysing credit based ARAN to detect selfish nodes in MANET," 2014 International Conference on Advances in Engineering & Technology Research (ICAETR - 2014), Unnao, 2014, pp. 1-5.
- [9] S. Biswas, P. Dey and S. Neogy, "Trusted checkpointing based on ant colony optimization in MANET," 2012 Third International Conference on Emerging Applications of Information Technology, Kolkata, 2012, pp. 433-438.
- [10] D. Das, K. Majumder and A. Dasgupta, "A game-theory based secure routing mechanism in mobile ad hoc network," 2016 International Conference on Computing, Communication and Automation (ICCCA), Noida, 2016, pp. 437-442.
- [11] T. Poongothai and K. Duraiswamy, "Intrusion detection in mobile AdHoc networks using machine learning approach," International Conference on Information Communication and Embedded Systems (ICICES2014), Chennai, 2014, pp. 1-5.

- [12] D. A. Varma and M. Narayanan, "Identifying malicious nodes in Mobile Ad-Hoc Networks using polynomial reduction algorithm," 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), Chennai, 2016, pp. 1179-1184.
- [13] Bandana Mahapatraa and Prof.(Dr) Srikanta Patnaik, "Self Adaptive Intrusion Detection Technique Using Data Mining concept in an Ad-Hoc Network," 2nd International Conference on Intelligent Computing, Communication & Convergence (ICCC-2016)
- [14] Manjula C. Belavagi and BalachandraMuniyal, "Performance Evaluation of Supervised Machine Learning Algorithms for Intrusion Detection," 12th International Multi-Conference on Information Processing-2016 (IMCIP-2016).
- [15] PreetiAggarwala and Sudhir Kumar Sharmab, "Analysis of "KDD" Dataset Attributes - Class wise For Intrusion Detection," 3rd International Conference on Recent Trends in Computing 2015 (ICRTC-2015).
- [16] Ciza Thomas, Vishwas Sharma and N. Balakrishnan, "Usefulness of DARPA Dataset for Intrusion Detection System Evaluation
- [17] P.Natesan and P.Balasubramanie, "Multi Stage Filter Using Enhanced Adaboost for Network Intrusion Detection," International Journal of Network Security & Its Applications (IJNSA), Vol.4, No.3, May 2012
- [18] M. Tavallae, E. Bagheri, W. Lu and A. A. Ghorbani, "A detailed analysis of the "KDD" CUP 99 data set," 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications, Ottawa, ON, 2009, pp. 1-6

Designing Next Generation Recommender Systems

Chaitali Choudhary, Dr. Inder Singh
Ph.D. Scholar, Assistant Professor (S.G.)
School of Computer Science, UPES Dehradun
chaitali.choudhary@gmail.com, inder@ddn.upes.ac.in

Abstract: The core of the recommender system is based on the idea that, as online shopping has progressed, it has become more usual for users to share their purchasing, activity, reviews, and rating with other users (both online and offline), which is at the heart of the recommender system. All of this is dependent on a foundation of trust and mutual interest, which the recommendation system captures. When all of the pieces come together, you have a rich information repository that can be utilized to enhance the user experience and boost engagement if done correctly. Data management systems that can handle the massive quantities of information produced on a daily basis are required in order to provide a customized user experience in social networks. Because of the high number of users in the recommender system, the system confronts a number of challenges, the most significant of which are the following: data sparsity, cold start, diversity, and scalability. While this paper concentrates on different technological solutions to address key problems encountered by recommender systems, it also briefly describes how they function and what their limits are in a straightforward manner.

Keywords: Recommendation system, Cold Start Problem, Data Sparsity, Scalability, Framework, machine learning techniques, Community Detection

Introduction

Another approach to describe the word "recommender system" is as follows: software that helps users in narrowing down their wide options and then providing them with the best suggestions based on their requirements or choices. Duke University researchers were the first to create a recommender system in the 1970s, and the Xerox Palo Alto Research Center was the first to further develop the technology in the 1980s. The Palo Alto Research Center of the Xerox Corporation has been credited for developing the world's first recommender system in the 1960s. Consumers welcomed recommendation systems as a means of helping them in choosing the most suitable products from among the vast array of available options when the Internet became widely available in the 1990's. As a result of their increasing popularity, recommender systems have become essential for large Internet corporations working on various areas such as social networking, entertainment, ecommerce, tourism, matchmaking, and other areas of business. Some of the most well-known companies that have used recommender systems in their operations include Facebook, Amazon, Netflix, Google, YouTube, and Tripadvisor to name a few. Candidates who have been recommended for promotion must achieve four secondary objectives in order to be successful: All of the considerations mentioned below must be taken into account: All of the following factors are essential to consider: data sparsity, cold start, variety, and scalability.

Cold Start Issue:

When the information on a piece of content or a service is not readily available, recommendation engines work less efficiently. When a visitor does not have prior information, which can be sub-divided into two categories: not knowing anything about the user and not knowing anything about the user's current session, the two most common types of extreme cold-start issues are: first, the problem of not having enough feedback or rating information from other visitors; and second, the problem of not having enough feedback or rating information from other visitors. For example, someone looking to get their first smartphone may find themselves on an ecommerce website and continue their search from that point forward. He made up his mind about buying a phone in less than a week, and he's no longer exploring any other alternatives for the time being. So, what kind of information should the recommendation engine include in its recommendation at this point?

In an online e-commerce network, using community detection tools allows for the following: 1) identifying community members similar to the target user; 2) predicting the active user's rating by looking at similar

community member ratings; 3) recommending new products to the target user; and 4) making recommendations on popular movies to users in the real-time system, due to the tool's usefulness for commercial purposes. Solution: (RS). You must use the user-movie matrix to identify the community and then give a brief description of the community's most recent movie in order to assist in solving the issue of Item-Cold Start. Similar to how its description may be used to plant seeds across a community, the description of a film can serve as inspiration for the creation of a whole new film. In the case of a user who is totally unfamiliar with the recommender system, they may encounter a problem known as the User Cold-Start issue (X). The recommender systems, on the other hand, have access to the information included in the user's profile (Y). In recommender systems, the modularity maximization technique, which is accessible based on the user-movie matrix that has been given, may be used to identify a community of interest by maximizing the number of modules in the system (Y). Example: In recommender system Y, we may take use of the external community profile that is available in recommender system Y, such as those from Amazon and eBay, in order to recommend an item for the new user in another recommender system (such as eBay). It is possible to identify a group of users in a system who are similar to a new target user who has joined the system by using the community detection method, which is described below. Therefore, depending on the scenario, we may conclude that the two users have similar book ratings and that they also have similar movie preferences; or that the two of them are friends with one another, as the case may be. In order to predict that the new user's interest in another dimension will be utilized as a solution to the user cold-start problem, it is necessary to collect this information in order to develop a solution that will be executed.

This new deep learning-based matching method, which they claim is less expensive, solves the cold-start and sparsity issues in CF-based recommendation systems while requiring little to no modification of the current system, they claim in their paper (Yu et al., 2018). A hybrid recommendation model developed by Jian Wei and colleagues to address the cold start problem employs a deep learning neural network to discover item content features, which are then used to the temporal SVD++ CF model to provide recommendations. The technique employed by Jian Wei and his colleagues is quite similar to the one that was used in this research. Several information retrieval solutions, including the ones discussed here, have been published. They employ a hybrid recommender, which makes recommendations based on user similarity, and a boosted recommender, which also employs interaction-based collaborative filtering to deal with the cold start and sparsity issues in information retrieval, as well as a boosted recommender, which also employs interaction-based collaborative filtering to deal with the cold start and sparsity issues in information retrieval (Palla et al., 2005).

It is possible to connect information to the user's external social information if the user's external social information is accessible. If the user's external social information is not accessible, the information may not be linked. One of the most creative approaches has been successfully deployed, and it takes use of information from social networks to fill in the gaps and find commonalities among users. Information from social networks is used to fill in the gaps and identify commonalities among users in this method, which uses information from social networks. The gathering of characteristics from different components of social networks as a consequence of this community-based approach will be utilised to help recommendation systems in resolving the cold-start problem based on the latent similarities that were identified. As part of their research, Li and Tang (2016) examined the most effective way to make a recommendation to an unknown user based on the opinions of an existing group of users, using methods from social choice theory as a basis for their conclusions. In the area of social choice theory, for example, models for aggregating individual preferences and judgments in order to arrive at a collective decision have been created to assist people in their decision-making. Socially aware systems benefit users because they enable them to find new peers in whom they can place their confidence since they have already established trust and connections with others as a consequence of their past interactions. Pitsilis and Knapskog developed a trust-based approach for leveraging latent relationships between users that is based on mutual trust and is based on the concept of trust (Pitsilis & Knapskog, 2009). Its foundation was built on the idea of trust, as well as the concept of trust, which served as the basis for this method. By doing so, users who do not need to be familiar with one another may discover and benefit from the points of view of participants who are physically separated from one another.

Data sparsity:

Data sparsity arises as a consequence of the fact that users of ecommerce websites choose to interact with a limited number of goods (e.g., adding an item to the basket, making a purchase, or leaving a review). While the majority of recommendation engines take into account evaluations from similar users, the reported user-

item matrix contains up to 99 percent empty or unknown ratings, owing to a lack of incentives for users to evaluate items or a lack of sufficient knowledge about a specific item to rate it appropriately (Jang et al., n.d.). As a result, those who do not provide comments or assessments get recommendations that are completely irrelevant to their requirements.

Solution: Popularity bias has been recognized as an issue in recommendation and ranking systems. This refers to an algorithm's tendency to prefer a small number of popular items while not representing the vast majority of other items. In the past, a number of techniques for decreasing popularity bias and enhancing the recommendation of long-tail, less popular products were explored. In the future, the issues of popularity bias and data sparsity won't be a worry because of a widely recognized idea known as the Pareto Principle (Abdollahpouri et al., 2021).

According to the "80/20 Rule," roughly 80% of outcomes may be ascribed to 20% of the causes (Albert & Barabási, 2002). For example, this rule could be formalized as a common rule of thumb in the information security sector, where it's widely accepted that the adoption of 20 percent of available security measures can effectively handle 80 percent of threats. This rule reinforces the commonly held belief that distributions of possible values for risks or resources will seldom be equal, which is supported by other rules. Current recommendation methods, as opposed to the previous section's approach of considering the Web application as a complete economic system – in which users are mutually linked in terms of allocations – place a strong emphasis on providing allocations that correspond to the preferences of each individual user. Pareto inefficiency is caused by a lack of a global perspective, which implies that we may be able to significantly enhance the suggestions by benefiting some users while not harming others, suggesting that the system did not achieve the most optimal possible allocation. If a resource's whole quantity can only be allocated to one (or a small number of) users, this issue is particularly severe since it results in the exclusion of all other users from the resource. As a result, the notion of Pareto Efficient Recommendation (PER) has been established, which refers to a recommendation system that gives the best possible (that is, Pareto optimum) suggestions when there is no possibility for any user to obtain any more advantages without harming the other users (Rani, 2014).

A Multi-Objective Optimization (MOO) framework is given to maximize the surplus of each user at the same time, and suggestions are made based on the Pareto optimal solutions that are discovered as part of the process. A Pareto Improvement Process (PIP) is also provided in order to profit from the many existing recommendation algorithms that are now accessible. This procedure will transform their suggestions into recommendations that are Pareto efficient. In experiments using real-world datasets, it has been demonstrated that PIP outperforms current algorithms in terms of both recommendation performance and consumer surplus, whereas the direct PER approach outperforms current algorithms in terms of both recommendation performance and customer surplus.

Data sparsity arises as a consequence of the fact that users of ecommerce websites choose to interact with a limited number of goods (e.g., adding an item to the basket, making a purchase, or leaving a review). While the majority of recommendation engines take into account evaluations from similar users, the reported user-item matrix contains up to 99 percent empty or unknown ratings, owing to a lack of incentives for users to evaluate items or a lack of sufficient knowledge about a specific item to rate it appropriately (Jang et al., n.d.). As a result, those who do not provide comments or assessments get recommendations that are completely irrelevant to their requirements.

Solution: Popularity bias has been recognized as an issue in recommendation and ranking systems. This refers to an algorithm's tendency to prefer a small number of popular items while under-representing the vast majority of other items. In the past, a number of techniques for decreasing popularity bias and enhancing the recommendation of long-tail, less popular products were explored. In the future, the issues of popularity bias and data sparsity won't be a worry because of a widely recognized idea known as the Pareto Principle (Abdollahpouri et al., 2021).

According to the "80/20 Rule," roughly 80% of outcomes may be ascribed to 20% of the causes (Albert & Barabási, 2002). For example, this rule could be formalized as a common rule of thumb in the information security sector, where it's widely accepted that the adoption of 20 percent of available security measures can effectively handle 80 percent of threats. This rule reinforces the commonly held belief that distributions of possible values for risks or resources will seldom be equal, which is supported by other rules. Current recommendation methods, as opposed to the previous section's approach of considering the Web application

as a complete economic system – in which users are mutually linked in terms of allocations – place a strong emphasis on providing allocations that correspond to the preferences of each individual user. Pareto inefficiency is caused by a lack of a global perspective, which implies that we may be able to significantly enhance the suggestions by benefiting some users while not harming others, suggesting that the system did not achieve the most optimal possible allocation. If a resource's whole quantity can only be allocated to one (or a small number of) users, this issue is particularly severe since it results in the exclusion of all other users from the resource. As a result, the notion of Pareto Efficient Recommendation (PER) has been established, which refers to a recommendation system that gives the best possible (that is, Pareto optimum) suggestions when there is no possibility for any user to obtain any more advantages without harming the other users (Rani, 2014).

A Multi-Objective Optimization (MOO) framework is given to maximise the surplus of each user at the same time, and suggestions are made based on the Pareto optimal solutions that are discovered as part of the process. A Pareto Improvement Process (PIP) is also provided in order to profit from the many existing recommendation algorithms that are now accessible. This procedure will transform their suggestions into recommendations that are Pareto efficient. In experiments using real-world datasets, it has been demonstrated that PIP outperforms current algorithms in terms of both recommendation performance and consumer surplus, whereas the direct PER approach outperforms current algorithms in terms of both recommendation performance and customer surplus.

Scalability:

As a consequence of the rapid growth of the ecommerce sector, scalability problems are becoming more difficult to resolve: current-generation recommendation engines are required to deliver real-time results for large-scale applications, which is getting increasingly difficult to do. Therefore, it is essential to evaluate the performance of a recommendation model by evaluating its throughput (the number of inferences produced per second) and latency (the amount of time it takes for each inference to be processed) (Sardianos et al., n.d.).

A significant body of research has established the capability of recommender (or recommender) systems to anticipate users' reactions to a huge number of choices. In online services that are linked to music and video, as well as in online shopping and other activities, it is common to see this kind of cookie being used. Information on the preferences of different users is gathered in order to anticipate their future choices, and it is used to make predictions about them. Facebook and other social networking sites have lately included recommender algorithms, which are becoming more popular in the corporate world (Schifferer et al., 2020).

Over the course of the last several years, matrix factorization techniques have been used to the development of a range of recommendation systems. In order to do this, the algorithms set out to find two new user-item interaction matrices W^T and H^T that are as close to the original matrix A in terms of how they interact with the objects that they are given as is reasonably possible. H^T matrix represents user profiles, whereas W^T matrix represents product categories and commodity categories. WT matrix represents product categories and commodity categories. Users' preferences with regard to an item j may be established in a straightforward manner by using the variables W^T and H^T in the most basic of ways. As an added benefit, since these recommendation systems are based on matrix operations, they are flexible enough to be performed in parallel with other applications. In light of the enormous size of WT and HT and the nature of these algorithms, several authors have chosen to parallelize their work. While still generating high-quality outcomes, they are able to save both time and labour by using this method. A number of well-known algorithms, like the ALS and the SGD, are available in parallel versions for both shared- and distributed-memory systems. Two well-known algorithms, Alternating Least Squares (ALS) and Stochastic Gradient Descent (SGD), are both available in parallel versions for both shared-memory and distributed memory architectures. Alternating Least Squares (ALS) and Stochastic Gradient Descent (SGD) are both available in parallel versions for shared-memory and distributed memory architectures. Both the Alternating Least Squares (ALS) and the Stochastic Gradient Descent (SGD) algorithms are available in parallel versions for both shared-memory and distributed memory systems, and both methods are available in parallel versions for both shared- and distributed memory architectures.

In recent study, it has been shown that using coordinate descent-based approaches (CCD) is more efficient than employing adaptive local search methods (ALS). When compared to adaptive local search (ALS)

techniques, coordinate descent-based approaches (CCD) are more efficient. (ALS). Data matrix operations have been successfully implemented on general-purpose graphics processing units (GPGPUs) as a result of the increasing popularity of general-purpose graphics processing units (GPGPUs) and their suitability for data-parallel programming. Taking use of the hundreds or thousands of processing units available on these systems, the algorithms may be implemented on a broad variety of platforms without the need for any extra hardware to be purchased (cores). Regarding recommendation systems, we are only aware of one paper, which was published in 2001 and was based on a neighborhood-based approach for graphics processing units. Two recommendation methods that make use of the graphics processing unit (GPU) of the computer are presented in this paper on the basis of the matrix factorization methodology).

Parallel ALS(Rodrigues et al., 2015):

Input: A,W,H λ , T

1. Initialize ($H \leftarrow$ (small random numbers));
2. for iter $\leftarrow$ 1 to T Step=1 do:
3. Compute in parallel the W using equation:

$$w_i^* = (H_{\Omega i}^T H_{\Omega i} + \lambda I)^{-1} H^T a_i // \text{Sync}$$

4. Compute in parallel the H using equation:

$$h_j^* = (W_{\Omega j}^T W_{\Omega j} + \lambda I)^{-1} W^T a_j // \text{Sync}$$

It is possible to parallelize ALS by spreading the matrices W^T and H^T across several threads. Whenever several matrices are changed at the same time, synchronization is required [23]. Iteration 4 of the sequential ALS algorithm illustrates the changes made. There is also a parallelization method for ALS that makes use of the CUDA GPU. Data is transferred to the GPU, and the host is in charge of ensuring that everything is in sync. When the calculation in the GPU is completed, the variables W^T and H^T are transferred from the device to the host computer(Jallouli et al., 2017).

GPU ALS(Rodrigues et al., 2015):

Input: A,W,H λ , T

- 1 Allocate GPU memories to matrices A, W & H
- 2 Copy matrices A, W and H to the GPU from host
- 3 begin:
- 4 Initialize ($H \leftarrow$ (small random numbers))
- 5 for iter $\leftarrow$ 1 to T Step=1 do
- 6 Compute the W using equation:

$$w_i^* = (H_{\Omega i}^T H_{\Omega i} + \lambda I)^{-1} H^T a_i$$

- 7 Compute the H using equation:

$$h_j^* = (W_{\Omega j}^T W_{\Omega j} + \lambda I)^{-1} W^T a_j$$

- 8 Copy matrices W and H to host from GPU

Diversity:

In order to provide the most accurate results possible, algorithms that make suggestions based on similarities between persons or objects are used. In as a result, users will be exposed to a more limited range of goods, with the vast majority of them being well-known items. Specialty things that are very important to the topic may, nevertheless, be missed. With the variety of suggestions, it is possible for consumers to discover things that they would not have discovered on their own otherwise. Machine learning (ML) feature extraction in particular is one of the most essential and time-consuming parts of constructing machine learning (ML) models, and it consumes a considerable amount of computational resources (or engineering). Everyone has heard it before: "It's nothing new." According to some estimates, the preparation of data accounts for 80 percent of the work performed by a data analyst.

In order to determine whether or not the feature extraction phase was successful, each application has its own set of criteria. As part of the feature extraction phase, descriptions of various things are obtained from their related objects, which is referred to as the feature retrieval stage. When combined with the preceding discussion, it was proposed that the idea of a feature store be used to address this time-consuming issue in the machine learning life cycle. This was accepted. An interface between feature engineering and model development, as well as a centralized data warehouse of features, is provided by a feature store, which enables data science teams to work on two problems at the same time while using the same technological platform.

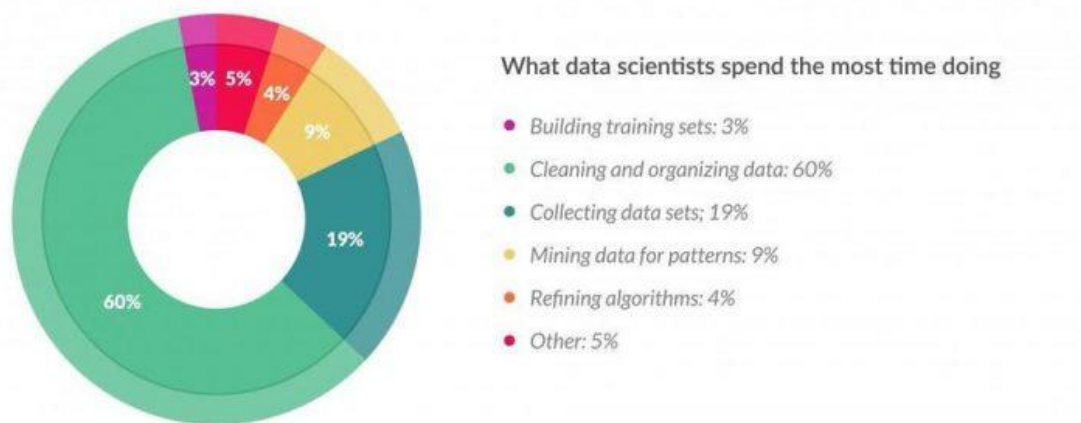


Fig 1. Feature Engineering as one of the important task for data scientists(Sarih et al., 2019)

Data science teams can save both time and money by addressing two issues at the same time. By tackling two problems at the same time, they can save both time and money. In DS/ML, a massive quantity of data can be stored and retrieved much more quickly than was previously possible. Because of this, machine learning models can be reused, experimented with, and then packaged considerably more rapidly than was previously feasible. The low latency and high throughput supplied by this software also allows it to provide capabilities to other programs or databases, enabling them to function more efficiently as a consequence of the low latency and high throughput provided by the software(Möller et al., 2018).

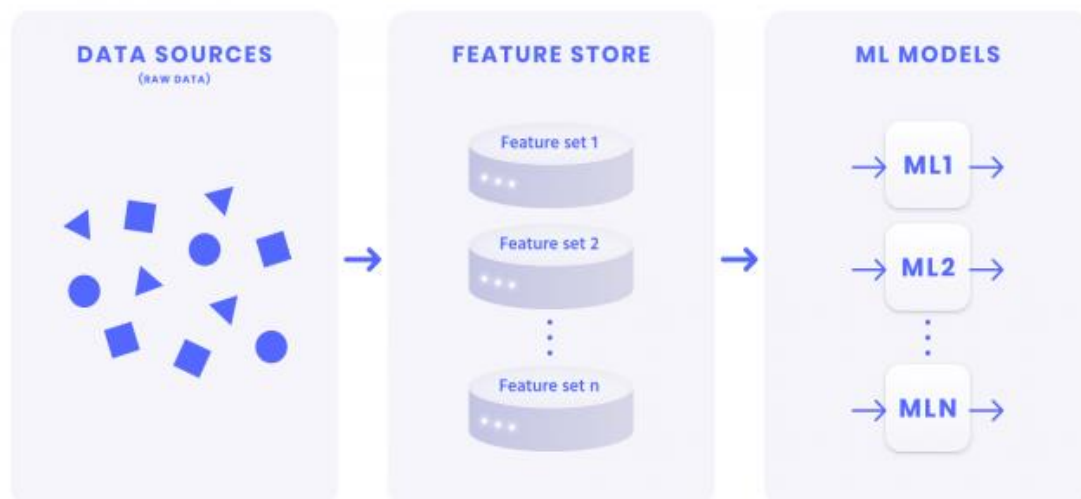


Fig 2. Using Feature Engineering to solve Diversity issues (*The Anatomy of High-Performance Recommender Systems – Part 1 _ Algolia Blog, n.d.*)

In contrast, when data scientists and machine learning experts are working on several machine learning projects at the same time, manual feature engineering may result in duplicate recommender systems in the recommender systems that are created by these engineers. Thus, a feature store is first and foremost a centralized repository for all of the features that are currently accessible on the market.. This catalogue may be used by any data scientist who is starting on a new project in order to quickly identify the characteristics that they will need for that project. Rather than just serving as a data layer or a data transformation service, feature stores are data transformation services that allow customers to modify raw data and store it as features that may be utilized by any machine learning model in the future(Resnick & Varian, 1997).

Proposed Framework:

We receive item and user information as input in our suggested framework and prune data to find the most essential data from a huge collection of data, as we know from experience that the most important data is located within the top 20% of all data. We are able to solve one of the most difficult elements of the recommendation model, namely data sparsity, by doing so. The next step is to build a community from user and item data so that whenever a new user or item is introduced, we can calculate their similarity measure and map them to existing communities to resolve the cold start issue and recommend new items to users as well as new users various items that were previously recommended to similar users. One of the most important issues is diversity; every item/user has a variety of features that can be extremely diverse; as a result, we must perform feature selection and select only the most important features; this can be done using a variety of methodologies, including auto encoders, PCA, and others. Finally, we may apply a number of machine learning algorithms suitable for the scenario at hand to see which approach gives the best recommendation for our model.

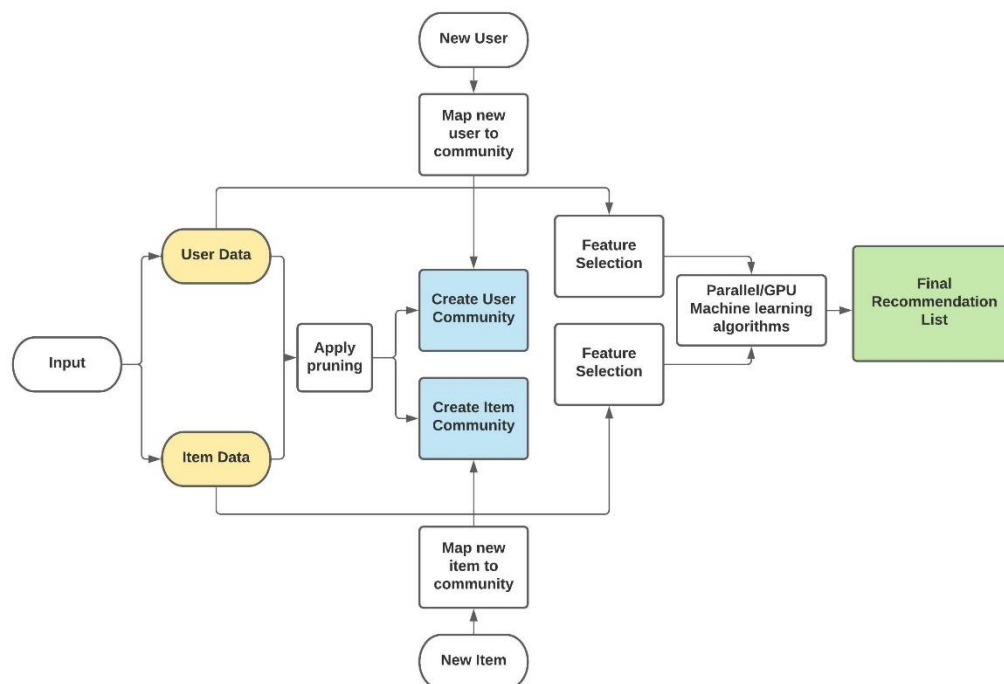


Fig 3. Proposed framework

Conclusion:

This paper concentrates on different technological solutions to address key problems encountered by recommender systems, it also briefly describes how they function and what their limits are in a straightforward manner. The major problems which we discussed here is

- cold start which we suggested here to be resolved by community detection problem.
- Diversity issues can be resolved by either identifying important features or using various methods to merge features using various methods like auto encoders, PCA etc.
- Data sparsity issues which can be resolved by identifying top 20% of the data as we are suggesting Pareto principle can be used here to quote that this 20% data is going to work for rest of 80%
- Scalability issues can be resolved by using parallel algorithms or GPU models.

Reference:

- [1] Abdollahpouri, H., Burke, R., & Malthouse, E. (2021). *User-centered Evaluation of Popularity Bias in Recommender Systems*. 119–129.
- [2] Albert, R., & Barabási, A. L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, 74(1), 47–97. <https://doi.org/10.1103/RevModPhys.74.47>
- [3] Jallouli, M., Lajmi, S., & Amous, I. (2017). Designing Recommender System: Conceptual Framework and Practical Implementation. *Procedia Computer Science*, 112, 1701–1710. <https://doi.org/10.1016/j.procs.2017.08.195>
- [4] Jang, S. W., Kim, S., & Ha, J. (n.d.). *Graph-based Recommendation Systems : Comparison Analysis between Traditional Clustering Techniques and Neural Embedding*. 1–11.
- [5] Möller, J., Trilling, D., Helberger, N., & Es, B. Van. (2018). *Do not blame it on the algorithm : an empirical assessment of multiple recommender systems and their impact on content diversity*. 4462. <https://doi.org/10.1080/1369118X.2018.1444076>
- [6] Palla, G., Derényi, I., Farkas, I., & Vicsek, T. (2005). Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043), 814–818. <https://doi.org/10.1038/nature03607>
- [7] Pitsilis, G., & Knapkog, S. J. (2009). Social trust as a solution to address sparsity-inherent problems of recommender systems. *CEUR Workshop Proceedings*, 532, 33–40.

- [8] Rani, A. (2014). *Tag based Recommender System using Pareto Principle*. June.
- [9] Resnick, P., & Varian, H. R. (1997). Recommender Systems. In *Communications of the ACM* (Vol. 40, Issue 3). <https://doi.org/10.1145/245108.245121>
- [10] Rodrigues, A. V., Jorge, A., & Dutra, I. (2015). Accelerating recommender systems using GPUs. *Proceedings of the ACM Symposium on Applied Computing*, 13-17-April, 879–884. <https://doi.org/10.1145/2695664.2695850>
- [11] Sardanios, C., Tsirakis, N., & Varlamis, I. (n.d.). *A Survey on the Scalability of Recommender Systems for Social Networks*. 89–110.
- [12] Sarih, H., Tchangani, A. P., Medjaher, K., & Pere, E. (2019). Data preparation and preprocessing for broadcast systems monitoring in PHM framework. *2019 6th International Conference on Control, Decision and Information Technologies, CoDIT 2019, Cm*, 1444–1449. <https://doi.org/10.1109/CoDIT.2019.8820370>
- [13] Schifferer, B., Titericz, G., Deotte, C., & Pereira, G. D. S. (2020). *GPU Accelerated Feature Engineering and Training for Recommender Systems*. 16–23.
- [14] *The anatomy of high-performance recommender systems – Part 1 _ Algolia Blog*. (n.d.).
- [15] Yu, M., Zhang, X., Lee, D., & Kreager, D. (2018). *Community-Based Recommendation for Cold-Start Problem: A Case Study of Reciprocal Online Dating Recommendation*. 201–222. https://doi.org/10.1007/978-3-319-78196-9_10

Protection of Transformer using Accelerated Convolution Neural Network

Harish Balaga¹, Marrapu Deepthi², Venkata Sravani A³

^{1,3} Vardhaman College of Engineering, Hyderabad

² Chaitanya Bharathi Institute of Technology, Hyderabad

harish.balaga@vardhaman.org

Abstract. An improved method of transformer inrush currents discriminating from internal fault is tried in this work by considering differential current signal as an image instead of a waveform. This is done with the help of an accelerated convolutional neural network (CNN). This method eliminates the necessity of a specialized feature extractor before feeding the data to a decision-maker. The signal feature extraction & fault detection are fused into a single deep neural network (DNN). The designed deep neural network is tested with multiple cases of operating conditions of the transformer for speed and accuracy as well.

Keywords: Accelerated CNN, Convolutional neural network (CNN), Deep learning, Differential protection, Inrush current, Internal fault, Power transformers.

1. Introduction

Power transformers are key parts of electric power system and their protection from internal grounding faults is of immense importance. As per CIGRE technical report, more than 1/10th of the power transformer failures are due to internal winding faults [1]. Differential current scheme is the most popular method for transformer protection. But inrush current issue is the major hurdle in this method [2], [3]. Many advanced methods were proposed in the last three decades to address this issue. Some of them include signal preprocessing methods in which wavelet transforms (WT) are used [4]–[6] as feature extractors. However, these methods have high computational burden and they are sensitive to noise. Some, like [7], are model-based methods, which use estimation techniques like Kalman filters. The main issue with these methods is they cannot represent the precise physical behavior of an electromagnetic devices [8]. All the aforementioned issues were addressed to some extent in the recent research works based on AI and ML. These methods are independent of mathematical models. In recent times, design and development of ANN based intelligent protective relays has become a growing interest among various researchers. ANN applies the knowledge acquired during training process and responds when acted upon a new event in an excellent manner. Similar works using wavelets along with ANNs are presented in [4], [9], [10]. An approach based on decision tree is presented in [11], with processing time of 1 cycle and 97.7% accuracy. Some of the other recent works related to use of ANN for the protection of transformer include [12]–[17]. A method based on SVM with GA for fault diagnosis of a power transformer is proposed in [18]. All these approaches are based on the pattern recognition technique. They have considered various other problems and operating conditions in transformer relaying like current transformer (CT) saturation. The modifications of algorithms that these articles discuss include the use of RBF network, probabilistic neural network, PSO trained ANNs etc. Similar pattern recognition based differential protection schemes can also be implemented for generator protection also [19], [20].

Few more recent works based on higher order statistics(HOS) [21] and Empirical Fourier transform(EFT) [22] seems promising. however, the HOS method treated the incipient faults created post-clearance of turn-to-turn faults as inrush.

However, all these intelligent methods cannot extract time-domain features form inputs before applying them to classifiers. Thus, it is necessary to have a classifier without any specialized feature extractor. Convolutional neural network (CNN) is mostly used in image processing applications. It is a type of Deep learning Neural Network, which is specialist in feature based classification applications [23]. However, CNNs have high computational complexity [24].

In this work, an accelerated CNN is designed to speed up the basic CNN and the need for expert feature extractor is eliminated.

2. Simulation of Differential Protection

2.1 Simulated network

The MATLAB/SIMULINK platform is used to simulate a power system network with a transformer of a rating 150MVA, 11/33kV. (see Fig.). In all operating conditions, the samples are divided into training and testing sets at 70:30 ratio randomly.

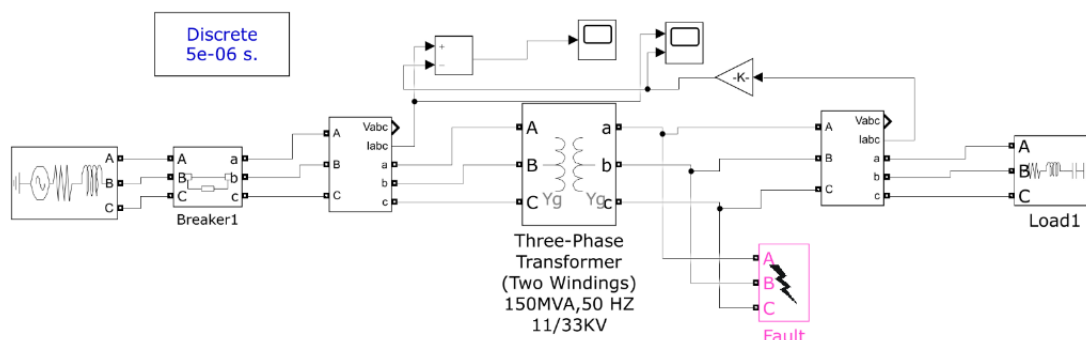


Fig.1. Simulated power system model

2.2 Input Data Preparation

To prepare the required training and testing database for the CNN, the simulated differential current signals are first sampled at 200 samples per cycle and only half cycle data (100 samples) is fed to the CNN at a time after normalizing as follows:

$$I_{D_{norm}} = \frac{I_D}{\text{Max}(|I_D|)}$$

Where I_D is the instantaneous differential current given by $I_{primary} - I_{secondary}$

More sets of such training data is created by sliding window approach in order to consider all possible fault inception angles in the training/testing data. This normalized data is revealed to an accelerated CNN, which outputs the Y, whose value is 1 for fault and 0 otherwise.

A total of 2432 sets of data are generated for various operating conditions of the transformer. The differential currents are simulated for various internal fault types, with different source impedances and fault inception angles. Similarly, the inrush currents are simulated at different switching angles.

3. Accelerated CNN

Regular neural networks utilize conventional activation functions like sigmoid, which require computing an exponent. They increase both training and testing times significantly. ReLU activation function and batch normalization are used instead to prevent overfitting and the usual vanishing gradient problem. To further reduce the computational burden, a faster and tiny version of the CNN is presented in this article. The accelerated CNN minimizes the error and decomposes the parameters of individual layers, as in [24], by quantizing the weight matrices.

3.1 Overall Structure

The CNN requires input to be an image of one channel or more with a consistent size. So, the normalized input data (half cycle or 100 samples) are first converted from 1-D signal form to 2-D signal of size 10×10 , and then converted to a $10 \times 10 \times 1$ size image as shown in Fig.. The signal then applied with 100 filters of 5×6 kernel size after passing through the accelerated layers. The signal flattens to a 5×100 signal and then passes through 3 fully connected layers as in Fig..

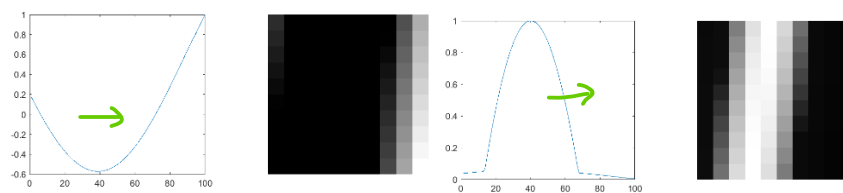


Fig.2. Sample waveforms and their normalized 2D vectors

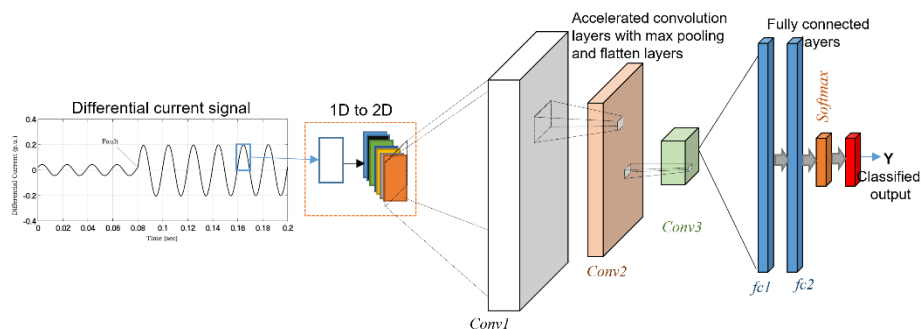


Fig.3. Procedure of the proposed scheme

In the last fully connected layer, final outout is generated. The differential relay issues a trip signal, only if the output is 1. Observe that, except the *softmax* activation function in the last FCN layer, *ReLU* is used for activation at all other layers. No padding is used, and the stride is 1.

4. RESULTS

Two cases are discussed in this section to show the advantages of accelerated CNN over regular CNN. Same sets of data are used in both cases. The indexes used to evaluate the results are:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100$$

$$\text{Dependability} = \frac{TN}{TN + FP} \times 100$$

Here, *TN* is number of faults detected, *FP* is the faults left undetected, *TP* is the number of healthy conditions detected, and *FN* is the number of healthy conditions those issued false alarm.

Table 21. Performance of the proposed accelerated CNN

Cases	LG Fault	LL Fault	LLL Fault	P-S Fault	Inrush	External Fault	Over Excitation
ACC%	99.76	99.15	99.40	99.28	100	100	100
DEP%	99.23	98.46	99.23	98.46	100	100	100

The numerical results of the proposed method in various operating conditions are shown in **Table 21**. One can observe here that all the indexes achieve around 99% fault detection, with 100% accuracy for the inrush current and external faults. Not much variation is observed in the overexcitation conditions, and it performs pretty accurately both with simple CNN and accelerated CNN as well.

5. CONCLUSION

In this paper, to improve the speed of the convolution layer and FCN layers an accelerated CNN is designed based on product quantization technique. This kind of CNN's speed is around four times compared to basic CNN. Accuracy of the accelerated CNN is increased by 1%. One more advantage of this protection technique is that, if the CNN structure is set once, it can be applied to any rated transformer. This accelerated CNN based protection method is validated using simulation and experimental study is yet to be done. The results thus obtained are compared with other methods and the superiority of the proposed method is seen in the previous section. The main drawback in the proposed method is its requirement of a very large database which is very difficult to obtain, so we use artificial methods for the generation of data. Generative adversarial neural networks may be used for this purpose in future works.

Acknowledgement

This research work was carried out at EEE department, Vardhaman College of Engineering, Hyderabad, India. The authors would like to thank the management and faculty for their constant support throughout.

REFERENCES

- [1] S. Tenbohlen et al., "Transformer Reliability Survey: Interim Report," 2012. Accessed: Oct. 16, 2021. [Online]. Available: https://e-cigre.org/publication/WGR_261_1-transformer-reliability-survey-interim-report.
- [2] S. Bagheri, Z. Moravej, and G. B. Gharehpetian, "Classification and discrimination among winding mechanical defects, internal and external electrical faults, and inrush current of transformer," *IEEE Trans. Ind. Informatics*, vol. 14, no. 2, pp. 484–493, Feb. 2018, doi: 10.1109/TII.2017.2720691.
- [3] A. Ashrafian, M. Mirsalim, and M. A. S. Masoum, "Application of a Recursive Phasor Estimation Method for Adaptive Fault Component Based Differential Protection of Power Transformers," *IEEE Trans. Ind. Informatics*, vol. 13, no. 3, pp. 1381–1392, Jun. 2017, doi: 10.1109/TII.2016.2620441.
- [4] J. P. Marques, C. Lazaro, A. P. Morais, and G. Cardoso, "A reliable setting-free technique for power transformer protection based on wavelet transform," *Electr. Power Syst. Res.*, vol. 162, pp. 161–168, Sep. 2018, doi: 10.1016/j.epsr.2018.05.002.
- [5] R. A. Ghunem, R. El-Shatshat, and O. Ozgonenel, "A novel selection algorithm of a wavelet-based transformer differential current features," *IEEE Trans. Power Deliv.*, vol. 29, no. 3, 2014, doi: 10.1109/TPWRD.2013.2293976.
- [6] R. P. Medeiros, F. B. Costa, and K. M. Silva, "Power Transformer Differential Protection Using the Boundary Discrete Wavelet Transform," *IEEE Trans. Power Deliv.*, vol. 31, no. 5, pp. 2083–2095, Oct. 2016, doi: 10.1109/TPWRD.2015.2513778.
- [7] F. Naseri, Z. Kazemi, M. M. Arefi, and E. Farjah, "Fast Discrimination of Transformer Magnetizing Current from Internal Faults: An Extended Kalman Filter-Based Approach," *IEEE Trans. Power Deliv.*, vol. 33, no. 1, pp. 110–118, Feb. 2018, doi: 10.1109/TPWRD.2017.2695568.
- [8] S. Jazebi et al., "Duality derived transformer models for low-frequency electromagnetic transients - Part I: Topological models," *IEEE Trans. Power Deliv.*, vol. 31, no. 5, pp. 2410–2419, Oct. 2016, doi: 10.1109/TPWRD.2016.2517327.
- [9] A. Sahebi, H. Samet, and T. Ghanbari, "Identifying internal fault from magnetizing conditions in power transformer using the cascaded implementation of wavelet transform and empirical mode decomposition," *Int. Trans. Electr. Energy Syst.*, vol. 28, no. 2, 2018, doi: 10.1002/etep.2485.
- [10] M. O. Oliveira, A. S. Bretas, and G. D. Ferreira, "Adaptive differential protection of three-phase power transformers based on transient signal analysis," *Int. J. Electr. Power Energy Syst.*, vol. 57, pp. 366–374, May 2014, doi: 10.1016/j.ijepes.2013.12.013.
- [11] S. R. Samantaray and P. K. Dash, "Decision Tree based discrimination between inrush currents and internal faults in power transformer," *Int. J. Electr. Power Energy Syst.*, vol. 33, no. 4, pp. 1043–1048, May 2011, doi: 10.1016/j.ijepes.2011.01.021.
- [12] H. Balaga, N. Gupta, and D. N. Vishwakarma, "GA trained parallel hidden layered ANN based differential protection of three phase power transformer," *Int. J. Electr. Power Energy Syst.*, vol. 67, pp. 286–297, May 2015, doi: 10.1016/j.ijepes.2014.11.028.
- [13] H. Balaga, D. N. Vishwakarma, and A. Sinha, "Numerical differential protection of power transformer using ANN as a pattern classifier," in *IEEE International Conference on Power, Control and Embedded Systems*, Nov. 2010, pp. 1–6, doi: 10.1109/ICPCES.2010.5698716.
- [14] D. N. Vishwakarma, H. Balaga, and H. Nath, "Application of genetic algorithm trained master-slave Neural Network for differential protection of power transformer," in *9th IEEE International Conference on Computer Engineering & Systems (ICCES)*, Dec. 2014, pp. 164–169, doi: 10.1109/ICCES.2014.7030950.
- [15] H. Balaga, D. N. Vishwakarma, and H. Nath, "Artificial Neural Network Based Backup Differential Protection of Generator-Transformer Unit," *Int. J. Electron. Electr. Eng.*, vol. 3, no. 6, pp. 482–487, 2015, doi: 10.12720/ijeee.3.6.482-487.

- [16] N. Gupta, H. Balaga, and D. N. Vishwakarma, "Numerical Differential Protection of Power Transformer using GA Trained ANN," *Int. J. Emerg. Trends Electr. Electron. (IJETEE)*, vol. 3, no. 1, pp. 1–9, 2013, Accessed: Oct. 06, 2014. [Online]. Available: http://iret.co.in/Docs/Volume_3/Issue1/1. Numerical Differential Protection of Power Transformer using GA Trained ANN.pdf.
- [17] H. Balaga, D. N. Vishwakarma, and A. Sinha, "Application of ANN based pattern recognition technique for the protection of 3-phase power transformer," in *Swarm, Evolutionary, and Memetic Computing*, vol. 7076 LNCS, B. K. Panigrahi, P. N. Suganthan, S. Das, and S. C. Satapathy, Eds. Visakhapatnam: Springer Berlin Heidelberg, 2011, pp. 358–365.
- [18] S. Fei and X. Zhang, "Fault diagnosis of power transformer based on support vector machine with genetic algorithm," *Expert Syst. Appl.*, vol. 36, no. 8, pp. 11352–11357, Oct. 2009, doi: 10.1016/j.eswa.2009.03.022.
- [19] M. Zielichowski and T. Szlezak, "A new digital ground-fault protection system for generator–transformer unit," *Electr. Power Syst. Res.*, vol. 77, no. 10, pp. 1323–1328, Aug. 2007, doi: 10.1016/j.epsr.2006.10.001.
- [20] M. Khan and M. A. Rahman, "Development and Implementation of a Novel Fault Diagnostic and Protection Technique for IPM Motor Drives," *IEEE Trans. Ind. Electron.*, vol. 56, no. 1, pp. 85–92, Jan. 2009, doi: 10.1109/TIE.2008.2004662.
- [21] L. L. Zhang, Q. H. Wu, T. Y. Ji, and A. Q. Zhang, "Identification of inrush currents in power transformers based on higher-order statistics," *Electr. Power Syst. Res.*, vol. 146, pp. 161–169, 2017, doi: 10.1016/j.epsr.2017.01.029.
- [22] S. K. Murugan, S. P. Simon, K. Sundareswaran, P. S. R. Nayak, and N. P. Padhy, "An Empirical Fourier Transform-Based Power Transformer Differential Protection," *IEEE Trans. Power Deliv.*, vol. 32, no. 1, pp. 209–218, Feb. 2017, doi: 10.1109/TPWRD.2016.2575981.
- [23] N. Liu, L. Wan, Y. Zhang, T. Zhou, H. Huo, and T. Fang, "Exploiting Convolutional Neural Networks With Deeply Local Description for Remote Sensing Image Classification," *IEEE Access*, vol. 6, pp. 11215–11227, Jan. 2018, doi: 10.1109/ACCESS.2018.2798799.
- [24] J. Cheng, J. Wu, C. Leng, Y. Wang, and Q. Hu, "Quantized CNN: A Unified Approach to Accelerate and Compress Convolutional Networks," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 29, no. 10, pp. 4730–4743, Oct. 2018, doi: 10.1109/TNNLS.2017.2774288.

Performance and Comparison of Various Machine Learning Classifier for Sentiment Analysis on Amazon Dataset

* S.Madhusudhanan¹, Jyothi N.M², M.Ramesh³, N.Arun Vignesh⁴

Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation,
Vaddeswaram, Guntur (DT), Andhrapradesh,India
¹*ssmadhu80@gmail.com

Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation,
Vaddeswaram, Guntur (DT), Andhrapradesh,India
²jyothiarunkr@gmail.com

Department of Computer Science and Engineering, Koneru Lakshmaiah Education Foundation,
Vaddeswaram, Guntur (DT), Andhrapradesh,India
³rameshcsit24@kluniversity.in

Department of Electronics and Communication Engineering, Gokaraju Rangaraju Institute of Engineering
and Technology, Hyderabad
⁴arunvignesh44@gmail.com

Abstract. The advent of web 2.0 provides an ideal platform for individuals to share their ideas, opinions, and feelings. Web Opinion Mining / Sentiment Analysis is a text mining task that aims to build a system that automatically extracts, identifies, and categorizes user opinions from natural language text, user created material, or user generated media. In this work, performance of different classifiers was analyzed and compared. SVM with 96% and Logistic Regression with 94% on Electrnics review dataset, were outpermed well with best classification accuracy compared with all other classifiers.

Keywords: Machine Learning; Sentiment Analysis; Classification; Bayesian; Support Vector Machine; Random Forest;

1 Introduction

In Data Science, Artificial intelligence (AI) has become more popular and focused on recent research work. Machine Learning (ML) is a part of AI for various applications to build a classification model [9]. During the training phase, Features and patterns are identified and classifiers are trained based on the features set. The Classifier predicted the test by using training data [13]. ML is used in day-to-day activities. E.g., the railway reservation system uses classification algorithm to predict seat confirmation probability, product recommendations in the e-commerce website are based on purchase history [11][12]. Prediction of words in a Google search engine, diagnosis based on medical image analysis, dynamic fare in airlines, and so on. In ML, mathematical model like probability and statistics are used to predict the patterns, facts, associations, and trends. In the decision-making process, to recognize unseen patterns ML techniques are used. Classification is task which is a part of ML. It is a process to build a model that separates test data label based on the training phase [11][7][8][10]. Sentiment analysis is one of the applications of text mining that purpose to detect opinions, emotions, and perspectives in the given text. Sentiment Analysis requires considerably more knowledge of the natural language than text examination and subjective analysis [20]. The previous algorithm considers the only frequent occurrence of the words in a document but fails to identify the target opinion especially when messages are short. It is the process of identifying subjective information as positive or negative. It helps the business organization and individual to make the decision. Recently a lot of research works focus on the automatic identification of sentiment by using machine learning algorithms.[21][22].

In this paper, five popular classification techniques namely Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF), Naive Bayes (NB), Decision Tree (DT), and K Nearest Neighbors (KNN) are analysed and compared its performance in sentiment analysis.

2 Related Work

The authors proposed feature selection method called unigram and bigram with χ^2 (Chi-Square) and Singular Value Decomposition for dimensionality reduction with four types of scaling methods and machine learning classifier is used with accuracy rate of 84.14%. [21].

The author performed sentiment classification on product review dataset using SVM and Naïve Bayes Classification methods by constructing Bag-of-Word Model[13].

Abbasi et.al [1] proposed hybrid method which combines information gain and entropy weighted genetic algorithm (EWGA) for feature selection with SVM classification. The proposed method would product better accuracy of 91% for movie review data set.

The problem of constructing a domain-dependent sentiment lexicon with supervised learning method was explored for sentiment classification is implemented by the authors [23]. They utilized unigram and bigram feature selection techniques with SVM classification algorithm and achieved 88% accuracy.

To improve the performance of sentiment classification, the aauthor [9] reevaluates objective sentiment words in SentiWordNet. They used Term frequency with SVM algorithm to analyse the performance.They achieved accuracy of 77.9% .

3 Method

3.1 Data set

Sentiment analysis uses many opinion datasets for classification. The popular movie review dataset which is available online is (<http://www.cs.cornell.edu/People/pabo/movie-review-data>) [4][5][14]. Data repositories such as Kaggle, Data.world, UCI, ASU, Open Knowledge Labs, Kdnuggets, and SNAP.

3.2 Classificati`on Algorithms

3.2.1 Logistic Regression (LR)

The most popular supervised algorithm is used to envisage the probability of the target variable. The logistic function is used to calculate likelihood probability for independent variable. This function is called relation function since it relates changes in the independent variables to a higher (or lower) probability of occurrence of the event by the dependent variable [22] [8].

3.2.2 Naive Bayes Classifier (NB)

This method uses the Bayes theorem to calculate the most frequent class for the new instance. It can handle some features or classes [5][6][16][22]. The Bayes hypothesis is a method to calculate difference between likelihoods $P(a|b)$ from $P(a)$, $P(b)$ and $P(b|a)$. likelihoods $P(a|b)$ from $P(a)$, $P(b)$ and $P(b|a)$. The formula to calculate probability is as follows :

$$P\left(\frac{a}{b}\right) = \frac{[P\left(\frac{b}{a}\right)*P(a)]}{P(b)} \quad (1)$$

Where $P\left(\frac{a}{b}\right)$ is posterior probability, $P\left(\frac{b}{a}\right)$ is conditional probability of b given a known as likelihood probability, $P(a)$ is a prior probability of a.

3.2.3 Support Vector Machine (SVM)

SVM examines data, describes decision boundary points, and uses the segments for the estimation, which are acted in the input space. In the ML algorithm, SVM outperforms well for classification [22]. The essential data is converted into two vectors of size m. As of now, every datum (passed on as a vector) is referenced into a class. By then, the machine perceives the separation between two labels which is far from any space on the trained data. Support Vectors are essentially the coordinates of individual perception.[24][13][17]

3.2.4 Decision Tree (DT)

The decision tree works by creating a binary tree having a condition on internal nodes and class labels on the leaf nodes.[21] First Information Gain (IG) and entropy are calculated on each attribute, the attribute with maximum IG is selected as a split node. Now the new attributes are labeled and calculate IG and entropy for each node repeatedly to form the internode with the leaf node. Leaf node contains the majority label [15][18].

3.2.5 Random Forest (RF)

RF classifier is based collection of tree-structured classifier [2][19]. Many Decision trees predict decision on dataset and final decision is done by majority voting process. Classifier predicts the result on the data set. The working principle of RF are discussed below:

Step-1: Choose random T input points.

Step-2: Build the decision tree for the input points.

Step-3: Forecast the decision which comprise two or more Decision trees.

Step-4: Repeat Step 1 & 2.

Step-5: Predictions of each decision tree are involved in the testing data; RF chosen the final decision based on majority voting process.

3.2.6 K Nearest Neighbors (KNN)

The K-Nearest-Neighbors (KNN) is a straightforward and powerful classification method. In this method, data are arranged by a dominant part vote of its Neighbors. It acknowledges the new data similarity and is assigned to the available similar category. For classification of test data, it retrieves the stored information, based on that it assigns the class to test data.[3] It is also known as lazy learners, because of the storing the data and retrieves it for classification. [20]

4 Experiments

In this work, datasets are collected from the repository [3][4]. The Python software is utilized for implementation. Confusion matrixes are generated to find the correctly classified class. The same environment variable is used for all algorithms.

4.1 Dataset-I: Clothing, Shoes, and Jewelry [14]. Rating 5 and 4 are considered as positive reviews and 1 and 2 are considered as negative reviews. Rating 3 is neglected. 25000 reviews of positive and negative are selected randomly and make it balanced. 80:20 for training and testing the dataset. Similar procedure is followed for all data set. Table 1 shows Dataset Statistics. The performance of various classifiers is shown in

Table 1: Statistics of Dataset			
Dataset	Total Reviess	Positive Review	Negative Review
Clothing, Shoes and Jewelry	278677	221597	26655
Electronics	308277	230674	77603
Health and Personal Care	346355	279801	33300
Sports and Outdoors	296337	253017	19249
Toys and Games	167597	140235	11005

Table2: Experiment Result of Clothing, Shoes and Jewellery					
Algorithm	Classification Accuracy (in %)	TP Rate (in%)	FP Rate (in %)	Precision	Recall
Logistic Regression	88	44.23	44.12	0.88	0.88
SVM	87.88	44.06	43.82	0.88	0.88

Random Forest	82	40.25	41.84	0.81	0.84
Naive Bayes	74	29.35	43.20	0.84	0.58
Decision Tree	72.43	30.94	41.49	0.79	0.62
KNN	50.54	18.97	31.57	0.51	0.38

In above Table2, Logistic Regression classifier outperforms maximum accuracy.

4.2 Dataset-II Amazon Electronics Product Review dataset [14]. The performance of various classifiers is shown in Table3.

Table3: Experiment Result of Electronics

Algorithm	Classification Accuracy (in %)	TP Rate (in%)	FP Rate (in %)	Precision	Recall
Logistic Regression	94	47.58	46.74	.96	.93
SVM	96	47.89	47.72	.96	.95
Random Forest	92	47.80	43.86	.89	.95
Naive Bayes	79	47.12	37.70	.72	.94
Decision Tree	73	25.21	47.58	.92	.50
KNN	84	45.07	39.20	.81	.90

In above table3, SVM classifier outperforms maximum accuracy.

4.3 Dataset-III Amazon Health and Personal Care Review dataset [14]. The performance of various classifiers is shown in Table4.

Table4: Experiment Result of Health and Personal Care

Algorithm	Classification Accuracy (in %)	TP Rate (in%)	FP Rate (in %)	Precision	Recall
Logistic Regression	84	42.13	41.98	.84	.84
SVM	84	41.92	41.81	.84	.83
Random Forest	74	37.00	37.40	.75	.74
Naive Bayes	66	24.58	41.34	.75	.49
Decision Tree	60	14.30	45.93	.79	.28
KNN	50	22.95	27.13	.50	.46

In above Table4, SVM and Logistic Regression classifiers outperforms maximum accuracy.

4.4 Dataset-IV Amazon Sports and Outdoors Review dataset [14]. The performance of various classifiers is shown in Table5.

Table5: Experiment Result of Sports and Outdoors

Algorithm	Classification Accuracy (in %)	TP Rate (in%)	FP Rate (in %)	Precision	Recall
Logistic Regression	85	43.35	41.67	.85	.85
SVM	84.8	42.95	41.83	.85	.85
Random Forest	76.3	38.02	38.28	.78	.75
Naive Bayes	67	24.63	42.22	.78	.48
Decision Tree	62	31.37	35.90	.70	.62
KNN	50	22.80	27.55	.51	.45

In above Table5, Logistic Regression classifiers outperforms maximum accuracy.

3.5 Dataset-V Amazon Toys and Games Review dataset [14]. The performance of various classifiers is shown in Table6.

Table6: Experiment Result of Toys and Games

Algorithm	Classification Accuracy (in %)	TP Rate (in%)	FP Rate (in %)	Precision	Recall
Logistic Regression	88	45.02	43.11	.88	.88

SVM	88	44.82	42.89	.88	.88
Random Forest	79	40.00	39.09	.80	.78
Naive Bayes	68	25.64	42.73	.80	.50
Decision Tree	68	28.70	39.61	.75	.56
KNN	49	12.70	36.02	.49	.25

In above Table4, SVM and Logistic Regression classifiers outperforms maximum and same accuracy. Evaluation and Comparison of different Classification algorithms are depicted in graph on Figure 1:

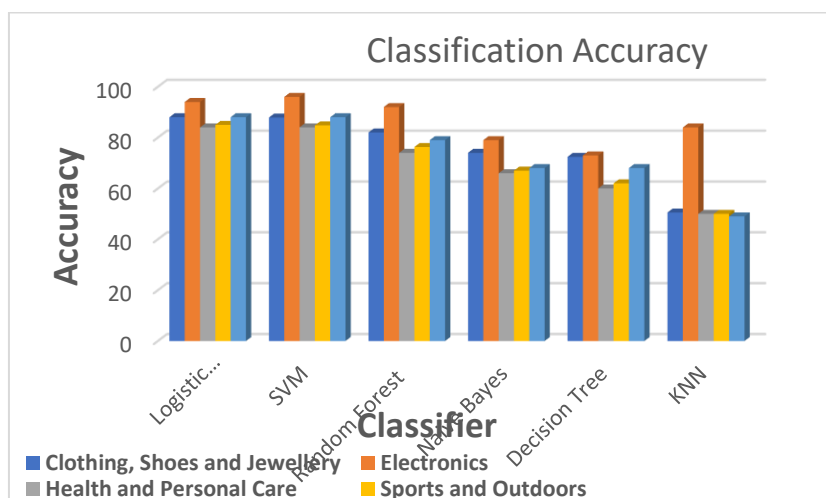


Fig1: Various Classifier with different dataset

5. Conclusion

In this work, the performance of various classifications with different datasets is evaluated and compared. Feature selection technique Term Frequency and Inverse Document Frequency (TF-IDF) is used to convert the dataset into frequency matrix for selecting the best feature. From the result we conclude that classifiers SVM and LR produced best classification accuracy compared with all other classifiers. Different feature selection techniques need to be implemented in the future.

Reference

- [1] Abbasi, H. Chen, and A. Salem, "Sentiment Analysis in Multiple Languages: Feature Selection for Opinion classification in Web Forums," ACM Trans. Information Systems, vol. 26, no. 3, pp. 1-34, (2008).
- [2] Anjali Devi, S., Sapkota, P., Rohit Kumar, K., Pooja, S., Sandeep, M.S., "Comparison of classification algorithms on twitter data using sentiment analysis", International Journal of Advanced Trends in Computer Science and Engineering, Vol. 9, Issue-5, (2020), pp:8170-8173
- [3] Bahrawi, "Sentiment Analysis using Random Forest Algorithm Online Social Media Based", Journal of Information Technology and its Utilization, Vol. 2, Issue 2, (2019), 29-33
- [4] Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," Proc. 42th Ann. Meeting on Assoc. for Computational Linguistics (ACL), (2004), pp. 271-278.
- [5] Bo Pang and Lillian Lee, "Opinion mining and sentiment analysis" Foundations and Trends in Information Retrieval, Vol. 2, No 1-2, (2008), pp 1-135
- [6] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques", Proceedings of EMNLP, (2002).
- [7] B. Liu, M. Hu, and J. Cheng, "Opinion Observer: Analyzing and Comparing Opinions on the Web," Proc. Int'l Conf. World Wide Web, (2005), pp. 342-351.
- [8] Bulusu, A., Sucharita, V. Research, "Research on machine learning techniques for POS tagging in NLP", International Journal of Recent Technology and Engineering, Vol. 8, Issue 1, Special Issue- 4, (2019).
- [9] Hung and H. Lin, "Using Objective Words in SentiWordNet to Improve Word-of-Mouth Sentiment Classification," in IEEE Intelligent Systems, vol. 28, no. 2, pp. 47-54, March-April (2013), doi: 10.1109/MIS.2013.1.
- [10] F. Alattar, K. Shaalan, "Survey on Opinion Reason Mining and Interpreting Sentiment Variations", IEEE Access, Volume 9, (2021).
- [11] Karthikeyan, C., Sahaya, A.N.A., Anandan, P., Prabha, R., Mohan, D., Vijendra, B.D., "Predicting Stock Prices Using Machine Learning Techniques", Proceedings of the 6th International Conference on Inventive Computation

- Technologies, ICICT 2021, .(2021).
- [12] Koyel Chakraborty;Siddhartha Bhattacharyya;Rajib Bag,“A Survey of Sentiment Analysis from Social Media Data”, IEEE Transactions on Computational Social Systems, Volume: 7, Issue: 2, (2020).
 - [13] Morinaga, S., Yamanishi, K., Tateishi, K. and Fukushima, T., (2002), “Mining product reputations on the web”, in Proceeding of the eighth ACM SIGKDD, international conference on Knowledge discovery and data mining (pp. 341-349). ACM.
 - [14] Ni, J., Li, J. & McAuley, J.,“Justifying recommendations using distantly-labeled reviews and fine-grained aspects”, In Proc. 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 188–197, (2019)
 - [15] Qixuan Hou; Meng Han; Zhipeng Cai,” Survey on data analysis in social media: A practical application aspect” Big Data Mining and Analytics , Volume: 3, Issue: 4, (2020).
 - [16] Sajana, T., Narasingarao, M. R.,“Classification of Imbalanced Malaria Disease Using Naïve Bayesian Algorithm”, International Journal of Engineering & Technology,7(2.7) ,786-790,(2018)
 - [17] S Sakhare, N.N., Sagar Imambi,“Performance analysis of regression-based machine learning techniques for prediction of stock market movement" International Journal of Recent Technology and Engineering 7 (6), 655-662, (2019)
 - [18] Sanjay Bhargav, P., Nagarjuna Reddy, G., Ravi Chand, R.V., Pujitha, K., Mathur, A., “Sentiment analysis for hotel rating using machine learning algorithms” International Journal of Innovative Technology and Exploring Engineering, Vol. 8,Issue.6,pp 1225-1228.
 - [19] Shaozhong Zhang; Haidong Zhong,“Mining Users Trust from E-Commerce Reviews Based on Sentiment Similarity Analysis”, IEEE Access ,Volume: 7, Page(s): 13523 – 13535, (2019).
 - [20] Surbhi Bhatia,“A Comparative Study of Opinion Summarization Techniques”, IEEE Transactions on Computational Social Systems, Vol. 8, No. 1, (2021).
 - [21] Shyamasundar L B., Jhansi Rani P, “A Multiple-Layer Machine Learning Architecture for Improved Accuracy in Sentiment Analysis”, The Computer Journal , Volume: 63, Issue: 1, Jan. 2020,pp 395 – 409,(2020).
 - [22] Vavilapalli, S.S., Reddykorepu, P., Saggam, S., Pentyala, M., Devi, S.A,” Summarizing Sentiment Analysis on Movie Critics Data”, Proceedings of the 6th International Conference on Inventive Computation Technologies, ICICT 2021, (2021).
 - [23] Yassine Al-Amrani,Mohamed Lazaar, Kamal Eddine Elkadiri,“Sentiment Analysis using supervised classification algorithms”, Proceedings of the 2nd international Conference on Big Data, Cloud and Applications, Association for Computing Machinery, Article No.: 61, Pages 1–8,(2017).
 - [24] You Li, Yuming Lin, Jingwei Zhang and Guoyong Cai, Constructing Domain-Dependent Sentiment Lexicons Automatically for Sentiment Analysis. Information Technology Journal, 12: 990-996, (2013).

The Perceived Benefits and Risks of Cloud Computing in Terms of Organizational Cost and Security: An Exploratory

Veshala Lokesh Kumar^{1*}, Veshala Likhitha², Spuritha M³

¹ Gokaraju Rangaraju Institute of Engineering & Technology, Department of Electronics and Communication Engineering, Bachupally, Hyderabad, India

² Ramakrishna Degree College, Department of Statistics, Korutla, Telangana, India

³ G.Narayanamma Institute of Technology and Science, Dept. of Information Technology, Hyderabad, Telangana, India

Abstract. Recently, cloud computing has received increasing popularity within the business sector, presenting important benefits for any business organizations that incorporate the technology into their information system. Business organizations that adopted cloud computing enjoyed enhanced data accessibility, improved quality of information, increased operational productivity, and effective management. However, several business organizations are still lagging behind in implementing and utilizing cloud computing for their business operation. This is not unconnected to the existing vagueness in terms of cost risks and security benefits associated with the novel technology. Although some of the perceived edges of cloud computing for businesses include cost effectiveness and the ability to access infrastructures and resources that would otherwise be inaccessible, there has not been any evidence to prove its cost risks. More so, the main source of concern, though, remains the security threats. To yet, little effort has been placed into examining the advantages and dangers of cloud computing in terms of price and security for businesses. As a result, the perceived benefits and dangers of cloud computing in terms of organizational cost and security were explored in this study report. The findings, which were based on empirical data collected through an interview and a well-structured questionnaire, revealed that cloud computing, particularly the hybrid cloud, is more cost-effective for small and medium-sized businesses, and that the cost and security risks associated with cloud computing cannot be eliminated without addressing the issue of inaccurate cost models, among other factors.

Keywords: Cloud Computing, IaaS, Hybrid Cloud, Community cloud service

1. Introduction

Just as its name depicts, cloud computing involves storing data or things in the “cloud” rather than on desktop computers or databases. The cloud is “a nebulous assemblage of computers and servers accessed via the Internet.” Since the introduction of cloud computing, the technology has helped in improving business operations and growth for small, medium, and big business organizations, ever since introduction, cloud computing has aided in the improvement of company operations and growth for small, medium, and large businesses, allowing them to remain competitive. Cloud computing has improved and automated the conventional manner of conducting business, allowing businesses to run their operations more effectively. Cloud computing is now one of the most recent, exciting, and widely discussed IT advancements in today's technological industry.

The way businesses utilize computers and the Internet has evolved dramatically thanks to cloud computing. This modification primarily affects how data/information is kept, software programs are installed and utilized, and cloud services are accessible. Once internet - connected, users may now quickly access all of their data, documents, and apps from anywhere and at any time. They can also cooperate more easily from different locations.

Organizations may simply conduct numerous activities from different places using cloud computing, as well as work from anywhere using the same sort of application. IaaS, SaaS, and PaaS are the three primary components of cloud computing. All of these aid in the reduction of an organization's administrative, procedural, computer, software, and hardware expenditures and expenses. As a result, companies will be able to save more money and spend more in their operations to automate their processes.

However, before determining whether or not to adopt cloud computing technology and employ it, corporate organisations must carefully weigh the benefits and dangers of cloud computing [2]. While one school of thought is primarily concerned with the economic lead of cloud computing, another is concerned with the security threats. Although several articles have been released to look into the cost and security implications of cloud computing for businesses, the results are mixed [3]. As a result, there is ambiguity about the potential benefits of cloud computing for businesses [4].

As a result, in attempt to address a well-known research gap in terms of cost risks and security benefits associated with the novel technology, this paper aims to investigate the perceived benefits and risks of cloud computing by responding to the following study questions about organisational cost and security:

- i. Do we have any financial dangers associated with cloud computing and what are the economic benefits of cloud computing for businesses?
- ii. Security concerns of cloud computing for organisations, and any advantages in terms of security?
- iii. Do various types of cloud computing services provide varying levels of perceived ease?
- iv. Does the size of a company have an impact on the perceived value of cloud computing?

As a result, the purpose of this research paper is to provide a comprehensive and in-depth knowledge of the impact on business organisations in terms of cost and security and also to explore all benefits and concerns in this field, and contribute to broader study topics.

2. Literature Review

To properly comprehend how cloud computing may impact corporate cost and security, you must first understand what cloud computing mean. The concept of cloud computing, its key properties, use and delivery methods, security trade-offs, and pricing hazards are all covered in this section.

A. What is Cloud Computing

It involves “a nebulous assemblage of computers and servers accessed via the Internet.” Once internet - connected, cloud computing allows businesses to access all of their information/data, documents, and apps remotely, as well as allowing customers to engage at any time and in any place.

Cloud computing is a sort of application as well as a platform. Information/data and application are saved in the cloud on servers that are most times a virtual machine instead of physical structure. Cloud computing platforms provision, manage, and reconfigure these sophisticated servers on which cloud applications and data are hosted and accessible through the Internet [5].

With cloud computing, business organizations can effectively and efficiently use and maximize their software and hardware investments. This allows them to overcome the separated systems' physical barriers and automate the administration of high number of machines as a single entity. Cloud computing may be defined as a virtual system that allows data centers to evolve naturally.

B. Characteristics of Cloud Computing

It contain a large spread of services through the internet, including networking, software, predictions, analytics, servers, storage, and more. These services are provided by cloud providers, who are third-party service providers. Businesses can engage with these providers and inform them of their requirements. These cloud providers will then rely on the information given by the business organization to provide to them their needed services or available services.

Business organisations who utilize cloud computing services pay per usage and do not have to pay any fees if they do not have any operations on the day. For instance, if the business organization does not have any

operations that require cloud computing for a period, the organization would not incur any charges for that period.

Unlike physical database servers, corporate organisations are not responsible for the safety and security of their customers' data; instead, cloud providers are responsible for providing that protection and security. These cloud providers are concerned with the backing up of data. Hence, users have no trouble with securing and retrieving data.

C. Service Models of Cloud Computing Technology

There are three basic service models/components in cloud computing:

- **Software-as-a-Service(SaaS):** These has to do with the provision of software services through cloud computing. Instead of installing software applications on the individual computer of an organization or on the database server, which both require consistent updates, frequent version upgrades, and regular patches, etc., software applications are provided on demand through cloud computing. Software for commercial organisations such as ERP (Enterprise Resource Planning), CRM (Customer Relationship Management), Word processing, and so on can be delivered via the internet using cloud technology. This means that organizations do not need to bother themselves about the development and maintenance of the software app as this would be handled by the cloud providers.
- **Platform-as-a-Service(PaaS):** This involves the provision of an environment or platform where business organizations can develop, test, or deliver any application. Organizations do not need to buy the software licenses for platforms/environments like middleware, databases and operating systems. Instead, these platforms, software development kits, and tools are all made available over the Internet through cloud technology. With PaaS, developers no longer have difficulties in creating an environment/platform for the development of their applications. But such an environment/platform is made available on demand to the software developers, testers via cloud organisations such as ERP (Enterprise Resource Planning), CRM (Customer Relationship Management), Word processing, and so on can be delivered via the internet using cloud technology. This means that organizations do not need to bother themselves about the development and maintenance of the software app as this would be handled by the cloud providers.
- **Infrastructure-as-a-Service(IaaS):** This refers to the use of virtual components to replace real computer equipment. Physical computing devices, such as storage devices, processing servers, virtual computers, operating systems, and network transmission, that were once housed in a data center, may now be accessible and used via the internet with simply a login authentication system. Because all of these services are available through cloud providers, companies would simply have to hire them on a rental basis using IaaS. These three basic service models/components enable business organizations to employ the “new business strategies for the organizational growth” [1].

D. Cloud Computing Deployment Models

Public cloud service, private cloud service, community cloud service, and hybrid cloud are the four types of cloud computing deployment methods. Each of these development approaches offers a unique set of services to users. These deployment models are classified based on several factors like location where the services are hosted, security requirements, cloud services sharing, and customization capabilities [6].

Public Cloud Service: This service is available to the general public over the Internet, as the name implies. Third-party service providers also make this service available to the general public. Small and medium-sized company enterprises can use the public cloud service to deploy IT solutions at a low cost. Third-party service providers, on the other hand, have their own set of values, policies, pricing, and costing models [7]. Google Apps, email, and media storage services are examples of public cloud services.

Private Cloud Service: This service is intended for usage by organisations, and it is commonly utilized by major corporations. The security of a private cloud service is higher than that of a public cloud service. This is due to the fact that the private cloud service is only available to the organization's trusted users. This cloud service is used by businesses to share client data throughout their branches [8]. Private cloud service is more reliable, more efficient, less risky, and offers higher security and lower complexity [9]. This is very suitable

for any business organization; however, more cost would be incurred by the organization to purchase the necessary equipment and software to have their private cloud computing system.

Community Cloud Service: A collection of corporate entities with shared interests utilizes the community cloud service. Though it is similar to the private cloud in that it is not utilized by a single company and the data saved is shared among companies with similar interests, it differs in that it is not used by a single organization. Organizations or a third party's agent may administer the community cloud services. The cost of setting up the community cloud service is cheaper than that of private cloud, this is because the cost is shared among the organizations that will benefit from its use. A better level of privacy and security is provided by a community cloud service. The educational and research cloud system, which may be shared across multiple institutions for academic purposes all over the world, is an example of a community cloud service [9].

Hybrid Cloud Service: A hybrid cloud service combines two or more cloud services into one. This sort of cloud service provides a service using real hardware and servers. The advantages of the hybrid cloud service are nearly identical to those of the private and public cloud services. Because the hybrid cloud service saves sensitive data of corporate organisations in private cloud storage, it is simpler to manage data security. The hybrid cloud is typically utilized for backup [10].

E. Advantages of Cloud Computing

Cloud computing is currently playing a significant role in the corporate sector. Cloud computing is a cutting-edge technology that has provided several benefits to businesses and is assisting them in remaining competitive [11]. Cloud computing offers enhanced and novel capabilities not seen in traditional IT solutions.

Cloud computing offers great benefits to all business organizations [12] – micro businesses, small and medium scale businesses, and enterprises, which include the following:

Flexibility: Cloud computing increases the flexibility of business organizations by enabling each worker more work flexibility within and without their workplace. For instance, workers who are not physically present in the office can access necessary documents, files, and data on any device as long as they are connected to the Internet [13]. More so, since the data and files are stored in the cloud, the same file can be accessed simultaneously by different workers.

Cost Reduction: One of the major reasons why business organizations opt in for cloud computing is cost reduction [14]. Compared to the traditional IT system, cloud computing is more cost efficient. Business organizations do not need to spend money on setting up physical structures and they only pay for the cloud service they use.

Automatic Software/Hardware Upgrades: Traditional IT solutions are always plagued with high cost of maintenance, and continuous updates and upgrades of software and hardware [15]. This sometimes affects the productivity of the business organization. However, these problems have been solved by cloud computing. There is no hardware to maintain or upgrade and the software update is done by the cloud provider over the cloud. This does not in any way affect the operation and productivity of the business organizations.

Agility: To be competitive nowadays, businesses must maintain a highly competitive business climate. One of the ways of creating a unique business environment and remaining competitive is for an organization to be able and timely respond to their customers' fast changing needs [16]. These can be achieved in a more effective manner through cloud computing [17]. Cloud computing is readily available and accessible as it is available on the Internet. This enables organizations to deliver their services within the shortest time thereby staying agile.

Other advantages of cloud computing include scalability, no up-front cost, increased collaboration, improved file/document control, security, improved productivity via remote working, environmentally friendliness, disaster recovery, and competitiveness.

3. Research Methodology

The goal of this study is to look into the perceived benefits and dangers of cloud computing in terms of security and cost for enterprises. While this paper will not give precise measurements or enumerations for specific advantages and hazards, It will provide a thorough examination of the economic and security implications of cloud computing for organisations, outlining all existing benefits and dangers. As a result, the purpose of this study is to address a vacuum in the existing research, namely the absence of findings on the financial risks and security advantages of cloud computing.

This research paper will not generate its findings using qualitative data but will adopt the quantitative approach and base its findings on subjective assessment of opinions. Hence, this paper will collect empirical data through interviews and also review existing literature for the purposes of data analysis [18]. A research model is constructed based on the study objectives, as shown in Figure 1.

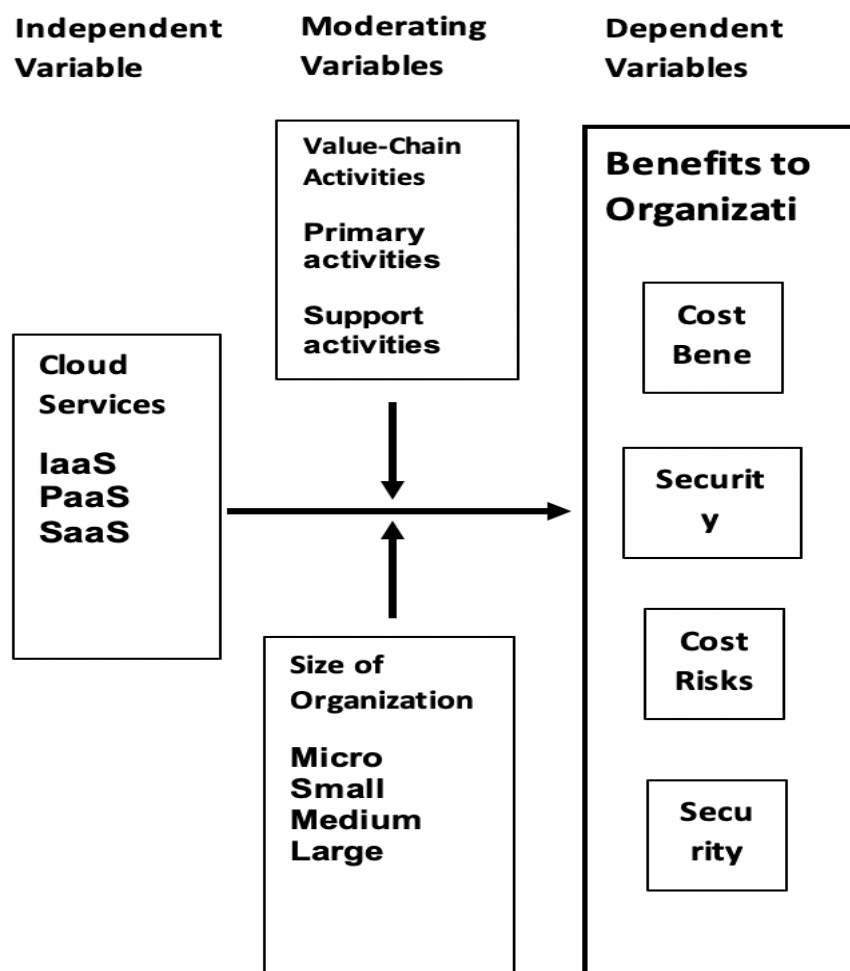


Fig. 1. Research model

A. Research Hypothesis

As previously said, CC services are divided into three categories: SaaS, PaaS, and IaaS. Based on the link between cloud computing services and their consumers, these three create a hierarchical structure. This hierarchical structure is SaaS, PaaS, and IaaS respectively, from the most direct service to the less direct one. It is generally believed that SaaS contributes the most and more directly to the users that the other two. This paper therefore proposes its first research hypothesis (H1) as follow:

H1: Many types of cloud computing services provide varying stages of cost and security protection to businesses.

It's also considered that company size has an impact on cloud computing adoption. Large companies are more resourceful, have an edge in adopting and implementing new technologies, and can bear greater risks. Micro, small, and medium businesses, on the other hand, are typically vulnerable to risk owing to a lack of or restricted resources, and should exercise caution while adopting and exploiting new technologies.

According to the premise, the size of the business will have an impact on the cost and security of this system. Therefore, the second research hypothesis (H2) proposed in this study (H2): H2: The cloud services differ based on the company's size.

These characteristics are assessed on a five-point Likert scale ranging from strongly disagree (1) to strongly agree (5), higher the score, higher the firm is successful.

B. Data Collection and Analysis

For this study, empirical data was gathered using both quantitative and qualitative methods. Empirical data is information gathered from people's personal experiences. This study will use a one-on-one interview procedure as well as a survey research employing a 5-point based questionnaire as a measuring instrument to gather data for this research.

The well-structured questionnaire has 15 items, including the respondents' background information, their industry, the number of people in the company, the kind and usage of cloud computing services used by the company, and so on. Ten business owners whose organizations have adopted the cloud computing service were interviewed. About 200 business organizations were randomly selected to participate in this research via questionnaire method but only 65 respondents returned valid responses.

IBM SPSS Statistics 23 is the analytical software used to analyze the data. The outcome of the findings are explained below.

4. Results and Discussion

The results of the interviews are presented in this section. It covers the respondents' opinions on the perceived economic and security risks and impacts of cloud computing.

Although earlier research has indicated that one of the primary edge of this computing model is cost effectiveness, this study revealed that cloud computing was a better alternative than the traditional server system because of its flexibility and variable pricing model [19]. Another cost benefit is the cost savings from hardware, software, and infrastructure [20]. Because these apps are now in the cloud, business organisations would no longer need to acquire or build up expensive and high-powered computer systems or infrastructure to operate them. This, therefore, eliminates maintenance costs. While all these are in support of previous findings, this research revealed that cost reduction has gone beyond just eliminating hardware and software but about staffing. That means, business organizations that adopt and use cloud computing save cost from the reduction of both operational and administrative costs.

In terms of cost risks, this study found that the most significant cost risk that cloud computing poses to businesses is the absence of strong and accurate cost models on the present cloud market, which makes cloud computing cost estimate difficult. Second, certain cloud-based programs may have limited capability compared to their traditional software equivalents.

The primary concerns of using cloud computing services have been security concerns [21]. Data loss, data breach, isolation failure, unauthorized access to sensitive and private information, backup vulnerabilities, management interface compromise, and insufficient data erasure are just a few examples. Other risks from other studies include inadequate monitoring, missing assurance and transparency, vendor lock-in, and gap in security. However, finding in this study shows that the risk of data loss does not exist with cloud computing but the risk of data breach. Other risks as found by this study include the lack of trust and integrity.

In terms of security benefits, some respondents claimed that cloud computing increases data safety. Others claimed that the cloud is more secure than in-house servers and data centres as the cloud has better improved

security maintenance and security software mechanisms than the traditional systems. Risk management, security maintenance, and support costs are cheaper for businesses that use cloud services.

Table 1. Cloud computing services that produce various levels of perceived advantages

	<i>SaaS</i>	<i>PaaS</i>	<i>IaaS</i>	
Primary activities	10	10	17	37
Support activities	12	5	11	28
Total	22	15	28	65

Cloud services are typically used by businesses for either core or secondary functions. Inbound logistics, operations, outbound logistics, marketing and sales, and services are examples of core activities, whereas support activities include firm infrastructure, human resource management, procurement, and technological development. Table 1 shows that, while various types of cloud computing services serve different objectives, they may all create the same amount of value. There is no significant difference in the amount of perceived benefit generated by each service as each service can be used for either activity.

Table 2. Perceived benefits of cloud computing services based on the size of organisation

Activity	Type of Cloud Computing			Total
	SaaS	PaaS	IaaS	
Large	6	6	10	22
SMEs	16	9	18	43
Total	22	15	28	65

Table 2 shows the perceived benefits of cloud computing services dependent on the size of the corporate organisation. As a consequence of the findings, it is clear that the perceived value of cloud computing services varies significantly depending on the size of the company. That is, the perceived benefits gained by business organizations from the different cloud computing services they adopted are influenced by the sizes of the business organization. Small and medium-sized businesses profit more from cloud services than large businesses since they don't require their own infrastructure, apps, or IT personnel to access them.

A. Scale Reliability and ANOVA

The influence of the independent factors and moderating variables on the dependent variable was evaluated using MANOVA in this study. Because this study is exploratory in nature and the final sample size is less than optimum, the Wilks' lambda value for a significant test was set at 0.1. [22]. In addition, to reduce the likelihood of type 1 mistake, this study used the Roy-Bargmann Stepdown Analysis technique [23]. The M value of Box = 74.338, F equals 2.068, and P equals 0.001 in the multivariate homogeneity of variance test, which is less than the specified significant threshold. This demonstrates that the independent variables' variances do not contradict the homogeneity assumption.

For the hypothesis test of the influence of different types of cloud computing services on perceived business advantages, the MANOVA result was $F(5, 115) = 0.339$, $P > 0.1$. (H1). This outcome isn't very noteworthy. This demonstrated that the benefits of various cloud computing services for businesses are not considerably

different. However, $F(3, 57) = 2.696$, $P = 0.1$, which is significant at the level of 0.1, is the finding for the benefits of CC across different company sizes (H2). This demonstrated that the value of this computing model varies with company size.

5. Conclusion

This is an exploratory research to look into the benefits and risks of cloud computing in terms of cost and security for businesses. This refers to the link between various forms of cloud computing services and the perceived value to businesses, which is moderated by the size of the firm. The main proposals are that different cloud computing services offer different benefits to business organizations. In addition, the link between the different types of cloud computing services and the advantages they provide is depending on the size of the company that uses them. Based on these ideas, four hypotheses were developed, and a survey was performed to test them.

The results of the data analysis showed that different cloud computing services do not offer any significantly different benefit to business organizations. That is, all cloud computing services offer slightly the same benefits to business organizations. The results also showed that SMEs gained higher benefits from cloud computing than large companies did, especially in the area of enhanced business scalability through large organizations benefited more in the area of cost saving than SMEs did

6. Recommendations

From the study findings, it can be seen that security concerns are the main issues that customers consider when adopting cloud computing. Consumers need assurance that cloud computing service providers adhere to solid security standards in order to mitigate possible hazards to both consumers and providers. This paper recommends that cloud computing service providers should guarantee respective customers that they would enjoy a good level of safety from any form of cyber-threats in the form of a solid and mutual agreement. Furthermore, there is a need for more research to be conducted focusing more on these gaps.

References

- [1] Milan, E., De MesquitaSpinola, M., & De Paula Pessoa, M. (2015). Opportunities for strategic alignment of IT to business with the adoption of cloud computing: Case studies in large organizations. *Management of Engineering and Technology (PICMET)*, 2015 Portland International Conference on, 64-74.
- [2] Hosseini, A. K., Sommerville, I. & Sriram, I. (2010). Research Challenges for Enterprise Cloud Computing. Unpublished, <http://arxiv.org/abs/1001.3257> (2010)
- [3] McAfee, A. (2011). What Every CEO Needs to Know About the Cloud. *Harvard Business Review*.
- [4] Greenwood, D., Khajeh-Hosseini, A., Smith, J. & Sommerville, I. (2011). The Cloud Adoption Toolkit: Addressing the Challenges of Cloud Adoption in Enterprise. *Cloud Computing Co-laboratory*, School of Computer Science, University of St Andrews, UK.
- [5] Boss, G., Malladi, P., Quan, D., Legregni, L. & Hall, H. (2007). Cloud Computing. IBM Corporation.
- [6] Rajaraman, V. (2014). "Cloud Computing," *RESONANCE*, pp. 242-258.
- [7] Adam, M. M. et al., (2014). Impact of Cloud Computing Adoption on Stock Price. Savannah, AISeL
- [8] Sumit, G., (2014). Public vs Private vs. Hybrid vs Community - Cloud Computing: A Critical Review. *International Journal of Computer Network and Information Security*, 3(1), pp. 20-29.
- [9] Ahmed, Y., (2012). Exploring Cloud Computing Services and Applications. *Journal of Emerging Trends in Computing and Information Sciences*, 3(6), pp. 838-847.
- [10] Tarannum, N. & Ahmed, N., (2013). Efficient and Reliable Hybrid Cloud Architecture for Big Database. *International Journal on Cloud Computing Services and Architecture*, 3(6), pp. 17-29.
- [11] Etro, F. (2009). The Economic Impact of Cloud Computing On Business Creation, Employment and Output in Europe. *Review of Business and Economics*, 54 (2), 179-208.
- [12] Ankeny, J. (2011). Heads in the cloud. *Entrepreneur*, 39(10), 50-51.
- [13] Nabil, S., (2010). Cloud Computing for Education: A New Dawn? *International Journal of Information Management*, 30(2), pp. 109-116.
- [14] Tiago, O., Manoj, T. & Mariana, E., (2014). Assessing the Determinants of Cloud Computing Adoption: An Analysis of the Manufacturing and Services Sectors. *Information & Management*, 51(5), pp. 497-510.

- [15]Piotr, P. & Zbigniew, P., (2013). Cloud Computing - A Case Study for the New Ideal of the IS/IT Implementation. Zadar, ToKnowPress.
- [16]Jafar, S. et al., (2013). Identifying Benefits and Risks Associated with Utilizing Cloud Computing. The International Journal of Soft Computing and Software Engineering, 3(3), pp. 416-421.
- [17]Ta-Tao, C., Kazuo, N. & Thoma, T. C., (2015). An Exploratory Study of Expected Business Value of Cloud Computing. Issues in Information Systems, 16(4), pp. 37-47
- [18]Kothari, C. R. (2004). Research Methodology. Second Revised Edition. New Age International Publishers.
- [19]Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., Lee, G., Patterson, D., Rabkin, A., Stoica, I. & Zaharia, M. (2009). Above the Clouds: A Berkeley View of Cloud Computing. Electrical Engineering and Computer Sciences, University of California at Berkeley.
- [20]Goncalves, V. & Ballon, P. (2011). Adding value to the network: Mobile operators' experiments with Software-as-a-Service and Platform-as-a-Service models. Telematics and Informatics. Vol. 28 (2011), 12-21.
- [21]Pearson, S. (2012). Privacy, Security and Trust in Cloud Computing. Springer.
- [22]Lavrakas, P. J. (2008). Encyclopedia of Survey Research Methods. SAGE Publications.
- [23]Huberty, C. J., & Morris, J. D. (1989). Multivariate Analysis versus Multiple Univariate Analyses. Psychological Bulletin, 105.2: 302.

Hyper-parameter Optimization of Image Classifiers using Nature-Inspired Algorithms

Rahul Sharma¹ and Amar Singh²

¹ Lovely Professional University, Phagwara, Jalandhar, India
prof.sharma.rahul@gmail.com

² Lovely Professional University, Phagwara, Jalandhar, India
amar.23318@lpu.co.in

Abstract. Nature-inspired algorithms are popularly used to handle complex optimization problems for which exact mathematical solution is either not available or is computationally very expensive. Nature-inspired algorithms and image classification have gained immense popularity and are applied to different fields like biology, biology, medicine, humanities, management sciences, etc. Soft computing techniques are applied to solve problems dealing with imprecision, uncertainty, partial tone, and approximation to achieve tractability, robustness, and optimum. In this paper, we have studied different applications of soft computing techniques for image classification and plant disease detection. Soft computing-based techniques can be used for improving the efficiency of traditional image classification methods. Selecting the best hyper-parameters of the model enhances the performance of the classifier. In this paper, traditional machine learning algorithms random forest (RF), support vector machine (SVM), k-nearest neighbor (KNN), artificial neural network (ANN) are used for image classification and their hyper-parameters are tuned using Grid Search, Particle Swarm Optimization (PSO), and genetic algorithm (GA). Experimental results demonstrate that GA optimized ANN classifier was able to classify images with 99.48% accuracy..

Keywords: Genetic Algorithm, Image Classification, Particle Swarm Optimization, Soft Computing, Optimization.

1. Introduction

Image classification techniques can be used to increase the productivity in agriculture sector [1]. Worldwide plant diseases cause major economic loss to farmers. The reduction of agricultural land, soil degradation, rising population, climate changes, biodiversity losses, etc are some of the challenges faced by the agriculture sector. Above 60% of the population worldwide depends upon the agriculture yield. Early and accurate plant disease detection and adopting timely preventive measures will help farmers save the crop and improve yield. The traditional method of monitoring vast crop fields is time-consuming and error-prone. Scalable image classification is a challenging task [2]. The image classification approach must be able to process images with different backgrounds, light conditions, scale variations, deformation occlusion, angle of capturing device, etc. Modern tools like machine vision and artificial intelligence are very useful for the recognition and classification of plant diseases. Machine learning algorithms (RF, SVM, KNN, ANN, etc), deep convolution neural networks (DCNNs), and fuzzy logic-based approaches are employed for image classification tasks. Selecting the best hyper-parameters of this machine algorithm is a challenging task. Manually choosing the hyper-parameters requires technical expertise and domain knowledge. Nature-inspired search and optimization techniques (GA, PSO, ABC, etc) can be used for automatically tuning the hyper-parameters of the classifier and improving the accuracy of the machine learning models. Nature always evolves and moves towards optimization, using the same concept solution of hard real-world problems can be analyzed. Nature's properties like self-organization, flexibility, robustness, collective work; etc are used to derive nature-inspired algorithms. Nature-inspired algorithms include bio-inspired and physics and chemistry-based algorithms. Bio-inspired algorithms are classified as swarm intelligence (SI) algorithms and evolutionary algorithms. Some of the popular SI algorithms are ACO, PSO, ABC, Firefly algorithm, Cuckoo Search, etc. Evolutionary algorithms like BBBC, PB3C, GA, P3PGA, ABC, Paddy Field, Genetic Algorithms, etc are derived from nature and are employed by different applications. Nature-inspired algorithms are used for optimizing applications ranging from engineering design, business planning, water distribution, etc. Optimization means tuning input values for a given function to minimize or maximize the output. The optimized solution can either be the best solution or the solution relatively close to the best solution. The objective of the optimization

is problem specific which can be minimizing cost, energy consumption or maximizing the profit of a business, or maximizing the network bandwidth utilization.

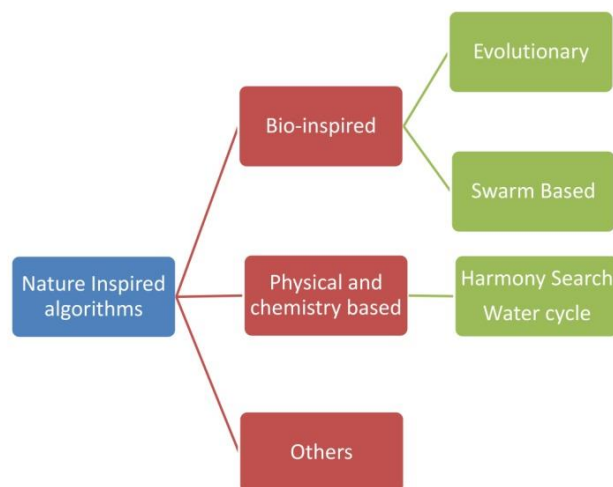


Fig. 1. Classification of Nature-inspired optimization algorithms

Nature-inspired algorithms (NIA) increase productivity as they can be used for feature selection, parameter selection. NIA play an important part in the field of image processing tasks. Optimization techniques are used for noise reduction, blurring of images, image enhancement, restoration of image, segmentation, edge detections, artificial image generation, image fusion, pattern recognition, etc. For optimization of image processing tasks capabilities of different NIA are utilized.

In addition to image classification, nature-inspired algorithms have diverse application areas like intrusion detection, travelling salesman problems, vehicle routing, scheduling problems, software testing, etc.

2. Literature review

Nature-inspired algorithms are integrated with machine learning and image feature extraction and reduction approaches. Table 1 presents different approaches proposed.

Table 1. Optimization of image classification based on Nature-inspired algorithms

Paper	Approach Used	Dataset	Advantages
[3]	A hybrid approach using principal component analysis and Whale optimization algorithm for feature extraction.	plant-village dataset	High testing accuracy of 94% and training accuracy of 99%
[4]	Optimized feature selection using bat algorithm. 11 features were extracted.	Leukocyte Images for Segmentation and Classification (LISC)	Performed better than other recent optimization algorithms with an accuracy of 97.69%
[5]	Hybrid model using modified optimized CNN and grasshopper feature optimization. Features extracted using SURF and CNN were studied	Plant village	Enhanced accuracy by 2.18% when compared with traditional image classifiers
[6]	A new exponential spider monkey optimization algorithm was used to reduce the features obtained by	Plant Village	Performed better than other nature-inspired algorithms

	SPAM (selecting 82 features from 686 features). The classification task was done by SVM		with an accuracy of 92.12%
[7]	Low-level features are obtained by scale-invariant feature transformation. SVM for image classification	Bud highres	Accuracy of 90.08%
[8]	Classified plant leaf diseases using bacterial foraging optimization based neural network	PlantVillage Disease Classification Challenge	High accuracy when compared with GA and KNN
[9]	Particle swarm optimization is used for the selection of optimized hyperparameters during segmentation and classification.	crowdAI database	For Segmentation 95.% accuracy. Classification accuracy of 97.59%
[10]	Particle swarm optimization based feature selection	WANG dataset	The proposed approach achieves a better performance matrix when combined with other well-known classifiers
[11]	PCA for dimensionality reduction, The bio-inspired hierarchical model integrated with SIFT for feature extraction The classification was done by the SVM model.	McGill color image dataset	Under the different complex conditions, the proposed algorithm was able to identify the images with an accuracy of 85.5%
[12]	PSO algorithm was employed for optimizing the CNN	MNIST, Rect340 angles, Convex, and MNIST-Fashion	The proposed approach illustrated better accuracy when compared to other state of art classifiers.
[13]	Inception neural network for feature extraction. Feature selection using Marine Predators Algorithm.	Two datasets are used in the study. 1.COVID-19 images by ICR 2.Kaggle COVID 19 X-ray datasets	1. Dataset 1 accuracy of 98.7% 2. Dataset 2 accuracy of 99.6%
[14]	Efficient approach employing Grey Wolf Optimization (GWO) and SVM	3064 MRI images	GWO performed better than PSO and the firefly algorithm
[15]	Improved bat optimization with multiobjective operators and proposed a new random walk based new searching technique	1. leukemia cancer data 2. Prostate cancer data 3. SRBCT dataset	Cancer datasets were explored to analyze the significance of biomarkers.

[16]	A new red deer algorithm (RDA) is proposed		Solved 12 benchmark functions and applied for solving many real-world problems including SMSP, TSP, FCTP, and VRP
[17]	A new improved crow search algorithm for Parkinson's disease prediction was proposed. Suitable for feature selection	20 different standard datasets were employed for testing the proposed approach.	Results illustrate that the modified crow search algorithm improves the performance of the original.
[18]	PSO algorithm for optimization and evolution of CNN model was proposed	MNIST and Rotated Digits Images and the Convex Sets (CS)	Outperformed 12 well-known existing algorithms.
[19]	A hybrid approach was proposed. The VGGNet model for feature extraction, feature selection using enhanced Salp Swarm Algorithm. Six well-known classifiers were used.	1.ALL-IDB1 2. C-NMC	The proposed algorithm performed reduced the complexity and improved accuracy
[20].	Autoencoder integrated with the deep neural network was employed for feature reduction and classification	RIDER Neuro- MRI Dataset	Overall Accuracy of 96%
[21]	SURF method along with SVM classification of soil dataset optimized by chaotic spider monkey is presented	Soil image dataset	With an accuracy of 79%, the proposed algorithm performed better than PSO, GSA, etc
[22]	Salp Swarm Algorithm along with traditional image classification methods like Decision Tree, KNN, and Naive Baye was proposed.	120 T2-type MRI images	Efficient and accurate classifiers
[23]	Proposed an approach VMVI-CNN for feature extraction and richness of feature information.	MNIST, Cifar-10, and Mini-ImageNet	Better identification performance compared to other well-known classifiers.
[24]	Hybrid feature selection by improving basic firefly optimization algorithm is proposed.	MNIST Cifar-10 Mini-ImageNet FDI-1000	Compared to artificial bee colony, cuckoo search, PSO the proposed approach illustrated better performance accuracy
[25]	Discrete wavelet transform for feature extraction from brain images PCA for dimensionality reduction Nature-inspired computing for ANN	60 MRI images from the local hospital	The proposed approach achieved the highest classification accuracy.

[26]	GLCM feature extraction, Lion Optimization Algorithm (LOA) Monkey Search Optimization Algorithm (MSO) and SVM for classification	Mammogram Image Analysis Society (MIAS)	LOA-SVM and MSO-SVM illustrate the application of nature-inspired techniques for image classification tasks.

3. Preliminaries

In this section different swarm, intelligence-based algorithms are discussed.

Genetic Algorithm

Genetic Algorithm (GA) is a meta-heuristic evolutionary method that was presented by John Holland in the early 1960s [27]. GA is based on the natural evolution of a biological species and can be applied for stochastic search or for finding an optimal or near-optimal solution. Solution space comprising of Individuals also known as chromosomes matures with every generation to provide an optimal or near-optimal solution. Individuals are selected based on a selection method (roulette wheel, Boltzmann selection, tournament selection, rank selection, steady state selection, etc) from the population for reproduction. The cross-over method (single-point crossover, two-point crossover, uniform crossover masked crossover, etc) is applied to produce a new chromosome by combining parts of parents. Produced chromosome retains the useful features of the parents and ignores less useful features.

To avoid local optimality [28] some (low probability) of the offspring are mutated. The mutation causes genetic diversity by bringing a small change in the chromosome's element allele. Some of the common mutation techniques are power mutation, uniform mutation, Gaussian mutation, shrink mutation, etc. Chromosomes with the best fitness value are retained in the new population. This process continues until the desired optimal solution is obtained or the process continues for a maximum number of generations.

Particle Swarm Optimization

PSO was introduced by Kennedy and Eberhart in 1995 and is an evolutionary computation technique based on research of the flock of birds searching a food source or school of fish protecting them from a predator. PSO algorithm a group of birds is searching for food in an area where one piece of food is available. Birds do not know about the exact location of the food source. But they know how far the food from their current location is. In this algorithm, all birds of the flock change their velocity based on their experience and by following the bird nearest to the food. Each bird or solution in PSO is called a particle. Searching the optimized solution in the search space is based on the particle's best position and global best position. The optimum solution is searched based on best fitness function.

Machine Learning and Hyper-parameter Tunning

Machine learning algorithms like Random Forest, SVM, KNN, ANN, etc are widely used in different applications. Building an accurate and efficient machine learning model is a complex and time-consuming task because it requires evaluating suitable ML algorithms for a given dataset and tuning the hyper-parameters. Hyper-parameters define the architecture and are set before training the ML model. Random forest important hyper-parameters are number of estimators, maximum features, maximum depth, minimum samples, split, minimum leaf samples, criterion etc. Important hyper-parameters of SVM are regularization parameter C, kernel type to be used, kernel coefficient gamma, etc. Learning rate, optimizer, activation function, batch size, number of neurons, epochs, etc are some of the ANN parameters. . Hyper-parameter tuning is a process of evaluating and selecting the optimum hyper-parameters of the ML model.

Manual tuning of hyper-parameters is based on a trial and error method and requires deep knowledge of ML algorithms and different configurations. Different techniques for automatically tuning hyper-parameters are random search, grid search, Bayesian optimization, nature-inspired optimization, etc. In this paper, we have explored grid search, genetic optimization, and PSO techniques for the optimization of image classifiers.

4. Results and Discussion

Experiments were conducted on the standard dataset to compare the performance of different optimization techniques. In this paper, we have used sklearn library digits dataset which is a copy of the UCI machine learning optical recognition of hand-written digits dataset. The dataset includes a total of 1797 images. Each

dataset image is of size 8×8 with 10 classes. 80% of images were used for the training model and 20% images were used for training.

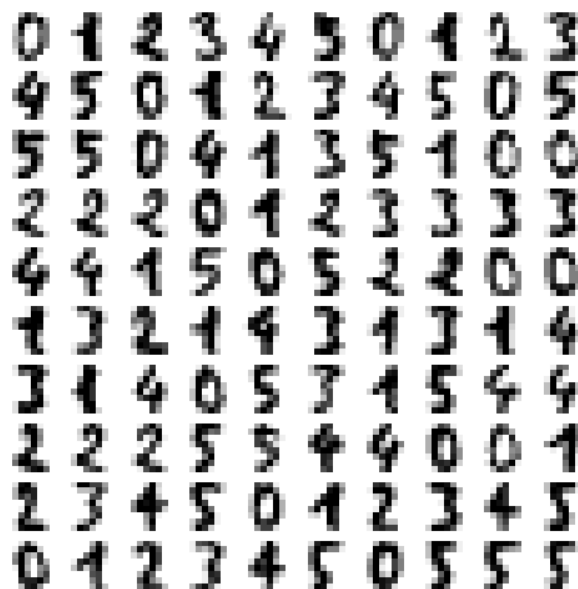


Fig. 2. UCI digits dataset

Traditional machine learning algorithms Random forest (RF), support vector machine (SVM), k-nearest neighbor (KNN), artificial neural network (ANN) were trained using digits dataset with default configurations. The accuracy obtained by the models was recorded and is shown in Table 2.

Table 2. Classification accuracy and default hyper-parameters			
Sno.	Traditional Classifier	Accuracy	Default Hyper-parameters
1.	RF	93.05	n_estimators=100, max_features='auto', max_depth=None, min_samples_leaf=1, bootstrap=True, criterion='gini'
2.	SVM	95.27	C=1.0, kernel='rbf', gamma='scale'
3.	KNN	93.33	n_neighbors=5
4.	ANN	96.01	optimizer= 'adam', activation='relu', batch_size=32, neurons=32, epochs=20, patience=3

Selecting the best hyper-parameter of the model enhances the classification accuracy of the models. Experiments were conducted for automatic tuning of hyper-parameter using the Grid Search algorithm and the results are presented in Table 3.

Table 3. Grid Search hyper-parameter tuning			
Sno.	Classifier	Accuracy	Grid Search optimized Hyper-parameters
1.	RF	93.33	n_estimators=30, max_features='sqrt', max_depth=15, min_samples_leaf=2, bootstrap=False, criterion='entropy '
2.	SVM	96.66	C=10, kernel='rbf', gamma='0.1'
3.	KNN	94.16	n_neighbors=3
4.	ANN	98.01	optimizer='adam', activation='tanh', batch_size=32, neurons=32, epochs=50, patience=5

In this paper, we have used PSO and GA for hyper-parameter tuning. The optimized hyper-parameter obtained using PSO and GA optimization algorithms are shown in Table 4 and Table 5 respectively.

Table 4. PSO optimized hyper-parameter

Sno.	Classifier	Accuracy	Grid Search optimized Hyper-parameters
1.	RF	93.34	n_estimators=68, max_features=10, max_depth=36, min_samples_leaf=3, 'min_samples_split'=4, bootstrap=False, criterion='entropy '
2.	SVM	97.44	C=3.417, kernel='rbf', gamma='0.24'
3.	KNN	96.82	n_neighbors=4
4.	ANN	98.92	optimizer='adam', activation='relu', batch_size=38, neurons=22, epochs=23, patience=3

Table 5. GA optimized hyper-parameters

Sno.	Classifier	Accuracy	Grid Search optimized Hyper-parameters
1.	RF	94.82	n_estimators=85, max_features=8, max_depth=39, min_samples_leaf=2, 'min_samples_split'=2, bootstrap=False, criterion='gini '
2.	SVM	97.44	C=6.96, kernel='rbf', gamma='0.19'
3.	KNN	96.82	n_neighbors=3
4.	ANN	99.48	optimizer='rmsprop', activation='tanh', batch_size=16, neurons=66, epochs=50, patience=13

5. Conclusion

In image processing, an efficient and accurate solution is desired. Nature-inspired algorithms can be applied to find the optimum solution. Nature-inspired algorithms need high computation power. Nature-inspired algorithms are employed for feature reduction, the evolution of neural networks, image segmentation, selecting the optimized hyperparameter of the classifier, etc. In this paper, we have presented an approach integrating the nature-inspired techniques PSO and GA and traditional machine learning algorithms RF, SVM, KNN, and ANN for obtaining the optimized hyper-parameters and enhancing the accuracy of the image classifiers. GA was used for obtaining the optimized hyper-parameters of ANN (optimizer = 'rmsprop', activation = 'tanh', batch size=16, neurons=66, epochs=50, patience=13). Experimental results illustrate that using the optimized hyper-parameter; ANN was able to classify digit images with an accuracy of 98.48%.

References

- [1] Hsieh, T. H., & Kiang, J. F. (2020). Comparison of CNN algorithms on hyperspectral image classification in agricultural lands. *Sensors*, 20(6), 1734.
- [2] Paoletti, M. E., Haut, J. M., Plaza, J., & Plaza, A. (2020). Scalable recurrent neural network for hyperspectral image classification. *The Journal of Supercomputing*, 76(11), 8866-8882.
- [3] Gadekallu, T. R., Rajput, D. S., Reddy, M. P. K., Lakshman, K., Bhattacharya, S., Singh, S., ... & Alazab, M. (2020). A novel PCA-whale optimization-based deep neural network model for classification of tomato plant diseases using GPU. *Journal of Real-Time Image Processing*, 1-14.
- [4] Gupta, D., Agrawal, U., Arora, J., & Khanna, A. (2020). Bat-inspired algorithm for feature selection and white blood cell classification. In *Nature-Inspired Computation and Swarm Intelligence* (pp. 179-197). Academic Press.
- [5] Al-bayati, J. S. H., & Üstündağ, B. B. (2020). Early and Late Fusion of Deep Convolutional Neural Networks and Evolutionary feature optimization for Plant leaf illness recognition. *JX Univ. of Archit. & Technol.*, 12(2), 1591-1610.

- [6] Kumar, S., Sharma, B., Sharma, V. K., Sharma, H., & Bansal, J. C. (2020). Plant leaf disease identification using exponential spider monkey optimization. *Sustainable computing: Informatics and systems*, 28, 100283.
- [7] Pérez, D. S., Bromberg, F., & Diaz, C. A. (2017). Image classification for detection of winter grapevine buds in natural conditions using scale-invariant features transform, bag of features and support vector machines. *Computers and electronics in agriculture*, 135, 81-95.
- [8] Chouhan, S. S., Kaul, A., Singh, U. P., & Jain, S. (2018). Bacterial foraging optimization based radial basis function neural network (BRBFNN) for identification and classification of plant leaf diseases: An automatic approach towards plant pathology. *IEEE Access*, 6, 8852-8863.
- [9] Kour, V. P., & Arora, S. (2019). Particle swarm optimization based support vector machine (P-SVM) for the segmentation and classification of plants. *IEEE Access*, 7, 29374-29385.
- [10] Garg, M., & Dhiman, G. (2021). A novel content-based image retrieval approach for classification using GLCM features and texture fused LBP variants. *Neural Computing and Applications*, 33, 1311-1328.
- [11] Deng, L., Wang, Y., Han, Z., & Yu, R. (2018). Research on insect pest image detection and recognition based on bio-inspired methods. *Biosystems Engineering*, 169, 139-148.
- [12] Junior, F. E. F., & Yen, G. G. (2019). Particle swarm optimization of deep neural networks architectures for image classification. *Swarm and Evolutionary Computation*, 49, 62-74.
- [13] Sahlol, A. T., Yousri, D., Ewees, A. A., Al-Qaness, M. A., Damasevicius, R., & Abd Elaziz, M. (2020). COVID-19 image classification using deep features and fractional-order marine predators algorithm. *Scientific reports*, 10(1), 1-15.
- [14] Kumar, A., Ansari, M. A., & Ashok, A. (2019). A hybrid framework for brain tumor classification using grey wolf optimization and multi-class support vector machine. *International Journal of Recent Technology and Engineering (IJRTE)*, 8(3), 7746-7752.
- [15] Dashtban, M., Balafar, M., & Suravajhala, P. (2018). Gene selection for tumor classification using a novel bio-inspired multi-objective approach. *Genomics*, 110(1), 10-17.
- [16] Fathollahi-Fard, A. M., Hajiaghaei-Keshteli, M., & Tavakkoli-Moghaddam, R. (2020). Red deer algorithm (RDA): a new nature-inspired meta-heuristic. *Soft Computing*, 24(19), 14637-14665.
- [17] Gupta, D., Sundaram, S., Khanna, A., Hassanien, A. E., & De Albuquerque, V. H. C. (2018). Improved diagnosis of Parkinson's disease using optimized crow search algorithm. *Computers & Electrical Engineering*, 68, 412-424.
- [18] Wang, B., Sun, Y., Xue, B., & Zhang, M. (2018, July). Evolving deep convolutional neural networks by variable-length particle swarm optimization for image classification. In *2018 IEEE Congress on Evolutionary Computation (CEC)* (pp. 1-8). IEEE.
- [19] Sahlol, A. T., Kollmannsberger, P., & Ewees, A. A. (2020). Efficient classification of white blood cell leukemia with improved swarm optimization of deep features. *Scientific reports*, 10(1), 1-11.
- [20] Mallick, P. K., Ryu, S. H., Satapathy, S. K., Mishra, S., Nguyen, G. N., & Tiwari, P. (2019). Brain MRI image classification for cancer detection using deep wavelet autoencoder-based deep neural network. *IEEE Access*, 7, 46278-46287.
- [21] Kumar, S., Sharma, B., Sharma, V. K., & Poonia, R. C. (2021). Automated soil prediction using bag-of-features and chaotic spider monkey optimization algorithm. *Evolutionary intelligence*, 14(2), 293-304.
- [22] Nejad, M. B., & Shiri, M. E. (2019). A new enhanced learning approach to automatic image classification based on Salp Swarm Algorithm. *Computer Systems Science and Engineering*, 34(2), 91-100.
- [23] Wei, B., Hao, K., Gao, L., Tang, X. S., & Zhao, Y. (2020). A biologically inspired visual integrated model for image classification. *Neurocomputing*, 405, 103-113.
- [24] Jothi, G. (2016). Hybrid Tolerance Rough Set–Firefly based supervised feature selection for MRI brain tumor image classification. *Applied Soft Computing*, 46, 639-651.
- [25] Yang, J., Govindaraj, V. V., Yang, M., & Wang, S. H. (2020). Hearing loss detection by discrete wavelet transform and multi-layer perceptron trained by nature-inspired algorithms. *Multimedia Tools and Applications*, 79(21), 15717-15745.
- [26] Badgujar, R. D., & Deore, P. J. (2019). Hybrid nature inspired SMO-GBM classifier for exudate classification on fundus retinal images. *IRBM*, 40(2), 69-77.
- [27] Holland J (1975) *Adaptation in natural and artificial systems*. The University of Michigan Press, Ann Arbor
- [28] Mitchell, M., Forrest, S., & Holland, J. H. (1992, December). The royal road for genetic algorithms: Fitness landscapes and GA performance. In *Proceedings of the first European conference on artificial life* (pp. 245-254). Cambridge: The MIT Press.

BERLin: Bidirectional Encoder Representations from Linformer

Sandeep Kumar

Associate Professor, Computer Science Department, Maharaja Surajmal Institute of Technology, New Delhi,
India sandeep.jaglan@msit.in

Jerin Verghese

Student, Computer Science Department, Maharaja Surajmal Institute of Technology, New Delhi, India
vjerin99@gmail.com

Ritesh Dutta

Student, Computer Science Department, Maharaja Surajmal Institute of Technology, New Delhi, India
duttaritesh99@gmail.com

Abstract : Traditional Transformer models have proven to be very accurate but are slow in pre-training due to the time complexity being $O(n^2)$ regarding the length of the sequence. Linformer is a type of Transformer has a time complexity of $O(n)$ and works at par with a traditional Transformer. In this paper, Linformer is used in a famous language model, BERT, instead of the existing transformer. In this work the self-attention system of BERT is modified. This modification results in a significant decrease in pre-training time while retaining the state-of-the-art accuracy of BERT.

Keywords: BERT, transformers, Linformer, NLP, time complexity, Natural Language Processing, Google

1. Introduction

Bidirectional Encoder Representations from Transformers (BERT) is a method published by researchers at Google AI Language. It uses transformers to pre-train language models and fine-tunes them for number of Natural Language Processing (NLP) tasks, including, Natural Language Inference, Question Answering etc. to produce state-of-the-art results [1].

The main innovation used in BERT is applying the ‘bidirectional training’ of Transformer, to language modeling. This is different from previous methods which looked at a sentence either from left to right or used both right to left and left to right training for the Transformer encoder. The paper’s results show that a language model that is bi-directionally trained can have a deeper sense of context and flow than single-direction language models [1]. But to get state of the art results time complexity is compromised. The attention mechanism of BERT takes $O(n^2)$ time with regarding to the sequence length which makes it highly accurate but slow. This is because the transformer uses a self-attention mechanism that compares each word (token) with every other word in the sequence. This maintains long term information and allows parallelization which makes transformers better than all current language models. But due to these transformers need high-end GPUs to pre-train and they take a lot of time and space.

Several attempts have been made to reduce time and space complexity. Some of them are:

- Z. Dai et al. ‘Transformer-xl: Attentive language models beyond a fixed-length context’ [5]. This paper aims to overcome the fixed-length context of Transformers in the setting of language modeling. The design they propose empowers learning reliance past a fixed length without meddling with transient communication. It uses information from the previous segment when processing the next fragment, allowing the next fragment to have a longer context. Fragments are fixed and cached. Gradient remains with the current segment only. With a recurrent mechanism applied to every 2 consecutive segments, it creates a segment level recurrence. It also resolves context fragmentation problems. Accordingly, Transformer-XL learns relations that are 80% longer than RNNs and 450% longer than customary Transformers, accomplishes better execution on both short and long successions, and is up to 1,800+ occasions quicker than common Transformers during assessment.
- Z. Dai et al. ‘Xlnet: Generalized autoregressive pre-training for language understanding’ [6]. This paper means to beat the deficiency of the BERT language model. BERT disregards reliance between the covered lattices and experiences a pre-train finetune error. It proposes a summed up autoregressive pretraining strategy that empowers learning bidirectional settings by boosting the normal probability over

all changes of the factorization request and beats the impediments of BERT on account of its autoregressive detailing. It additionally utilizes a few ideas from Transformer-XL in pretraining. It outflanks BERT on 20 NLP errands, regularly by a huge degree, under specific conditions.

- N. Kitaev et al. 'Reformer: The efficient transformer' [7]. This paper proposes strategies to work on the existence intricacy of existing enormous transformers. They utilized area touchy hashing and diminished the transformer's intricacy from $O(L^2)$ to $O(L \cdot \log(L))$, where L is the succession length and utilizes reversible leftover layers rather than the standard residuals, which stores enactments just a single time in the pre-preparing process rather than N times, where N is the quantity of layers. Thus its presentation is comparable to the Transformer however is a lot quicker and effective.

2. Linformer Self Attention

The Linformer is a model proposed by Sinong Wang, Madian Khabsa, Belinda Z. Li, Hao M., and Han Fang in their paper [3], in which they try to propose a model which brings down the complexity of attention of the transformer model to linear time and space from the normally quadratic complexity[3]. This is done by projecting the values and keys vectors to a smaller dimension vector essentially from an $N \times N$ matrix to an $N \times K$ matrix and by selecting a suitable value of K the complexity of the computation becomes linearly dependent on N . The idea behind this was that the attention matrix is approximately a low-rank matrix and it can be approximate to a matrix of smaller dimension without losing the data, this was proved by them in the paper empirically.

For taking the projections we constructed a function which outputs a linear layer with learnable parameters that maps the input matrix of size $N \times N$ to a matrix of $N \times K$.

We pass the keys and values from the input matrix through the linear layer which allows us to get the resultant $N \times K$ projections of key and value.

3. Implementation

Linformer in Bert

The BERT model in its core is also a transformer encoder and it uses attention for the generation of sentences, therefore it also calculates attention using $O(n^2)$ complexity and the improved attention mechanism of the Linformer model which uses the low-rank property of the attention matrix can be used for the BERT model also. Also using the Linformer attention mechanism would not have any impact on the bidirectionality of the BERT model since it does not depend on how attention is calculated.

Changes in masking: Normally in the BERT model or rather in any transformer which uses masking, the masking is done usually on the matrix A , where A is as given in the below equation

$$Attention(Q, K, V) = softmax\left(QK^T \frac{\tau}{\sqrt{d_k}}\right)V \quad (Eq. 1)$$

$$A = QK^T \quad (Eq. 2)$$

Here, Q = Queries, K = Keys, V = Values and d_k = Dimension of keys.

But in the case of Linformer we are doing transformations on the K and V matrix before calculation of matrix A , therefore the shape of the matrix A also changes, so we cannot mask the matrix A because the input mask is dependent on the input size N .

Therefore, for the modified attention mechanism, we have masked the K and V vectors instead before transforming them. We found that this did not cause any significant effects on attention mechanisms or the masking.

Using learnable parameters for projection matrix: Making the parameters or the weights of the projection matrix trainable helps increase the accuracy of the model as it adds more learnable parameters. We also found that this did not cause any reduction in the time efficiency of the Linformer mechanism.

Head-wise sharing of projection: Head-wise sharing of the projection matrix is one of the sharing methods described by the authors of Linformer, we have also used this in the BERT model. Essentially this means using the same projection matrix for all the attention heads. This also increases efficiency.

K-V sharing of projection: K-V sharing is also another method for sharing projections proposed by the authors of Linformer, this involves using the same projection matrix for the transformation of both the K and the V matrices. We found that using this method resulted in a reduction of time for every iteration in pre-training and training of the model.

Pre-Training Data

For pre-training, we have used the wikitext-2 dataset because this dataset has continuous related sets of sentences that are required for Next Sentence Prediction. Pre-training the model with next sentence prediction and LM masking is important to maintain the bi-directionality of BERT. Wikitext-2 is a relatively smaller dataset, we chose to use wikitext-2 because it can be pre-trained easily and since our primary objective was not to train a good model but to compare efficiencies of the models. Wikitext-103 is a much larger dataset and can be used to obtain better results in terms of accuracy and perplexities.

Fine-Tuning and Training Data

After pre-training the model on the wikitext data we have fine-tuned it by adding a linear layer with tan_h activation. This layer can be trained to act as a classifier. After fine-tuning it we trained the model on Glue MRPC data, Glue is a well-known NLP benchmarking dataset that has been used to benchmark many NLP algorithms, the original BERT being one of them.

4. Results

To verify the efficiency of our model we made two models BERT without and with Linformer and compared their times and losses. We pre-trained it on a Wikipedia text dataset with masked LM and next sentence prediction to maintain the bidirectionality of both the models, after that we trained both the models on the Glue MRPC data after fine-tuning them.

Table1: Pretraining (25 epochs)

Models	BERT	BERLin
Avg. iterations per second	1.32	1.70
Total Time for pretraining	6 hrs	4 hrs 30 mins

Table 2: Training (3 epochs and 1140 iterations)

Models	BERT	BERLin
Avg. iterations per second	1	1.06
Loss	0.628	0.631

Table 3: Inference

Models	BERT	BERLin
Iterations per second	8.34	10.75

- For pre-training, the BERLin model was 1.3 times more efficient than the original BERT model.
- For training, the BERLin model was 1.06 times more efficient than the original BERT model.

- In terms of inference, BERLin was 1.3 times more efficient than the original BERT model.

5. Conclusion

Based on the result we can conclude that ‘BERLin’ boosts the efficiency of the original BERT model while maintaining accuracy. BERT being a state-of-the-art model is used for many NLP tasks and is one of the most used NLP models, BERLin can be integrated into many of those tasks for better efficiency. Also, we believe the idea behind our model can be used to improve other BERT-based models also like RoBERTa.

6. References

- [1] J. Devlin, M.W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding”, arXiv preprint arXiv:1810.04805, 2018.
- [2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, and Ł. Kaiser, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, pp. 5998–6008, 2017.
- [3] S.Wang, B. Li, M. Khabza, H. Fang, and H. Ma, “Linformer: Self-Attention with Linear Complexity”, *arXiv preprint arXiv:2006.04768*, 2020.
- [4] C.S. Perone, R. Silveira, and T.S.Paula, “Evaluation of sentence embeddings in downstream and linguistic probing tasks”, arXiv preprint arXiv:1806.06259, 2018.
- [5] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q.V. Le, and R. Salakhutdinov, “Transformer-xl: Attentive language models beyond a fixed-length context”, arXiv preprint arXiv:1901.02860, 2019.
- [6] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q.V. Le, and R. Salakhutdinov, “Xlnet: Generalized autoregressive pretraining for language understanding”, *Advances in neural information processing systems* (pp. 5753-5763), 2019.
- [7] N. Kitaev, Ł. Kaiser, and A. Levskaya, “Reformer: The efficient transformer”, *arXiv preprint arXiv:2001.04451*, 2020.
- [8] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners”, *OpenAI Blog*, 1(8), 9, 2019.
- [9] K. Ethayarajh, “How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings”, arXiv preprint arXiv:1909.00512, 2019.
- [10] E. Maha, B. Laurent, V. Jakob, “Pervasive Attention: 2D Convolutional Neural Networks for Sequence-to-Sequence Prediction”, arXiv preprint arXiv:1808.03867, 2018.

An IoT based 4-Wheeled Safety Monitoring System

Harsh Yadav¹, *Gunjan Chhabra²

¹B. Tech, School of Computer Science, UPES, Dehradun

²Assistant professor, School of Computer Science, UPES, Dehradun.

¹ramharsh5400@gmail.com ²gchhabra@ddn.upes.ac.in

Abstract: Emerging technologies making motor vehicles smarter and upgraded with humanly features. With these new technologies, made easier to develop vehicles smarter with advance safety features. Besides road safety rules, safety alert systems, awareness programs and road surveillance systems, still road accidents deaths are increasing frequently and subsequently. The analysis report shows that many of these occurred due to lack of following the safety guidelines like over-speeding, not using seat belts, wrong side driving and so on. In four-wheeled vehicles, it is enforced to buckle up the seat belts to avoid life damages, and vehicles comes up with the feature of alarm generation to ensure the same. Still people are capable of forge these alerts systems and drive without wearing seat belts. Hence, to fill up this gap, the article represent an Internet of Things (IoT) based safety monitoring system mounted with a micro-camera, which is capable of ensuring the presence of seat belts on both driver and co-driver. In case to its unavailability, IoT device will discontinue the power supply and stop the ignition, until the seat belt is not buckled up by the users. In addition to that, system is also capable of monitor the yawning and eye movement to generate the required alerts ensuring that driver is in conscious condition to drive safely or not. Collectively, generating a real time monitoring system for the safety of four wheeler users. Further, these features can be embedded along with the driverless transportation vehicles ensuring the safety for end users.

Keywords: Internet of Things (IoT), Smart Vehicles, Real Time Monitoring, Face detection, Road Safety

1. Introduction

Technology is now embedded with in all the sectors and one of the major research area for the same is automobiles. Introducing smart features in vehicles are accepted and attracts lot many end users. Now users are also aware and concerned about the safety features therefore more attention is given on the safety features of the product. The past records says that approx. 90K deaths occur due to road accidents and more than 60% are due to not following safety protocols. Hence, now it's time to add some more strict smart features to be introduced inside a vehicle to reduce this death rate.

Road safety is a major concern, especially in case of four wheeled vehicles. The road accidents are increasing very frequently and the reason behind the same is not using safety measures. People are not wearing seat belts while driving, over speeding, tried but still driving, drink and drive and many more. Many strict laws are imposed but still the number of accidents and deaths caused by these accidents are increasing every year. With the advancement in technology, many smart features are imposed and now a days smart vehicles are high in demand. Many alert based features like alert for over-speeding, non-wearing of seat belts etc. are in use by the renowned automobiles brands but still, on the other side, end users are able to forge these systems. People can easily bypass these alerts by doing minor modifications. Thus, there is requirement of more advanced technological systems to overcome these issues. This article highlights the implantation of new technique that may help vehicle driver to stay safe without forging the system.

2. Related Work

This section elaborates on previously done research work under seat belt detection problem. This problem is somewhat similar to the object detection in computer vision. Many researchers has solved this issues by using machine learning and made huge progress. Wieli et al. [13] proposed a seat belt classifier using Adaboost have capacity of boosting the decisions from multiple weak classifier. The features are extracted via Haar-like features etc. HSV color model is employed for vehicle detection [14-15]. In research [16-17], enhanced Adaboost model was proposed to detect and recognize objects in much faster way. In [4], researcher proposed fully convolution networks for object detection and classification. In addition to that single shot multibox detector (SSD) based neural networks are also used for better results in object detection.

Many researchers in [6], has used YOLO neural networks for fast and better outcomes. Later on YOLO v3 has been released which works with the accuracy of single shot multibox detector algorithm but it is three times faster. Authors in [8] proposed CNN based network based on extracting multi-scale features and optimized the network by using support vector machine technique. Other researcher [15] used SSD and CNN combined to detect the seat belt on passenger, researcher in [13], used edge detection based algorithm along with HSV color model for the verification of presence of seat belt. Another model proposed by [14] is based on improved version of CNN, BN-AlexNet give accuracy of approx. 92% which is was quite high.

Authors of paper [10] came up with another new CNN architecture (NADS-Net) proposed for both pose estimation and seat belt detection inspired by the Feature Pyramid Network (FPN) [11]. The network was tested on more than 50 driving sessions including different illumination conditions. But unfortunately, the outcomes was not so accurate enough. The researcher in [20] proposed a method by dividing the problem into three stages that includes classification of the vehicle, secondly driver and passenger position and lastly seat-belt feature includes grayscale and texture analysis, edge detection and gradient orientation to make final decision. Many more researchers have contributed in the domain to proposed better model and most of the work has been done using road CCTV cameras. In this research, a micro-camera is implanted inside the car and used to detect the seat-belts of both driver and co-driver. An IoT device is also installed to control the ignition of the vehicle so that in case of regular alerts, IoT device will switch-off the ignition. The detailed process of the whole system is elaborated in the coming sections.

3. Proposed Methodology

In context to the highlighted problems and issues, a methodology has been developed as a solution which is based on the hybridization of latest technologies. This methodology includes Artificial Intelligence and its sub-domains with IoT device to give the better results and extra safety. The methodology includes video processing mechanism in real time. The main objective is to detect the seat belt availability along with the present condition, drowsiness and yawning, of the driver and co-driver. As illustrated in figure 1, there are two functionalities to improve the overall safety of the user. The algorithm designed is also have capability to detect the alertness of the driver. By capturing the real time frames, algorithm will classify the status of the user and generate the alarm accordingly. In addition to the alarming system, ignition controlling system is also included to enhance the overall safety feature of the vehicle.

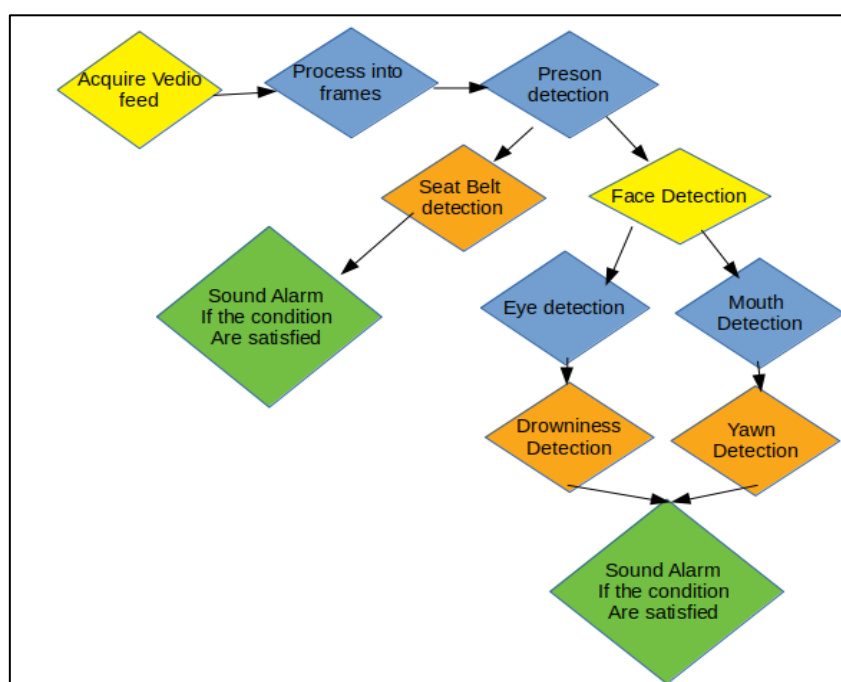


Figure 38 Proposed Methodology

A micro-camera is embedded on the dashboard of the four wheeler for real time monitoring. The night vision camera is also capable to capture the video of both front seaters and is connected with IoT device controller. The IoT device is connected with the internal mechanical system of the vehicle and it is able to communicate with mobile device network. For the communication purpose GSM module is available to carry the signals. The data is stored and processed on cloud environment, as shown in figure 2. The importance of IoT device is that it will control the ignition system of the vehicle i.e. in case seat belts are not worn properly, IoT enabled device will not allow the engine to start. Hence, making it as a mandatory event for the user to wear the seat belts. To enable and disable the ignition process, relay switch is available and it will work as per the command received from the controlling device.

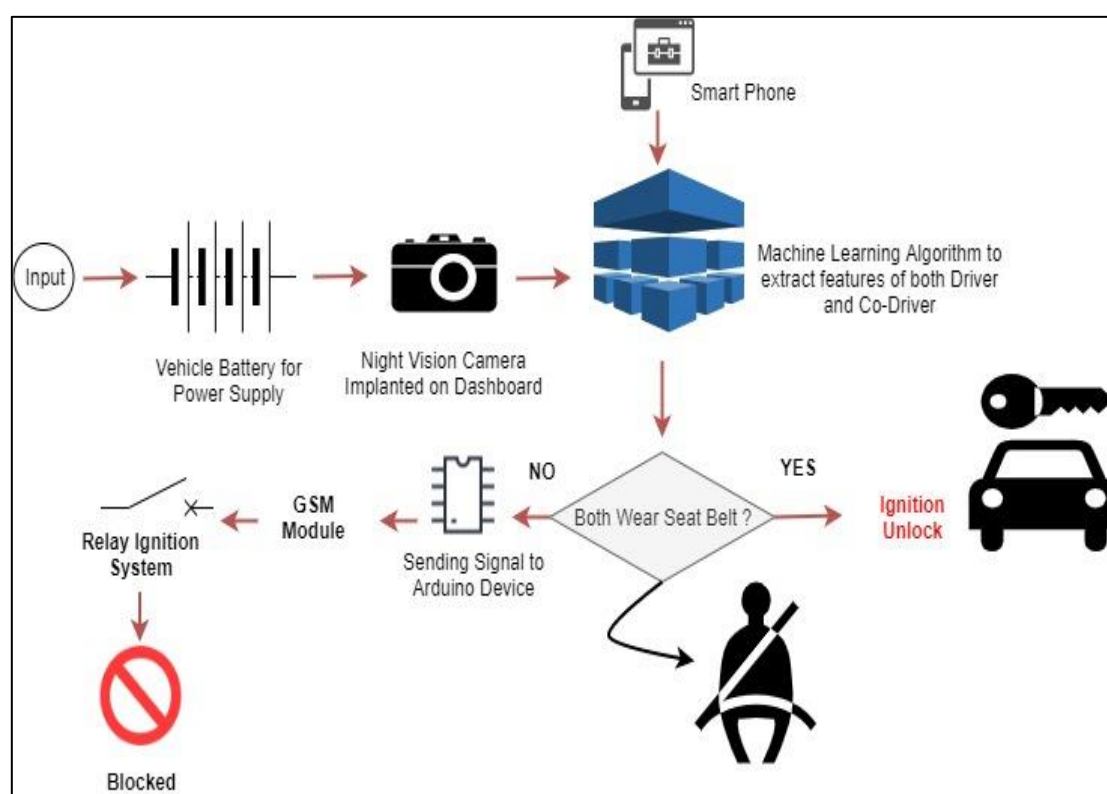


Figure 39 :Overall working of the proposed method

The entire process includes two sub-objectives, one is related with seat belt and second is drowsiness and yawning detection. The computation steps behind the whole implementation shown in the figure 3 below, which is a continuous process. The environment is very dynamic in nature hence need lot of computation for better outcome. Therefore the computing is performed on the cloud environment, linked through mobile network application.

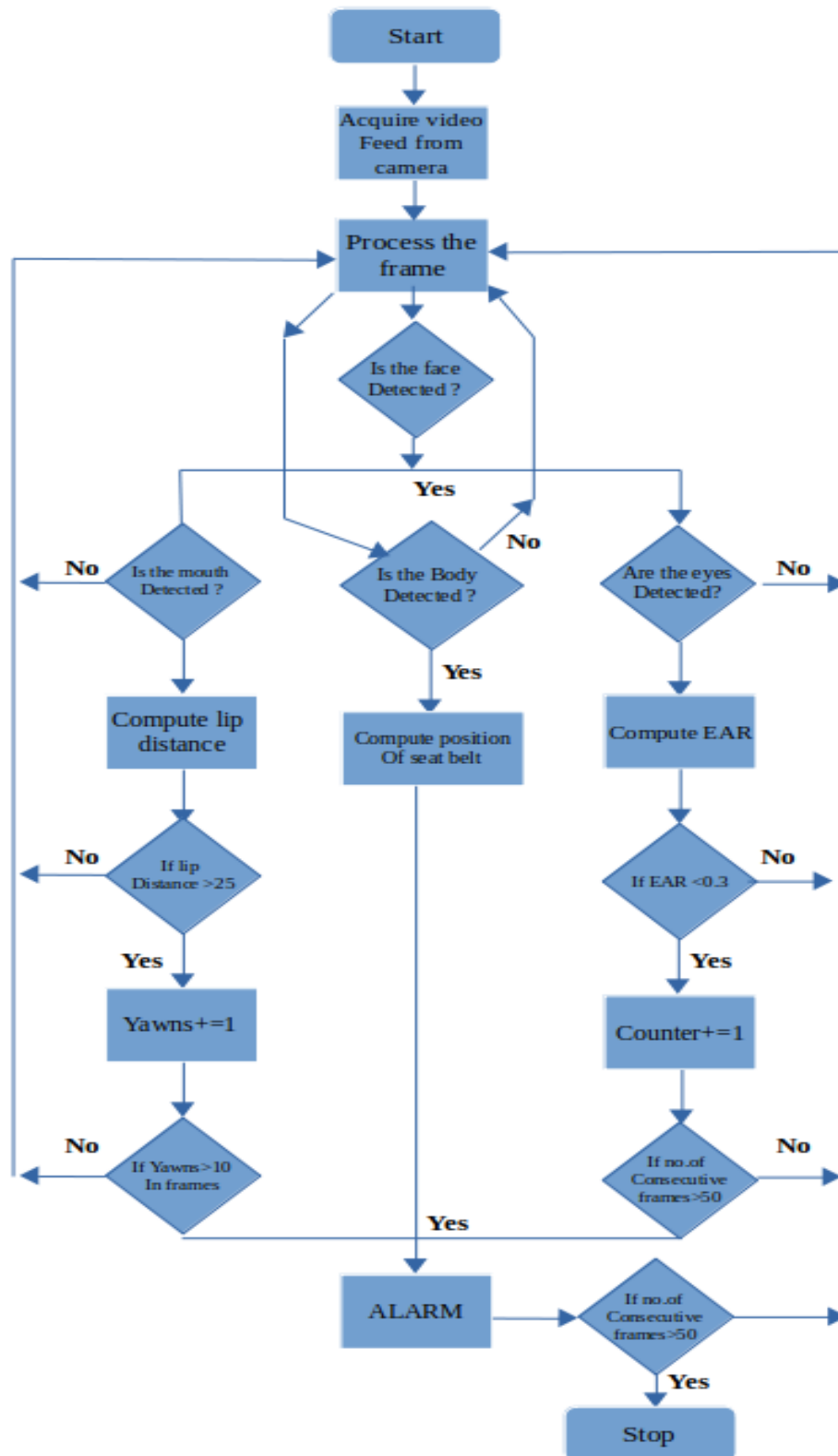


Figure 40 Working flowchart

For the detection of drowsiness and yawning status of the driver, facial feature extraction method using machine learning is used for accuracy. The flowchart illustrated in figure 4 shows the step by step procedure for face detection, its computation and other intermediate steps. Face detection method involves the machine learning technique and features are extracted using eigen values. After the detection of such an activity an alert alarm will be enabled to save the user from any incident or mishappenings.

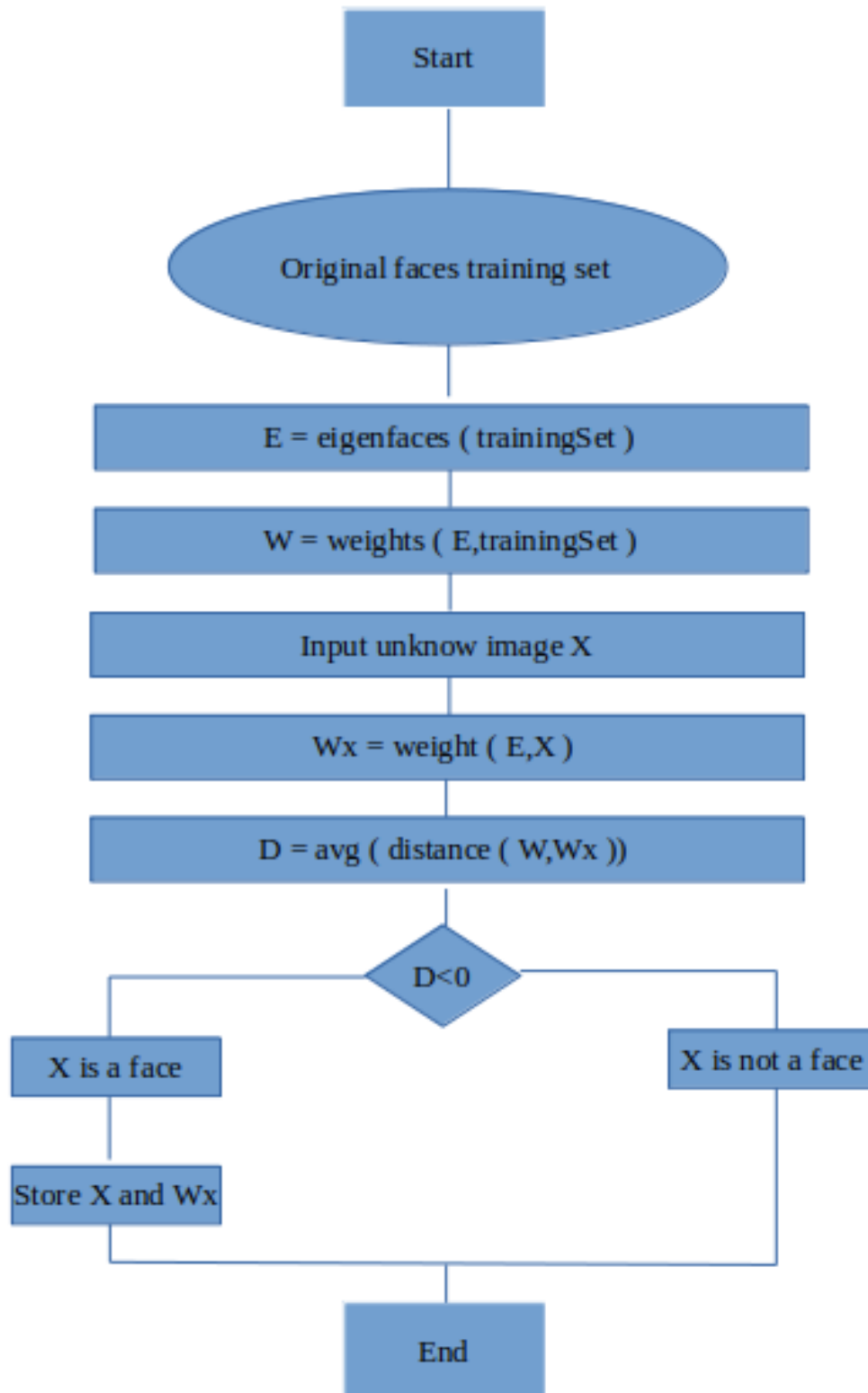


Figure 41 Flowchart for face detection

Pseudo Code

Input : Vedio(25_Frame_Per_Second)

Ouput : Facial_Activity(Eyes_open, Mouth_close, Eyes_close, Mouth_Open>4Seconds),
Seat_Belt(Wear_in_prope_way, Not_Wear_in_prope_way)

```

for lines <- loop line by line
    s=slope(x1,y1,x2,y2)
    calculate slope of line using slope Equation1
    update s = Equation1

    # If Current Line's Slope Is Greater Than 0.7 And Less Than 2
    if s > 0.7 and s < 0.2 then
        # And Previous Line's Slope Is Within 0.7 To 2
        previous line = ps
    if ps > 0.7 and ps < 0.2 then
        # And Both The Lines Are Not Too Far From Each Other
        if (x1 - px1) > 5 and x2 - px2 > 5 or (y1 - py1) > 5 and (y2 - py2) > 5
            output = Wear_in_prope_way
        else
            output = Not_Wear_in_prope_way
        end_if
    end_if
end_for
for i <- 1 to num of sub window's shift do
    for j <- 1 to num of cascaded stages do
        for k <- to num of filter do
            Calculate We face using equation
            update W-face using Equation2
            update W-facial_activity = Equation2
            if We < W-face then
                break for k loop
            elseif We < W-facial_activity then
                output = Eyes_close, Mouth_Open>4Seconds
            else
                output= Eyes_open, Mouth_close
            end_if
        end_for
    end_for
end_for

```

The complete process of seat belt detection and drowsiness and yawning alert system is designed to overcome the issues related to road accidents. Internet of things and machine learning technologies brought up with new emerging field of development known as internet of vehicles. Hence, the application stated in this article can be used in future technologies for the safe driving experience.

4. Results & Conclusion

This section showcase the outcomes obtained and concluding remarks of the experimentation. The objective of the research is to provide a solution for safe driving experience. Many road accidents causes death due to the absence of seat belts, drink and drive cases and many other circumstances. This research article gives a solution that will switch off the ignition of the engine whenever any emergency alert is detected. The accuracy of the present work is 92%. The figure 5 shown below illustrate the detection of seat belt on driver. In opposite case, if the seat belt is not properly worn by the driver, an alert will be generated and ignition of the vehicle will get automatically off for the safety.

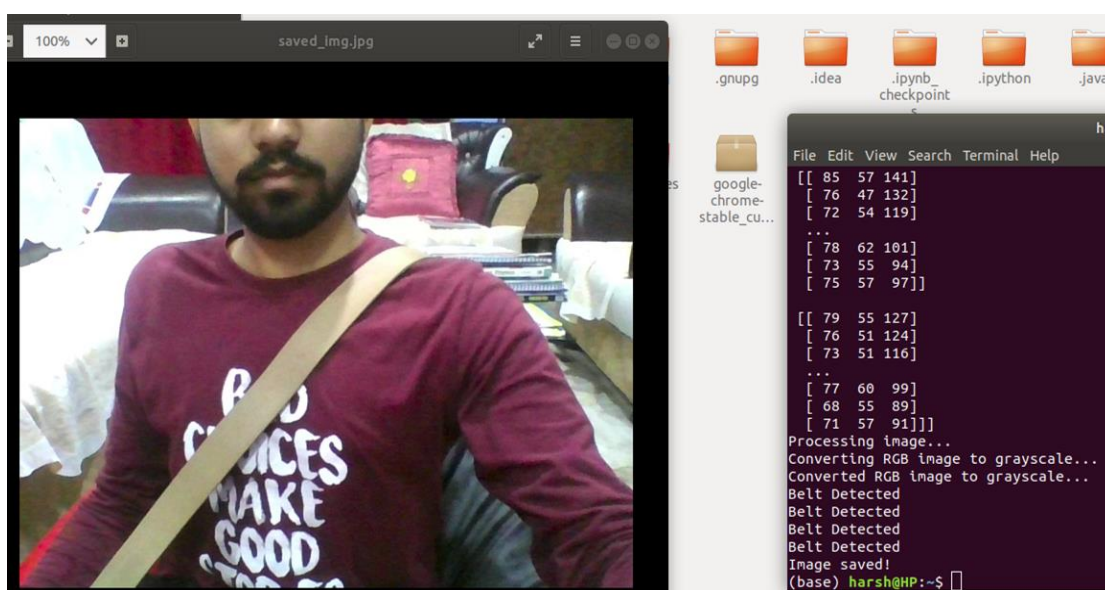


Figure 5 Illustrate successful detection of seat belt

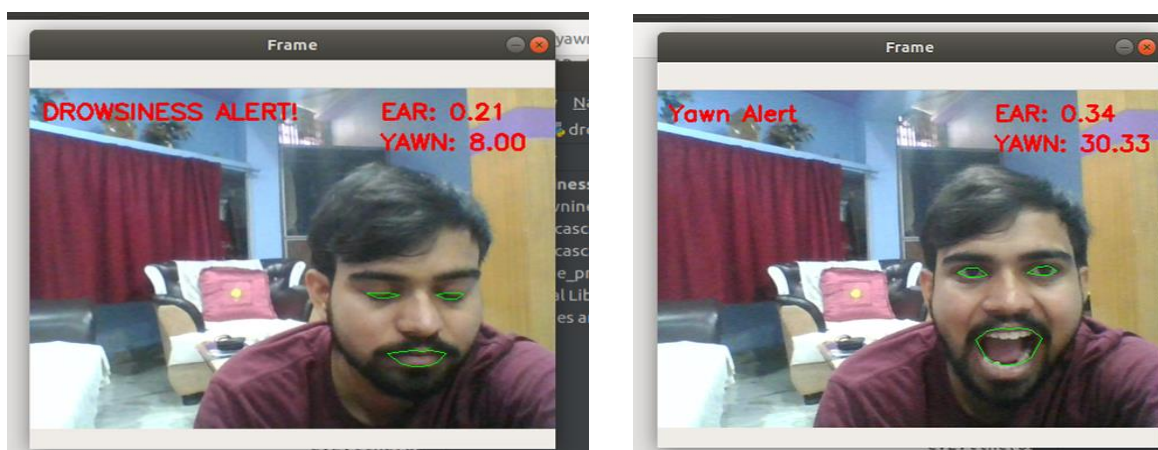


Figure 6 Drowsiness alert (left) and Yawning alert (right)

In the second phase, the face features being captured by the micro camera on regular intervals. The figure 6, shown above, provides the results related to the drowsiness and yawning alerts for the driver to safeguard. Previous research shows number of road accidents occurred due to this reason. While, a driver is over drunk falls under this scenario. The proposed system will generate an alert on regular basis. In any case, the driver caught by the system under same situation, the alert will be passed to the microcontroller; hence the ignition will get an interrupt, resulting the engine gets off. In both the scenarios, the vehicle's ignition gets off to safeguard the driver and the passengers (if any).

In future, other advance techniques including concepts of machine learning, intelligent computation, automation etc. can be incorporated to enhance the results. Also, one can add wearable sensors for real-time monitoring of the driver's health and at accordingly to save the life. Internet of vehicles and autonomous industries came up with many more features for safe and better driving experience that will reduce rate of deaths on road.

References

- [1] Guo, H., Lin, H., Zhang, S., & Li, S. (2011, July). Image-based seat belt detection. In Vehicular Electronics and Safety (ICVES), 2011 IEEE International Conference on (pp. 161-164). IEEE. <https://doi.org/10.1109/ICVES.2011.5983807>.
- [2] G.Y. Song, K.Y. Lee, J.W. Lee, Vehicle detection by edge-based candidate generation and appearance-based classification, in: Intelligent Vehicles Symposium, IEEE, 2008, pp. 428–433. <https://doi.org/10.1109/IVS.2008.4621139>.
- [3] B. Sun, S. Li, moving cast shadow detection of vehicle using combined color models, in: Pattern Recognition, IEEE, 2010, pp. 1–5 <https://doi.org/10.1109/CCPR.2010.5659321>.
- [4] BOERNSTEIN W. S. (1955). Classification of the human senses. The Yale journal of biology and medicine, 28(3-4), 208–215.
- [5] Bojarski, M., Yeres, P., Choromanska, A., Choromanski, K., Firner, B., Jackel, L., & Muller, U. (2017). Explaining how a deep neural network trained with end-to-end learning steers a car. arXiv preprint arXiv:1704.07911.
- [6] What is backpropagation really doing? | Deep learning, chapter 3, 3BLUE1BROWN SERIES S3 E3 (2019) <https://www.youtube.com/watch?v=Ilg3gGewQ5U>.
- [7] Rio-Alvarez, A., de Andres-Suarez, J., Gonzalez-Rodriguez, M., Fernandez-Lanvin, D., & López Pérez, B. (2019). Effects of Challenging Weather and Illumination on Learning-Based License Plate Detection in Noncontrolled Environments. Scientific Programming, 2019. <https://doi.org/10.1155/2019/6897345>.
- [8] El-Melegy, M. T., El-Magd, K. M. A., Ali, S. A., Hussain, K. F., & Mahdy, Y. B. (2019, February). Ensemble of Multiple Classifiers for Automatic Multimodal Brain Tumor Segmentation. In 2019 International Conference on Innovative Trends in Computer Engineering (ITCE) (pp. 58-63). IEEE. <https://doi.org/10.1109/ITCE.2019.8646431>.
- [9] Sun, W., Du, H., Nie, S., & He, X. (2019). Traffic Sign Recognition Method Integrating Multi-Layer Features and Kernel Extreme Learning Machine Classifier. <https://doi.org/10.32604/cmc.2019.03581>.
- [10] Ershadi, N. Y., Menéndez, J. M., & Jimenez, D. (2018). Robust vehicle detection in different weather conditions: Using MIPM. PloS one, 13(3), e0191355. <https://doi.org/10.1371/journal.pone.0191355>.
- [11] Pavel Surmenok, “ResNet for Traffic Sign Classification with PyTorch”. Twords data science, Nov 2018.
- [12] A. Khammari , F. Nashashibi , Y. Abramson , C. Laurgeau , Vehicle detection combining gradient analysis and Adaboost classification, in: Intelligent Transportation Systems Conference, IEEE, 2005, pp. 66–71
- [13] Li, W., Lu, J., Li, Y., Zhang, Y., Wang, J., & Li, H. (2013, December). Seatbelt detection based on cascade adaboost classifier. In 6th IEEE International Congress on Image and Signal Processing (CISP2013), (Vol. 2, pp. 783-787) <https://doi.org/10.1109/CISP.2013.6745271>

Blockchain based identity management system for food supply chain

Dr. Anuradha Konidena¹, Neha Bhatia²

¹Associate Professor, Department of CSE, College of Engineering and Technology, IILM Academy of Higher Learning, Greater Noida, U.P, India.

Assistant Professor, Department of Information Technology, Panipat Institute of Engineering and Technology, Panipat, Haryana, India.
akonidena75@gmail.com

Abstract: Proof of one's identity and its management have been considered very crucial in all online transactions and is key element in authorization and authentication. In the digital era of internet of things, existing identity management systems lacks key quality of service parameters like scalability, inter-operability and mobility. Yet these identity services are expected to ensure privacy of the users with flexibility in its operation. This paper describes significance of digital identity and its management i.e., creation, updating and revoking using block chain technology.

Keywords: Authentication, cloud computing, Internet of Things, Block chain, digital identity management

I. Introduction

In the digital era, information is wealth. Many people do share lots of information via emails and other social media platforms. As a consequence of data creation, copy, usage, forwarded, liking; amount of data is in hundreds of zettabytes. It required lot of computational elements and storage. It calls for cloud computing services. Cloud computing provides anywhere-anytime access to data. With the diversified applications of information technology, an immense growth in volume, velocity and variety of the information is being witnessed. These data are heterogenous in nature and also being shared on public networks which poses security problems. Under such circumstances, financial applications require a platform that provide, optimal quality of services to end users. To deal with financial data, public ledger networks were outlined in 1991.

The Properties of Distributed Ledger Technology (DLT)

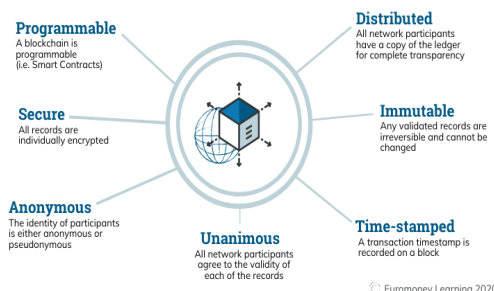


Fig1: Blockchain technology

(Source:

[https://www.euromoney.com/learning/~media/0E855B86EDF04F3C8EABAF42917C8C6.png?la=en&hash=B67568B63CBB2C0C7311DE742F5B9E48E86DC8B9\)](https://www.euromoney.com/learning/~media/0E855B86EDF04F3C8EABAF42917C8C6.png?la=en&hash=B67568B63CBB2C0C7311DE742F5B9E48E86DC8B9))

It prevents tampering of timestamping of documents. It operates in blocks, wherein each block is linked with a cryptographic algorithm to provide more security, as each block is encrypted with hashes. These are robust and scalable networks and prevents hackers to exploit vulnerability. Each block consists of information like block number, block id of current block, block id of previous block, time stamp, Merkle tree root hash, size of the block and transactions list in the block [1]. Nodes of the blockchain are taken by various organizations. Communication between nodes is dependent on ledger content. Transaction requests sent by users, as received by blockchain will perform the requested operation. As a result of this transaction, one or more

ledgers will store record of the transactions, that will never be updated. These block chains are empowered by asymmetric key cryptography, wherein private key is used to check the signature on transaction and its compatible public key checks signature generated by private key. It averts altering of data.

II. Literature Review

Evolutions of information technology, revolutionized various aspects of human life: ranging from managing day-to-day entertainment to healthcare, socializing to financial transactions. Digital identity is crucial, but it is cumbersome and challenging. Semantic based authentication mechanism used in cloud computing reduces access control, but requires checking of operations every time, whenever a user places a request [2].

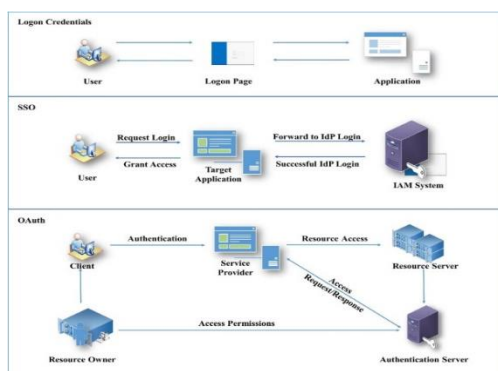


Fig. 2. Comparison of different authentication mechanisms in cloud environment.
(source: <https://www.sciencedirect.com/science/article/pii/S2215098617316750>)

With the use of encrypted session cookies, Enterprise Single Sign-On makes it easier to save and transmit user credentials across web-based apps. SSO generates a browser-based encrypted session cookie when a user is authenticated through a centralised authentication server. The new application checks for a valid session cookie as the user moves from one application to the next. The user is authenticated to the targeted application when the user's details are read from the valid session cookie. If an application can't find a valid session, the user is directed to the centralised authentication mechanism for re-authentication. There will be a new session created.

OAuth is an open standard for access delegation that allows users to access other websites/applications without revealing their passwords. OAuth 2.0 is the current version. In order to obtain client identifiers and client secrets, reliant parties (RP) must register with the OAuth token provider. When a user attempts to access a relying party, the user is forwarded to the identity provider for authentication. The identity provider shares the OAuth access token, refresh token, client id, and client secret with the user when they have been successfully authenticated. From the user, the relaying party obtains further user information.

Next session address authentication using blockchain technology.

III. Methodology

Traditional identity management systems like emails, user-ids, biometric based identify management systems requires user to first register themselves to obtain online services. And this practice is made mandate by all online service provides leading to creation of many digital identities, which at times becomes complex for human beings to remember. Moreover, internet of things is connecting millions of people, trillions of devices increasing sharing of sensitive information on public platforms.[4]. It is paving path for cyber attacks and need for stronger identity management techniques.

To construct and register a digital identifier (DID), users must first register with a self-sovereign identity and data platform. The user generates a set of private and public keys throughout this process. If a DID's public keys are hacked or need to be rotated for security reasons, they can be stored on-chain. To preserve scalability and compliance with privacy standards, additional data connected with a DID, like as attestations, can be anchored on-chain.

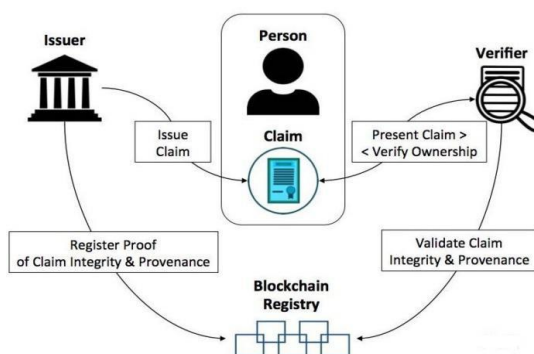


Fig3: Block chain based Identity Management

Techniques like location-based authentication and biometric based authentications are in place, because of their limitations like limited scope and copying of the sensitive information is defeating the purpose-compromising high value data, highly confidential data, highly sensitive data. Blockchain addresses identity theft management with the help of Decentralized identifiers. They are issued by peers by confirming specific details like location, social security number or any other known characteristic, but a unique one. Every entity on blockchain be it an organization or a person or any object should have a digital identifier to communicate with other identifier. These are stored and managed by DID (Decentralized identifier) owners themselves, rather relying on profile providers like Instagram or LinkedIn. Decentralized identities are secured by asymmetric key cryptographic algorithms. Encryption of the identifiers is performed by public key and it can be decrypted by its compatible private key only. That private key is known only to the owner. Requested services can be provided, once proof of control or ownership of the identifier is verified and found to be valid. These digital identifiers can be used for self-sovereign identity. With SSI, people or organizations can store their sensitive information on their personal devices, and apply minimalistic approach. That is they can decide what to share with whom. And as these identifiers can be created independent of nation, state or organization, it simplifies the process of Know Your Customer and reduces the time required in KYC. It aims streamlining and standardizing various schemes provided by nations to their citizens.

Cost implications associated with Blockchain technology are addressed by the authors of [5]. As per their understanding considerable costs are involved in key management and infrastructure. Implementation of Self-Sovereign Identity management involves cost in upgradation of processes and equipment, training costs.

IV. Digital identity management as a service:

Food supply chain management requires traceability, transparency, automated packing and shipping with minimized operational costs in addition to Customer's demand for details about the food like geographic source of the food, quality of the food, transit details, nutritional values and round the year availability.

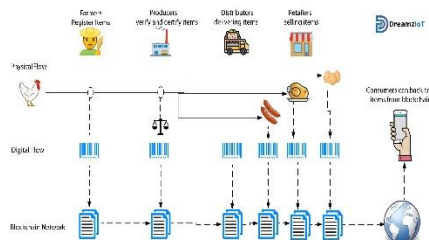


Fig4: Food supply chain using Blockchain

(Source: <https://dreamziot.com/wp-content/uploads/2018/11/Foodchain-Blockchain-3.jpg>)

Currently for the dissemination of this information to retailers and customers, cloud computing platforms are used by retailers and information can be obtained using websites or mobile applications or QR codes. Though cloud computing is advantageous to conventional FSC management in terms of more reachability and availability of information, it has its own challenges on its way like security of the information and various

regulations across the borders [3]. DID will address this issue. Additionally, food safety is grabbing utmost importance, as one 10 people per year fall ill and 42,000 dying globally due to food-borne diseases. So, food supply chain management should be transparent and effective, that can address issues related to data of food shelf -life. With the help of blockchain, suppliers can put the information about details of the food like: where it is being grow, how it is being grown/cultured, handled, stored and transmitted. Quality and shelf-life of food items can be assessed by applying various techniques of convolutional neural networks and distributed ledger technology(blockchain) can be used to store and disseminate the information thus obtained in a secure manner. IBM, Walmart, JD. Com, Tsinghua University are in collaboration with each other to give consistent user experience when accessing food safety information and traceability information. Usage of blockchain technology for safe and secure food supply chain management is at its inception stage and requires great deal of regularization of norms across the borders and lot of skill set that combines knowledge of food technology and blockchain technology.

According to authors of [6], block chain can be beneficial to food industry in various ways: Food traceability, Food supply chain collaboration, Food trading and in assessing Food supply chain efficiencies.

Food traceability is part of logistics management, according to [7]. It is concerned with the collection, storage, and transmission of information on food products at all stages of the FSC in order to facilitate food quality and safety control and ensure the backward compatibility of goods. Traceability is becoming more widely recognized as an important component of providing safe and healthy food as well as ensuring consumer happiness and trust.

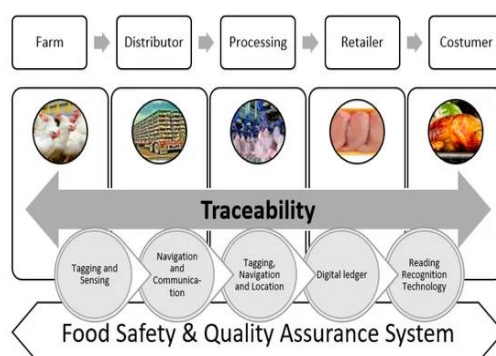


Fig5: Food traceability using blockchain

(Source: https://www.mdpi.com/sustainability/sustainability-12-03497/article_deploy/html/images/sustainability-12-03497-g001-550.jpg/)

When numerous exchange partners constitute specific FSC Interoperability challenges develop owing to non-compatible technologies required for information sharing. This issue can be addressed using Global Standard one (GS1) standards and help the FSC's various stakeholders communicate and collaborate with one another [8].



Fig6: Collaboration with Walmart using blockchain

(Source: <https://www.ibm.com/blockchain/>)

India's, one of the most important industry is Food processing industry, whose contribution is 32% of total food market of India. Food trading plays crucial role in economy for developing countries [9] like India, and can be efficiently managed by blockchain technology. Automatic reconciliation of invoices, trading details,

real-time tracking of food delivery can be safely and securely managed by blockchain technology [10]. Blockchain's capacity to cut logistical costs, expedite transportation processes, and legitimise trade ties can be beneficial to global FSC trading partners. Furthermore, blockchain has the potential to break down specific trade barriers and open up new market opportunities for businesses to expand, thrive, and become more competitive [11]

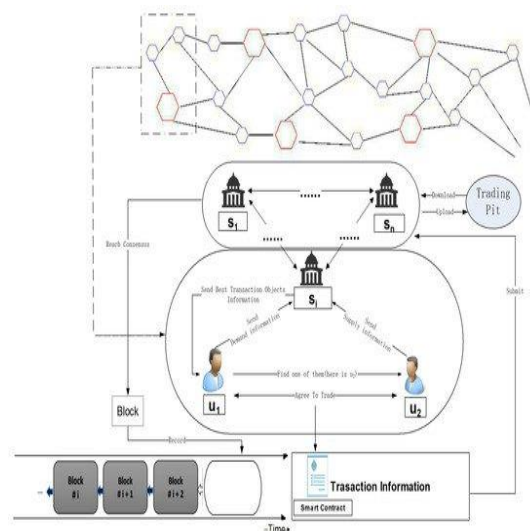


Fig7: Food trading using blockchain

Food trade is becoming increasingly reliant on blockchain technology. It could be a viable choice for farmers who are looking for a way to supplement their income. They feel forced to sell their products through marketing boards. The adoption of blockchain technology may be able to prevent this from happening [12]. It can also eliminate middlemen and cut transaction costs by using price coercion and retroactive payments.

Food supply chain using blockchain is effective in securely managing and transiting the data, but it has some limitations associated with it.

It is facing challenges like technological immaturity and reluctance to adapt new technology by organisations along with lack of sufficient skill force to scale and remove redundancies.

V. Discussion and Conclusion

Various decentralized and flexible block chain-based identity management systems like uPort empowers users to full control on creation and disclosure of IDs to other blockchains namely Ethereum and Bitcoin. Sovereign can be used as a platform for identity management of intermediate entities and support for other identity systems. This kind of self-sovereign identity management systems provides easy to use and implement identity management for individuals and organizations as well. Multiple modules of a product can be secured by using multi-factor authentication system -Wayport. Strongly secure digital identity provides authorities to implement welfare schemes to the correct person, secure payment for services and goods, enforcement of federal laws, easy to use healthcare applications etc. Though no mechanism is 100% safe and secure, block chain-based identity management system will address challenges being faced existing identity management systems. In this research paper we have described about identity management using block chain and its effectiveness in food supply chain. This research paper will be useful to those researchers who want to propose strong algorithms ranging from proof of identity to food trading in food supply chain. This work has some limitations as the content is retrieved based on keyword search from open access research material and peer reviewed journals, although has provided detailed content, but it might not be exhaustive. Future studies by Researchers, requires thorough reference to more articles from indexed databases.

References

- [1] Simanta Shekhar Sarmah, Application of Blockchain in Cloud Computing, International Journal of Innovative Technology and Exploring Engineering (IJITEE), ISSN: 2278-3075, Volume-8 Issue-12, October 2019
- [2] M. Auxilia and K. Raja, "A Semantic-Based Access Control for Ensuring Data Security in Cloud Computing", IEEE conference on Radar, Communication and Computing, 2012.

- [3] Sandeep Jagtap , Farah Bader, Guillermo Garcia-Garcia, Hana Trollman, Tobi Fadiji, Konstantinos Salonitis, Food Logistics 4.0: Opportunities and Challenges, Logistics 2021, 5, 2. <https://doi.org/10.3390/logistics5010002>.
- [4] Xiaoyang Zhu, Youakim Badr, Identity Management Systems for the Internet of Things: A Survey Towards Blockchain Solutions, December 2018, MDPI.
- [5] [Yang Liu](#), Debiao He, Mohammad S. Obaidat, Fellow of IEEE, Neeraj Kumar, Muhammad Khurram KhanKim-Kwang Raymond Choo, Blockchain-based identity management systems: A review, Journal of Network and Computer Applications
- [6] Abderahman Rejeb , John G. Keogh , Suhaiza Zailani , Horst Treiblmaier , Karim Rejeb, Blockchain Technology in the Food Industry: A Review of Potentials, Challenges and Future Research Directions, Logistics 2020, 4, 0027; doi:10.3390/logistics4040027.
- [7] Techane Bosona, Girma Gebresenbet, Food traceability as an integral part of logistics management in food and agricultural supply chain, [Volume 33, Issue 1](#), September 2013, Pages 32-48, <https://doi.org/10.1016/j.foodcont.2013.02.004>.
- [8] Ki-Hyun Um, Sang-Man Kim, The effects of supply chain collaboration on performance and transaction cost advantage: The moderation and nonlinear effects of governance mechanisms, International Journal of Production Economics

Crop Yield Prediction & Fertilizer Recommendation System

Suresh Chand Gupta¹, Shekhar Singh², Deepak Kumar³

¹²³CSE Department, Panipat Institute of Engineering and technology, Samalkha,
Panipat, Haryana, India,

¹sureshguptaji@yahoo.com, ²Shekhar.cse@piet.co.in, ³wadhwa123deepak@gmail.com

ABSTRACT: Agricultural sector is acting as an important contributor in every country's economic growth. Motivation for the current research came into existence due to the reason, that even in today's world; the farmers are still managing with manual techniques to predict yield production in advance. Main purpose of the current research was undertaken with the goal to help farmers in pre-yield prediction and recommendation of specific fertilizer for high yield production of one of the main staple food crop i.e. Wheat, Rice and Mustard. Machine learning based system is a decision support system for prediction crop yield; system also provide supporting decisions on what are the required fertilizer for crops to grow. Several ML techniques have been applied to develop crop yield prediction system. In this research, we proposed hybrid architecture for yield prediction and has been implemented with five most relevant and important ML Algorithms by integrating each of them with PCA. The experimentation was performed on the real time dataset collected from the directorate of agriculture department, talab tillo, Jammu under model village programmed 2019-20. The performance of these models has been evaluated on five performance measures such as recall, precision, specificity, f-score and accuracy. Experimental results predicted that KNN, naïve bayes and random forest gave better results when compared with other algorithms undertaken for the research.

I. INTRODUCTION

Agriculture sector is the primary source of livelihood and important contributor in economic growth which forms the backbone of our country. Wheat, rice, mustard, etc. are the most important crops sown in India. In order to meet the demand of agricultural production, it is needed for farmers to use digital technology and the adoption of new technologies in agricultural production but the farmers still manage to predict yield production in advance with manual techniques. The main purpose of this system is to assist farmers in pre-yield prediction from soil parameters and recommendation of specific fertilizers for high yield crop production. This system will help the farmers to know about the productivity of undertaken crops in advance. In case of low yield of undertaken crop, the farmers can use fertilizers to the soil to increase productivity of the crop. This system will also provide the information about the pesticides to be used for the crops if they get diseased. Also, prediction of crop yields depends on many factors such as soil fertility, water availability, climate and pesticide or fertilizer use. Soil fertility is one of the fundamental parameters which affect the health of any crop directly [3] [4] [5]. The proposed system "Crop Yield Predictor based on Machine Learning Techniques" will deal with yield prediction of the most commonly sown agricultural crops such as wheat, mustard, rice, etc. from soil parameters. This system will take soil nutrient parameters as input in order to predict crop yield as high, low or moderate by applying machine learning approaches. Objectives of the system are to carry out thorough research on soil nutrients affecting the crop yield, analyze machine learning techniques used for crop yield prediction, and compare the performance of machine learning techniques under research for crop yield prediction on the basis of different parameters.

II. LITERATURE REVIEW

By reviewing the relevant literature given in next section and problems related with the previous research, it is noticed that there is a requirement of selection of best suitable algorithms for crop prediction considering soil analysis which is one of the major parameter affecting crop yield. Bondre et al. (2019) proposed an approach based on machine learning algorithms to prediction of crop yield and fertilizer recommendation. This model enabled the analysts to recommend the fertilizer for crop productivity [3]. Lata et al. (2019) proposed models based on data mining techniques and machine learning for predication of crop yield [6]. Manimekalai et al. (2018) proposed an approach compute the soil parameters for prediction. In this approach crop yield prediction perform using neupper rule established algorithm [7]. Shastry et al. (2017) apply regression techniques in soil parameters for prediction of crop yield. They also used regression technique like linear regression and least square Median. They found least median squares regression produce better results for prediction than the classical linear regression technique [10]. Supriya et al. (2013) clustered the data using the K-means clustering algorithm is applied to classify the clusters. In this model implanted two data mining

techniques namely multiple linear regression and density based clustering techniques. The predict ants were 'Year', 'Rainfall', 'Area of Sowing', 'Yield', 'Fertilizers 'and the response variable was 'Production'[12]. Bhuyar (2014) proposed an approach where different classification algorithms were applied to soil data set to predict its fertility. This paper focused on classification of soil fertility rate using J48, Naïve Bayes, and Random forest algorithm. J48 algorithm gave better result (98.17% accuracy) than other algorithms. Decision tree formed by J48 algorithm helped the farmer and decision makers to identify the soil fertility rate and on the basis of nutrients found in the soil sample different fertilizers can be recommended [2].

During the review of literature it is noticed that machine learning algorithms play important role in crop yield prediction, improve the productivity and quality of the crops. Accurate prediction of crop yield plays an important role in crop production management [3] [4] [5]. So it is proposed to present a set of Machine Learning techniques, in order to predict the category yield of the primary analyzed soil dataset under model village program (2019-20) in Jammu region. The proposed system also aims at classification of soil into low, medium and high categories by adopting machine learning techniques. This research can help the soil analysts and farmers to decide sowing in which land may result in better crop production. A graphical user interface has been developed in order to help farmers and soil analysts to predict crop yield and recommend fertilizers.

III. PROPOSED SYSTEM

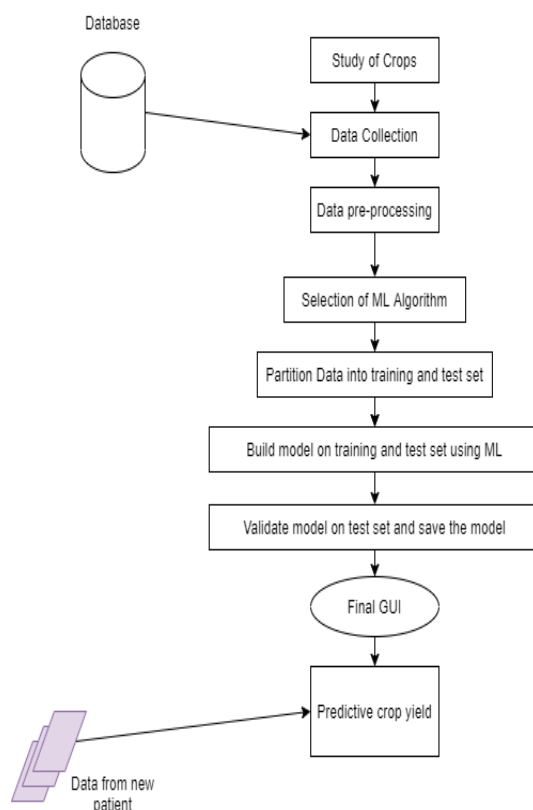
Early Crop Yield Prediction from soil nutrient analysis is one of the major parameter affecting crop yield. Agriculture Output prediction and comparison with other methods has been commonly done on already available secondary and small datasets but in this study a recent and expanded dataset will be used for this purpose. There is a need of selection of best suitable machine learning algorithms for crop yield prediction on huge dataset on different applicable and integrated tools. Fertilizer recommendation and knowledge based agriculture can be used to mitigate agricultural risk. Analyzing the Soil and Predicting the Category of Yield well in advance is one of the important prospects in the field of agriculture. Most Farmers in India lack knowledge about the soil type, its nutrient content, atmospheric conditions prevailing in area, best type of crop suited for soil, seed type and most importantly fertilizers content. Machine Learning in Agriculture can be efficiently used to predict crop yield well in advance and also help farmers to get higher yield. The solution for the current system is to develop a system which will sort out all the above problems. The proposed system will work upon the current issues and will overcome the problem that it has right now. Working of proposed system is explained & presented in figure 1.

Collection of Soil Sampling Data:

The experimentation was performed on the real time dataset collected from the Agriculture Department, Jammu under model village program 2019-20. The different samplings were conducted to understand the behavior of selected variables throughout the cropping season. Soil samples were collected using a soil auger from each field during growing season 2019-2020. The samples were analyzed by the agriculture department, Jammu using standard methods. The collection of data was a challenging task as the model was to be developed on real data. So, the collection of real data was a difficult task. After collecting the data when analysis of the data was done then at time it was noticed that some of the fields were blank, so we need to fill those columns by preprocessing the data.

Prediction of crop yield and fertilizers Recommendation:

We need to select the best technique from the available techniques to preprocess the data. The proposed model will predict the yield of crops like wheat, rice, mustard etc. It will take the soil parameters as input and based on those parameters it will predict the crop yield for the future production. The proposed model will also recommend the suitable fertilizer & Pesticides if required.



IV. RESULTS & ANALYSIS

The experiment results for mustard dataset can conclude that machine learning techniques KNN and Random Forest are found to be the most accurate algorithm for yield prediction. KNN, Naïve Bayes and Random Forest are found to be the most accurate algorithm for wheat yield prediction. Machine learning algorithms based system results for all datasets are presented in Figure 2, 3 &4. Random forest produces the best result with accuracy of 99.3% and for rice yield prediction, it can be noticed that all the algorithms can be used for accurate yield prediction of rice crop; but Naïve Bayes and Random Forest predicted better performance. These machines learning techniques based system will help the farmer in predicting yield in advance based on soil parameters. These models also incorporate Fertilizer recommendations which will help the farmers to make decisions accordingly. Performance Metrics for undertaken machine learning algorithm for all datasets are presented in Figure 5.

Figure 2: Proposed Implemented System

Figure 3: Crop Yield Prediction usin ML Techniques

Figure 4: Fertilizer Recommendation

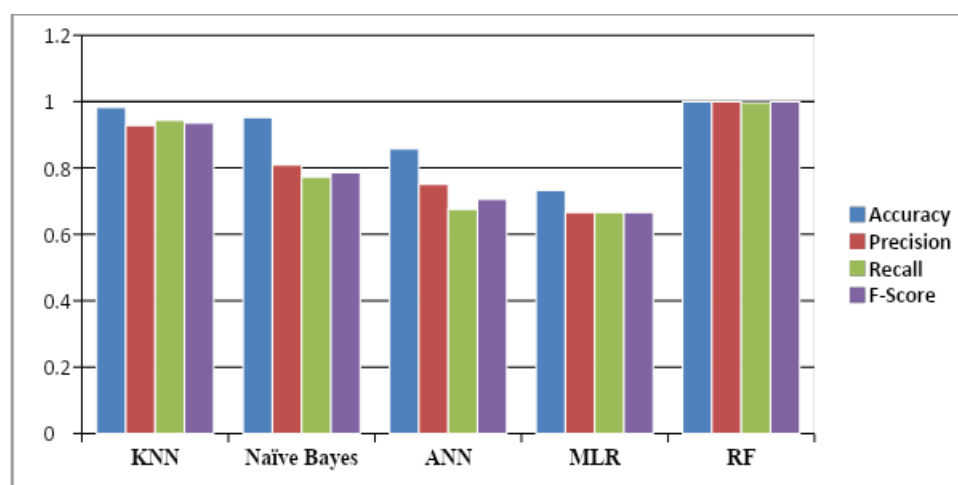


Figure 5: Comparison of the performance of machine learning algorithms for all the datasets

V. CONCLUSION

Yield Prediction using machine learning algorithms is one of its kind system which has been developed no such system is available in market with such functionalities in a single system with such a versatility and user friendly system with a high level of smart crop yield prediction. The proposed system has proved to be significantly useful for the end-user as it meets all the requirement of the user as well as fulfils the entire criterion that judge the quality and success of the system. The success for the proposed system is achieved as it has met the success criteria and established what it was supposed to. The proposed system after evaluation has been found to successfully guarantee a good prediction of crop yield before cultivation of crops and hence has solved the problem what was expected. System experimental results aid farmers and analysts in analyzing and providing qualitative judgment to the problem of crop prediction

VI. REFERENCES

- [1] Agriculture in India - statistics & facts. Retrieved November 30, 2020, from <https://www.statista.com/topics/4868/agricultural-sector-in-india/>.
- [2] Bhuyar, V. (2014), "Comparative Analysis of Classification Techniques on Soil Data to Predict Fertility Rate for Aurangabad District", International Journal of Emerging Trends and Technology in Computer Science, 3(2), 200-203.
- [3] Bondre, D. A. & Mahagaonkar, S. (2019), "Prediction of Crop Yield and Fertilizer Recommendation using Machine Learning Algorithms", International Journal of Engineering Applied Sciences and Technology, 4(5), 371-376.
- [4] Champaneri, M., Chandvidkar, C., Chachpara, D. & Rathod, M. (2020), "Crop Yield Prediction using Machine Learning", International Journal of Science and Research.
- [5] Klompenburg, T. V., Kassahun, A. & Catal, C. (2020), "Crop Yield Prediction using Machine Learning: A Systematic Literature Review", Computers and Electronics in Agriculture, 177.
- [6] Lata, K. & Chaudhari, B. (2019), "Crop Yield Prediction using Data Mining Techniques and Machine Learning Models for Decision Support System", Journal of Emerging Technologies and Innovative Research, 6(4).
- [7] Manimekalai, S. & Nandhini, K. (2018). Crop Yield Prediction from Soil Parameters through Neupper Rule Established Algorithm. International Journal of Engineering and Technology, 7, 908-912.
- [8] Manjula, E. & Djodiltachoumy (2017),"A Model for prediction of Crop Yield", International Journal of Computational Intelligence and Informatics, 6(4).
- [9] Priya, P., Muthaiah, U. & Balamurugan, M. (2018), "Predicting Yield of the Crop using Machine Learning Algorithm", International Journal of Engineering Sciences and Research Technology, 7(4).
- [10] Shastri, A., Sanjay, H. A. & Bhanusree, E. (2017), "Prediction of Crop Yield using Regression Techniques", International Journal of Soft Computing, 12(2), 96-102.
- [11] Suganya, M., Dayana, R. & Revathi, R. (2020)' "Crop Yield Prediction using Supervised Learning Techniques", International Journal of Computer Engineering & Technology, 11(2), 9-20.
- [12] Supriya, D. M. (2017), "Analysis of Soil Behavior and Prediction of Crop Yield using Data Mining approach", International Journal of Innovative Research in Computer and Communication Engineering, 5, 9648-9652.
- [13] www.bhu.ac.in
- [14] www.ijettcs.org

Clinical and Epidemiological Factors Analysis for COVID-19 by Mamdani-Fuzzy Inference System

Poonam¹, Monika Mangla², Nonita Sharma³, Reena³, Vaishali Mehta Wadhwa⁴

¹JC Bose University of Science and Technology, YMCA, Faridabad, Haryana, India

²Lok Manya Tilak College of Engineering & Technology, Navi Mumbai, India

³Dr. B.R.Ambedkar National Institute of Technology, Jalandhar, Punjab, India

⁴Panipat Institute of Engineering and Technology, Panipat, India

¹poonamgarg1984@gmail.com

Abstract: During this pandemic outbreak of COVID-19, the whole world is getting severely affected in respect of population health and economy. This novel virus has brought the whole world including the most developed countries to a standstill in a very short span like never before. The prime reason for this unexpected outburst of COVID-19 is lack of effective medicine and lack of proper understanding of the influencing factors. Here, authors aim to find the effect of epidemiological factors that influence its spread using a fuzzy approach. For the same, a total of 9 factors have been considered which are classified into *Risk* and *Preventive* factors. This fuzzy model supports to understand and evaluate the impact of these factors on the spread of COVID-19. Also, the model establishes a basis for understanding the effect of risk factors on preventive factors and vice versa. It is worth mentioning that this is the first attempt to analyze the effect of Clinical and Epidemiological factors with respect to COVID-19 using a fuzzy approach.

Keywords: Coronavirus, COVID-19, Fuzzy approach, Clinical factors, Epidemiological Factors, Risk factors, Preventive factors, Mamdani-fuzzy inference model.

1. Introduction

The first case of pneumonia of unidentified cause was confirmed to the WHO (World Health Organization), China on December 2019. Since then, the epidemic of this virus became a matter of severe concern and within 1 month, the outbreak was declared an international public health emergency on 30 January 2020. Also, WHO coined a name for new coronavirus disease: COVID-19 and declared it a pandemic on 11th March 2020. The emergence of the novel coronavirus (SARS-CoV-2) infection (COVID-19) had raised a wave of shock and concern around the world as on 1st May 2020, there have been 3,181, 642 cases and 2,24, 301 deaths affecting 215 countries and territories[1]. These numbers are expected to rise significantly in future as there is no successful vaccine or antiviral strategy available against the disease. Preventive measures like lockdown remain the only viable strategy to stop the viral transmission. In the midst, of this global pandemic accurate and timely information about the origin and epidemiological characteristics can prove to be a probable solution to contain this virus. Due to elevated transmissibility of these emerging infectious diseases of zoonotic origin, identifying the drivers and spatial diffusion is a great challenge. Efficient integration of tools like epidemiology, fuzzy approach and bioinformatics can provide solutions to these complex situations.

COVID-19's first case was associated with direct access to Wuhan's Huanan Seafood Wholesale Market, and from there the animal-to-human transmission was believed to be the main source of disease spread. Subsequent cases, however, were not found to be linked to the same exposure, and thus symptomatic human-to-human transmission was identified as the key cause of COVID-19 spread. At the same time, pre-symptomatic and asymptomatic transmission for the spread of the disease is equally likely [2].

Several demographic factors, including gender, age, and blood type, have previously been classified as risk factors for infection but the biostatistics studies have not been adequately advanced to examine complex associations between several variables until recently. With the emergence of fuzzy logic and genetic algorithms applications, computers can now find correlations between variables and find out the most promising ones.

The manuscript works on the broad classification of factors involved in the spread of COVID'19 as risk and preventive factors. A fuzzy model based on Mamdani approach is presented to model the inter-relationship and dependency of the factors. Three fuzzy models are presented; First based on the mapping of various factors on the risk evaluation; second based on the preventive factors mapping. The final model is the mapping of the risk and preventive factors to evaluate the susceptibility index.

The manuscript is organized as follows: After Introduction in section 1, section 2 covers the related work to COVID-19 domain. Section 3 describes the factors taken for the fuzzy modelling approach. Section 4 presents the proposal model. Section 5 presents the results and discussion of various factors on the susceptibility index, with conclusion given at the end.

2. Related Work

Researchers and health professionals have been fascinated by the spread of this epidemic virus. The pattern of unprecedented and unmatched spread of COVID-19 needs to be understood in order to devise efficient preventive mechanisms to curb its spread. The prime challenge for completely understanding the pattern of its spread is its short span. Hence, intensive study is being carried out by researchers to assimilate its epidemic nature which aids in accurately forecasting the increase in infected people. This section briefly discusses the recent studies related to the forecast of its outbreak.

Authors in [3] aim to elucidate coronavirus transmission at different stages by using Gaussian distribution. For the same, authors simulate the model considering the data related to the Hubei epidemic situation. Authors claim that the gap between model's predicted values and actual value is quite small. Thus, the proposed model can lay the basis for epidemic prevention and control in the affected countries. Another prediction model is proposed by Zifeng Yang et.al [4]. In this study, authors use the Susceptible-Exposed-Infectious-Removed (SEIR) model on migration data (before and after January 23). Additionally, authors used the most updated epidemiological data related to COVID-19 to understand epidemic curve. Further, artificial intelligence (AI) is also integrated with SEIR and this modified SEIR model is trained using 2003 SARS data. This modified SEIR epidemiological model also included some additional parameters apart from those considered for Ebola virus analysis in 2018. Authors guarantee that this dynamic SEIR model efficiently predicts the epidemic peaks and sizes.

Authors in [5] proposed a classifier model that uses a neural network to understand how a patient responds to the treatment. For the same, this model uses related data from February 20, 2020 to March 9, 2020 related to recovered patients and dead patients. The proposed model uses seven different variables namely country, region, infection reason, confirmation date, birth year, sex and group. The proposed classifier model aims to find the most effective variables to determine fatality or recovery. This model establishes three most influencing variables viz. infection reason, confirmation date and region that strongly predicts the future state of patient. Thus, the proposed NN based model helps to envisage the status of COVID-19 patients. The transmission trend of epidemic is also studied by Yichi Li et.al [6]. This understanding enables to devise short term prevention programs that helps in devising various control measures to limit further propagation of this virus. For the same, authors present a dynamic model for epidemic and a time series model for short-term forecasting of transmission of COVID-19. Authors claim to achieve accurate prediction of the forecasting.

Work by Joel Hellewell et.al [7] further assisted the belief that contact tracing and isolation of infected cases is effective to control the spread of this contagious disease. Authors concluded this by using a stochastic transmission mode that ascertains the effectiveness of isolation and contact tracing. This model works by finding the number of primary infected people generated and secondary infected people that helps in finding the reproduction number (R_0). R_0 is a crucial factor as it indicates the number of people who get infected by an infected person. If R_0 goes above 1 then the rate of infection grows exponentially. However, if R_0 remains below 1, then the rate of infection may decay exponentially [8]. Vitaly Volpert et al. [8] used the data analytics to R_0 can be controlled by restricting the contact between individuals through a quarantine model. However, it also claims that the mere quarantine method is not sufficient to curb its spread thus new strict measures need to be considered. Authors in [9] claim that R_0 is in the expected range of 1.4-3.9 for India. This claim is made by comparing data of India with several other countries in the US. Exponential and classic susceptible-infected-recovered (SIR) models are used to perform predictions (both long-term and short-term). This model predicts that equilibrium is achieved in India by May 2020. Authors also mention that this estimation stands invalid if community transmission takes place in India as this prediction is made considering sophisticated levels of social distancing.

Melanie Bannister et al. [10] considered the influence of environmental factors on the spread of COVID-19 and presented an interesting fact that mentions the impact of temperature on its spread in Europe. The study states that higher temperature diminishes the proliferation of this virus. Similarly authors in Jianyun Lu et.al

[11] studied the relationship of outbreak with ventilation. Here, authors concluded that droplet transmission is enhanced in an airconditioned environment as it persists longer in the air. Moreover, the direction of air flow is also an important factor in this regard. Apart from environmental factors, research also attempts to understand the relation of physical health. A such attempt is made to understand the impact of existing health conditions (including hypertension, diabetes and other respiratory disease) of the infected person by [12]. Authors in [12] present an analysis based on review of different databases i.e. PubMed, EMBASE and Web of Science to assess the prevalence of comorbidities in the COVID-19 patients. This study claims that the existing conditions add more risk to severe patients over non-severe patients.

3. Related Epidemiological and Clinical Factors

This section discusses the various epidemiological and clinical factors that have been considered to understand the spread of pandemic coronavirus. The contributing factors have been classified into risk and protective factors. Risk factors represent environmental and some other factors which are beyond human control. Examples of risk factors are age, humidity, and temperature etc. However, protective factors are under human control and thus can be adjusted in response to requirement. Few examples of protective factors are personal hygiene, social distancing etc.

I. VIRULENCY

Coronavirus (CoVs) is an unparalleled damaging virus for human, livestock, gastrointestinal, hepatic, infecting respiratory, and central nervous system [13]. During the 2002 SARS outbreak and 2012 MERS outbreak, the likelihood of transmission of CoVs from animals to humans was already confirmed [14]. Prior to 2019, there were only few CoVs that could infect respiratory system in human viz. HCoV-229E, HCoV-OC43, HCoV-NL63 and HKU1. These viruses could trigger only mild upper respiratory disease in normal conditions although it could also lead to serious infection under extreme circumstances. SARS-CoV and MERS-CoV are other similar viruses which can infect lower respiratory tract and cause severe respiratory syndrome [13][15].

This recent virus COVID-19 has been unanimously accepted to be the most severe virus till date as its spread is quite fast and uncertain. According to the recent guidelines by Chinese health authorities' [16] three are major transmission modes for COVID-19: 1) transmission through droplets, 2) transmission of aerosols and finally 3) transmission by contacts. Apart from these, digestive system is also considered to be a possible transmission mode for COVID-19. The distance covered by these droplets is normally limited to 2 meters. Nevertheless, this airborne transmission also takes place with much smaller droplets that can float and travel longer distance [17]. Moreover, these airborne droplets remain in environment for hours and even days under particular environmental conditions [18].

II. Host Defense Potential

Clinically, SARS-CoV-2 or any similar virus triggers the immunity system to work in two stages. Initially, it requires an immunological response to eradicate the virus during the incubation stage and thus prevent its progression to severe stages. These two stages are clinically known as anti-sera or pegylated IFN α . For the same, health of the host should be in good general health. Additionally, it also necessitates a favorable genetic history as it elicits adaptive antiviral immunity for the growth of an endogenous defensive immune response at the incubation and non-severe stages.

Genetic differences in the immune response to pathogens are believed to respond to human variations. However, when a defensive immune response is compromised, virus spreads and massive tissue destruction starts particularly in intestines and kidneys leading to severe viral infections [19] [20]. Thus, immune insufficiency or misdirection leads to viral replication and hence damages tissue. By study, it is evident that Immune insufficiency or misdirection can increase viral replication and cause damage to tissues [21].

III. Underlying Health Conditions

Health of an individual plays an important factor considering in susceptibility of COVID-19. It appeared prematurely in the COVID-19 epidemic that infants or healthy people are safe from this virus, but later it was

understood that it could be just because they were unlikely to have visited the Wuhan wet market. Another reason for the same could also be that they were more probable to have asymptomatic or benign infection and therefore implausible to have been screened for it. Nevertheless, COVID-19 affected infants (1 month old). Also no cases of negative child outcomes have been reported for COVID-19 positive moms pregnancy [15].

IV. Atmospheric Temperature and Humidity

According to a research, it is claimed that high temperature and humidity can reduce the spread of coronavirus. This claim is in agreement the similar claim for spread of Influenza [22]. At the same time, it is also too early to make any strong claim based on the small historical data of not more than 4 months. Primarily, this claim is based on the strong belief that the coronavirus primarily transmits through respiratory droplets [23]. For the same, the behavior of coronavirus at different temperatures and humidity is being studied intensively. According to the study in [24], authors claim using linear regression framework that high humidity and high temperature lead to reduction in transmission of COVID-19. Authors also claim that each 1^o Celsius increase in temperature and 1% increase in humidity results in reduction of daily reproductive value by 0.0225 and 0.0158, respectively.

Similar study for SARS coronavirus in [23] revealed that the SARS coronavirus can retain for 5 days at temperatures of 22–25°C and relative humidity of 40–50% . However, the virus is lost at higher temperature (>38°C) and relatively higher humidity (>95%). Hence, it is evident that low temperature and low humidity environmental conditions favors expediting the virus spread.

V. Airflow and Ventilation

During the intensive study of contributing factors responsible for the spread of coronavirus, a shocking conclusion was drawn while understanding the outbreak of the virus in a southern chinese city of Guangzhou [11]. In this incident, it was noticed that the direction of droplet transmission was consistent with airflow direction. This incidence sidelined the prevalent opinion about its spread by giving an eye opening result that coronavirus can also ride on currents created by air-conditioning. Resultantly, HVAC (Heating, Ventilation and Air Conditioning Systems) engineer's society itself made an official statement that airborne transmission of COVID-19 is most likely that leads to implementation of necessary changes in the building ventilation. Authors in [25] also attempted to establish the transmission mechanism of this virus. In the study, authors in [25] conclude that HVAC, if not implemented correctly, may enhance and expedite the spread of the virus. Hence, a major shift is mandated to prevent virus spread by implementing structural changes to buildings and HVAC systems.

VI. Personal Hygiene factors

It is quite evident since the explosion of COVID-19 that this virus transmits through physical contact. Hence, the spread of this virus may be controlled by maintaining sophisticated hygiene standards. Relation between personal hygiene and virus spread has already been established by authors in [26]. Moreover, the evidence of hand hygiene in preventing nosocomial infections was demonstrated as early as 1847. However, standard personal hygiene is not limited to hand hygiene only. It also considers safe waste management, respiratory hygiene, environmental cleaning and sterilization of patient-care equipment etc. [27]. For COVID-19, maintaining personal hygiene is most crucial as it spreads through air droplets of infected people. Hence, it is of paramount importance that the patient must cover their nose and mouth while coughing or sneezing [28].

VII. Social Distance

As the novel COVID-19 outbreak continues to proliferate around the globe, healthcare experts and government officials strongly advise individuals and families to position space between themselves and others. The idea of social distancing seems to have gone from emergency healthcare preparedness and mitigation jargon to a household expression till a proven vaccine is available. As COVID-19 is known to spread through human-to-human transmission, the transmission is known to affect contacts within 6 feet radius. The virus can be transmitted airborne and stay on objects for as many as three hours; hence social distancing is considered to be the viable and easiest solution to contain this virus.

VIII. Population Density

Another most important factor in analysing the susceptibility of the virus is population density. COVID-19 has taken hold and struck hard in different places of the world. For instance, largely dense cities, like New York, London, with the huge inflows of tourists, and hugely dense residential areas, industrial centers like Wuhan, Italy, Detroit have shown the spread of the epidemic exponentially. At the same time, its many other factors along with population density such as population size, age, education, level, work culture, also play prominent roles in the spread of epidemics.

4. Proposed Model

The literature survey shows that to ensure the susceptibility of a subject, a number of epidemiological and clinical factors need be considered and their collaborative effect has to be ensured. For the same, authors propose a fuzzy inference model in this section.

4.1 Objective

An integrated analytical approach is essentially required to model the inter-relationship and dependency of the factors. Hence, a model needs to be designed to analyse the impact of various influencing factors.

4.2 Proposal

The proposed model employs a fuzzy expert system based on Mamdani-fuzzy inference [29] system for predicting the susceptibility of the subject based on various epidemiological and clinical factors as discussed in the previous section. Here, authors attempt to give a brief introduction to the employed fuzzy expert system.

4.3 Fuzzy Expert System

Fuzzy logic is a mathematical tool to predict unknown and multi-factorial issues. Fuzzy expert system is illustrated through a block diagram as shown in following Fig. 42 containing various blocks.

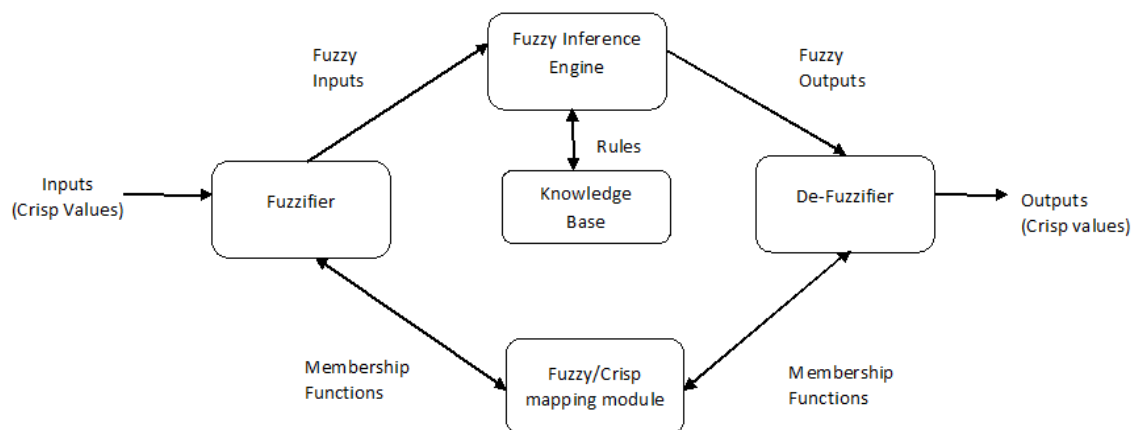


Fig. 42: Block diagram of Fuzzy Expert system

The description of basic blocks is as follows: The process begins by feeding crisp values of input parameters to the fuzzifier module for fuzzification.

Fuzzifier Module: This module transforms the crisp inputs into linguistic variables and their corresponding memberships.

Fuzzy/Crisp mapping module: This module contains the mathematical details for converting a crisp value to its corresponding fuzzy set and associated membership value and vice versa.

Knowledge Base: This module contains a library of conditional rules in 'IF-THEN' form.

Inference Engine: - It refers to the knowledge base and selects the relevant rules to be fired. The selected rules are placed in the conflict set and fired one by one. The inference engine used in our case is that of Mamdani-type.

Defuzzification Module: - This module converts the obtained crisp value using any of the chosen defuzzification methods.

Characteristics of Mamdani-based fuzzy inference system are presented in Table 1.

Table 22: Characteristics of fuzzy inference system Mamdani Model

Operation	Operator	Role	Formula
Sum (OR)	MAX	T-conorm	$\mu C(x) = \max(\mu A(x), \mu B(x)) = \mu A(x) \vee \mu B(x)$
Subscription (AND)	MIN	T-norm	$\mu C(x) = \min(\mu A(x), \mu B(x)) = \mu A(x) \wedge \mu B(x)$
Implication	MIN	T-norm	$\max(\min(\mu A(x), \mu B(x)))$
Aggregation	MAX	T-conorm	
Defuzzification	Centroid		$COA = Z = \frac{\int z \mu(z) dz}{\int \mu(z) dz}$

4.4 Proposed Model based on Mamdani based FES

The proposed model is implemented in MATLAB-R2013b. This model is based on a hierarchy of 3 FES where risk value and prevention value are calculated through the base FES and the results are fed into the inherited FES, which results into the susceptibility level determination as shown in Fig 43.

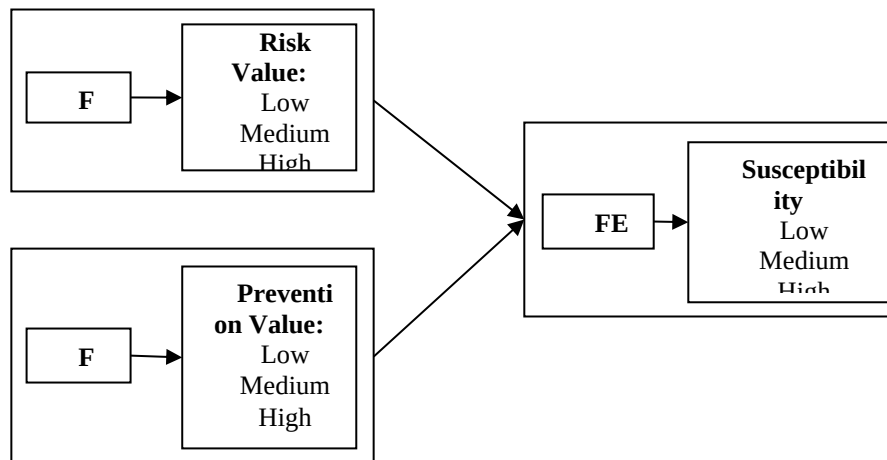


Fig 43: Hierarchy of FES in the Proposed Model

4.4.1 FES Implementation for Risk Value

Fuzzy pattern to calculate risk value is shown in Fig 44 where 5 major risk factors are considered as input to the FES. Each factor is an aggregate of multiple related risk factors like Immunity covers age, Current health position and other related health factors; similarly, atmosphere temperature also includes humidity and related factors.

These factors have their different range of variations like Virulency is scaled on a Likert Scale (0-10), Immunity is also scaled on Likert Scale (0-10), Atmospheric Temperature variate from 0-45 Degree Celsius, Population density variate from 0-1000 people/ KM² (2 or 3) and Ventilation variate from 0-50 CFM as shown in table 2 and crisp value of these variations are fuzzified into 2 or 3 linguistic variables low, medium and

high etc as per the context. As per the variation of each input factors into different membership functions, total rules formed are 162. On these rules, input value of these 5 parameters are mapped and risk value is calculated by Centre of Area (Centroid).

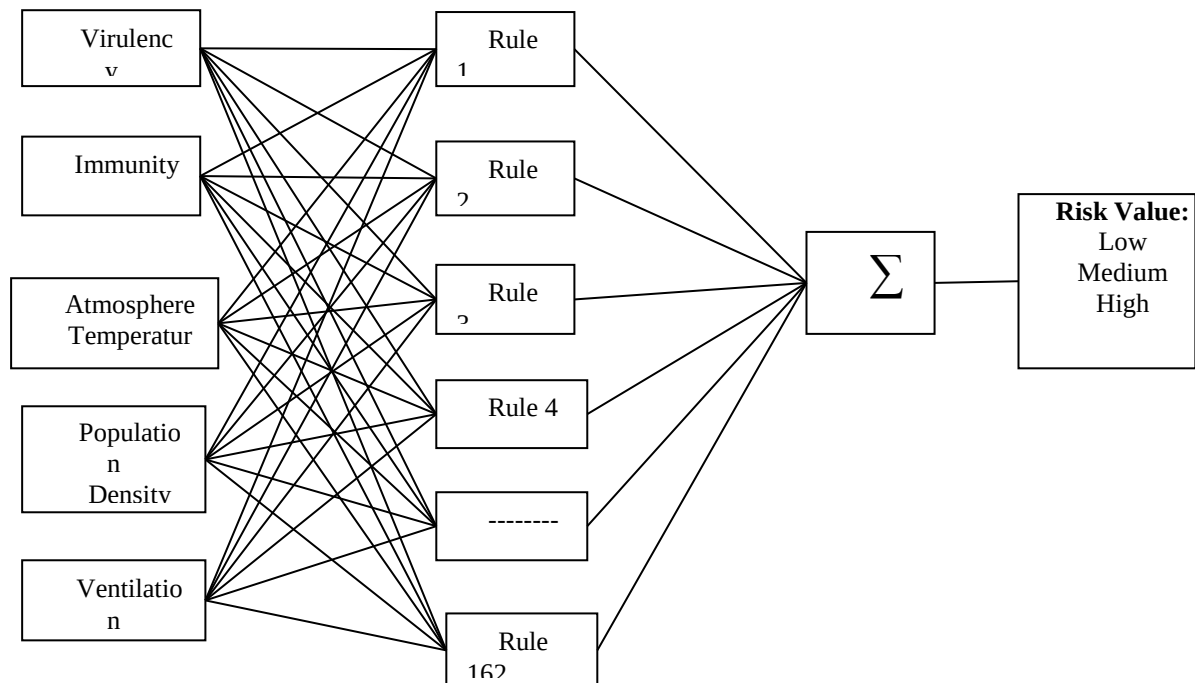


Fig 44: Fuzzy pattern to calculate Risk Value

The membership functions and the input range for each input parameter is represented in following table 2.

Table 23: Fuzzy classification of input and output for Risk variables

Parameters		Membership functions		
		L	M	H
Input	Virulency	Contagious [0 0 4]	More Contagious [3 5 7]	Most Contagious [6 10 10]
	Immunity	Adaptive [0 0 6]	Innate [5 10 10]	
	Temperature	Cold [0 0 20]	Mild [15 22 30]	Hot [22 45 45]
	Population Density	Sparse [0 0 300]	Moderate [200 350 600]	Dense [450 1000 1000]
	Ventilation	Suffocated [0 0 25]	Moderate [18 28 40]	Airy [30 50 50]
Output	Risk Value	Low [0 0 4]	Medium [3 5 7]	High [6 10 10]

Design of FES1 is done using fuzzy tool of MATLAB as shown in Fig 45 where a fis file of the type Mamdani is created and all the 5 factors are created as input variable and output is a single variable named, Risk Value. All 162 rules are designed in the rule base as shown in Table 24. These rules are If- Then conditions with logical operators as AND, OR and NOT.

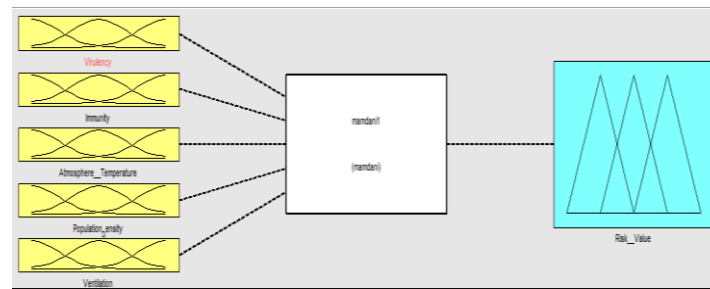


Fig 45: Snapshot of FES1 to Calculate Risk Value

Table 24: Rule Base for FES1

Inputs					Output
Virulency	Immunity	Temperature	Population Density	Ventilation	Risk Value
Contagious	Adaptive	Cold	Sparse	Suffocated	Medium
Contagious	Adaptive	Cold	Sparse	Moderate	Low
Contagious	Adaptive	Cold	Sparse	Airy	Low
Contagious	Adaptive	Cold	Moderate	Suffocated	Medium
Contagious	Adaptive	Cold	Moderate	Moderate	Medium
Contagious	Adaptive	Cold	Moderate	Airy	Low
Contagious	Adaptive	Cold	Dense	Suffocated	Medium
Contagious	Adaptive	Cold	Dense	Moderate	Medium
Contagious	Adaptive	Cold	Dense	Airy	Low
:	:	:	:	:	:
:	:	:	:	:	:
Most Contagious	Innate	Hot	Dense	Airy	Medium

4.4.2 FES Implementation for Prevention Value

Following Fig. 5 illustrates the Fuzzy pattern to calculate prevention value which takes 2 input parameters (Social distancing and personal hygiene) and gives 1 output parameter (prevention value). Here, the input parameter, Social distancing ranges from 0-12 feet and personal hygiene has its value on a Likert scale (0-10) and output variable Prevention Value has its value on a Likert scale (0-10) which has been classified into 3 different ranges thus giving 9 rules altogether. All these rules aggregate the values of two input parameters to evaluate the value of prevention value.

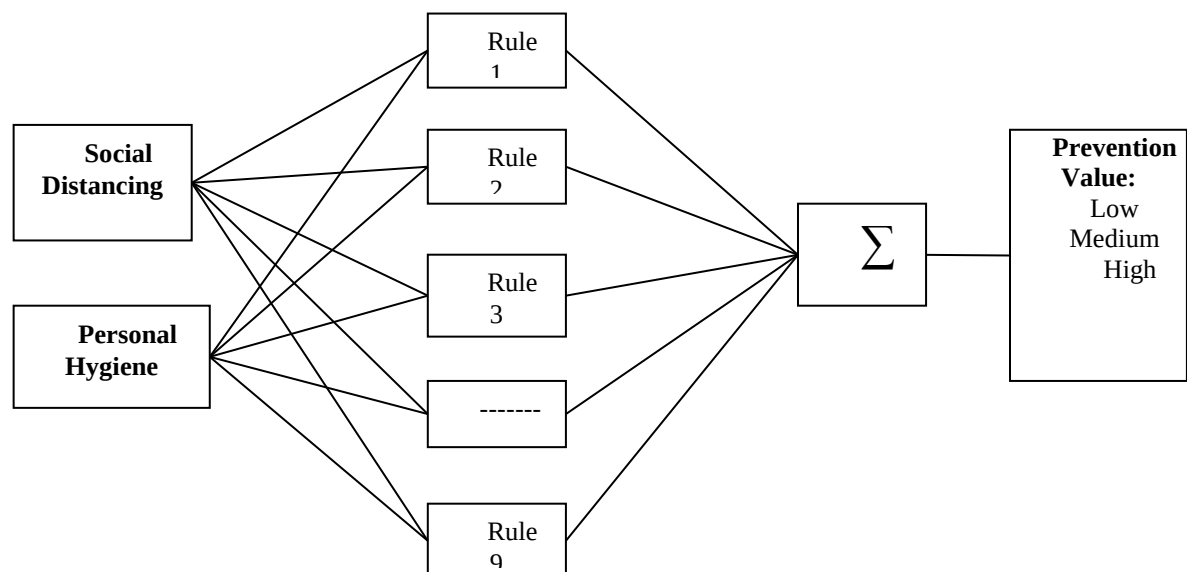


Fig 5: Fuzzy pattern to calculate Prevention value

The ranges of input and output variables for the above fuzzy pattern is given in following Table 4.

Table 25: Fuzzy classification of input and output variables for prevention value:

Parameters		Membership functions		
		L	M	H
Input	Social distancing	Inadequate [0 0 5]	Adequate [4 6 8]	Dispersed [79 12]
	Personal Hygiene	Unhygienic [0 0 4]	Requisite [3 5 7]	Sophisticated [6 10 10]
Output	Prevention Value	Low [0 0 4]	Medium [3 5 7]	High [6 10 10]

Similar to FES1, FES2 is also designed using fuzzy tool of MATLAB that uses social distancing and personal hygiene as input variable and give prevention value as output. The rule base in FES2 uses 9 rules comprising of If- Then conditions with logical operators as shown in Table 25 above. The snapshot of FES2 to calculate prevention value is illustrated in following Fig 6. The rule base is also demonstrated in Table 5.

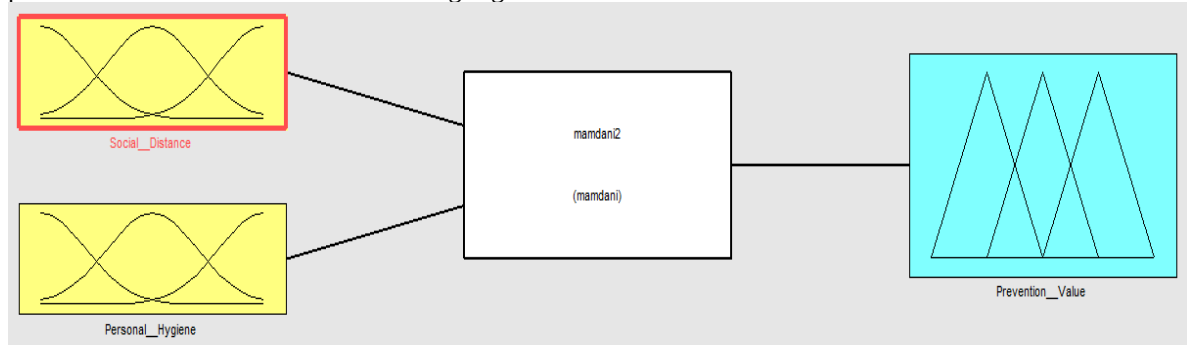


Fig 6: Snapshot of FES1 to Calculate Prevention Value

Table 26: Rule Base for FES2

Inputs		Output
Social Distancing	Personal Hygiene	Prevention Value
Inadequate	Unhygiene	Low
Inadequate	Requisite	Low
Inadequate	Sophisticated	Medium
Adequate	Unhygiene	Low
Adequate	Requisite	Medium
Adequate	Sophisticated	High
Dispersed	Unhygiene	Medium
Dispersed	Requisite	High
Dispersed	Sophisticated	High

4.4.3 Fuzzy model implementation for Susceptibility Index

Following Fig. 7 illustrates the Fuzzy pattern to calculate susceptibility index by using 2 input parameters viz. risk value and prevention value and gives 1 output parameter susceptibility index.

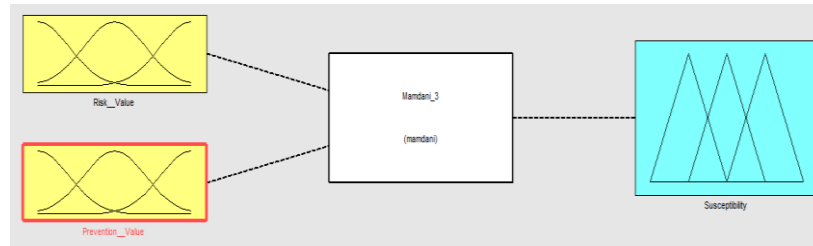


Fig.7: Snapshot of FES3 to Calculate Susceptibility Value

The corresponding Rule Base is demonstrated in following Table 27.

Table 27: Rule Base for FES3

Inputs		Output
Risk Value	Prevention Value	Susceptibility
Low	Low	Medium
Low	Medium	Low
Low	High	Low
Medium	Low	High
Medium	Medium	Medium
Medium	High	Low
High	Low	High
High	Medium	High
High	High	Medium

5. Result and Discussion

Following section demonstrates the results obtained from FES implementation of Risk factors, prevention factors, and the Susceptibility Index. A comprehensive discussion of the results section wise is presented here:

5.1 Impact of Risk Factors on Susceptibility Index

The output of FES1 i.e. Risk Value (Low, Medium and High) is converted to Crisp value by Fuzzifier and mapped on Likert scale (0-10) as in Fig. 8. Different values of Virulency, Immunity, atmospheric_temperature, population_density, and ventilation is adjusted to get different values of risk values.

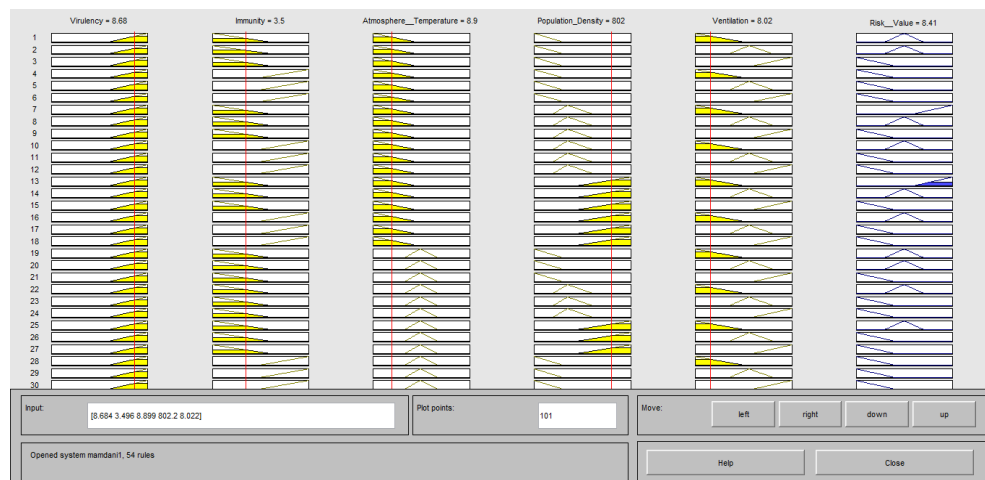


Fig. 8: Snapshot of Rule Base for FES1

As can be seen from Fig. 8, highest value of Risk_value i.e. 8.41 is obtained when virulency =8.68, immunity=3.5, Atmospheric_temperature=nearly 8.9 Degree Celsius, Population_density=802 people per KM2 and Ventilation=8.02 CFM.

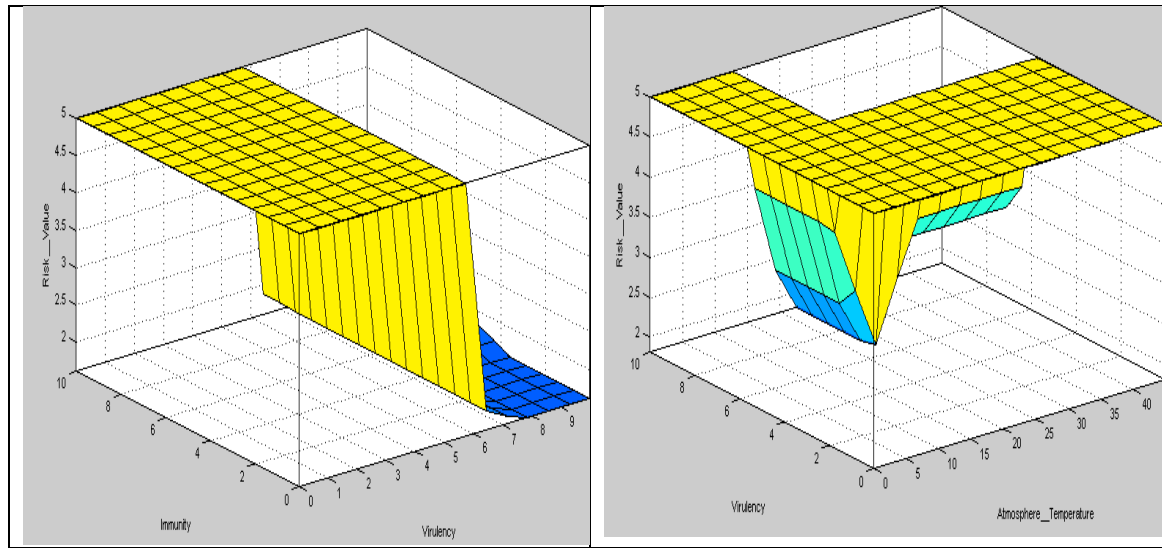


Fig. 9: 3-D Surface view of the FES1 Results

Surface view of FES1 results is shown in Fig. 9, where 2 input parameters can be taken into account at a time; similarly impact of other parameters can be seen in surface view.

5.2 Impact of Preventive Factors on Susceptibility Index

The output of FES2 i.e Low, Medium and High are the fuzzy results of prevention value which is converted to Crisp value using Fuzzifier and further mapped on Likert scale (0-10). Result of the same is illustrated in following Fig. 10 which demonstrates that when social_distancing is 8.08 and personal_hygiene is 5, then the obtained prevention value is 8.5 (High).

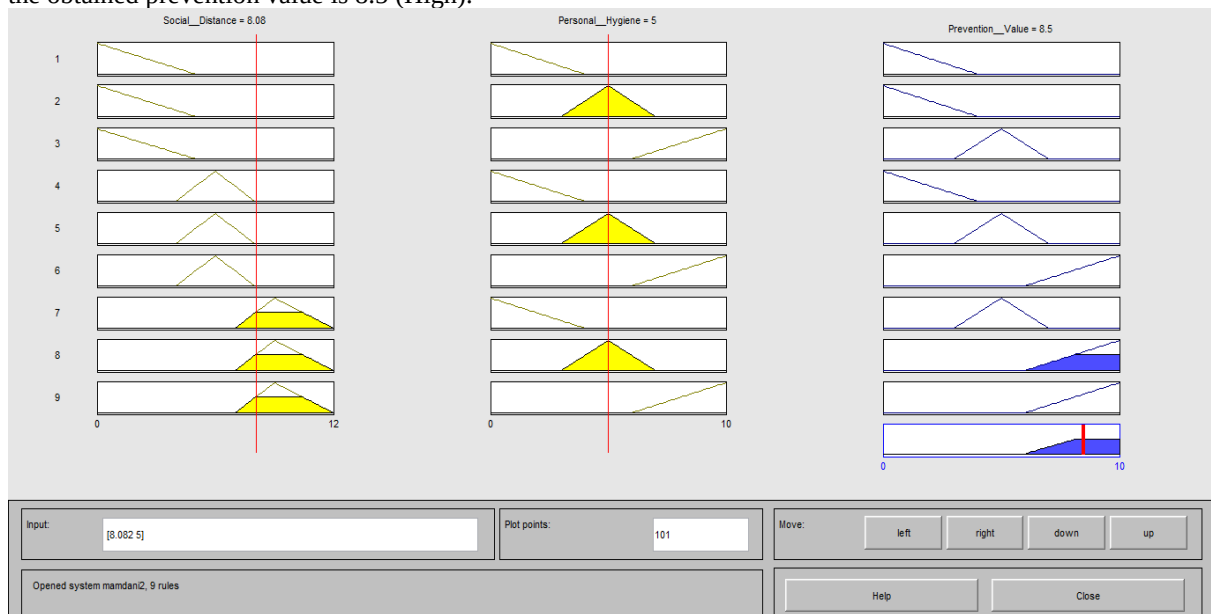


Fig. 10: Snapshot of Rule Base for FES2

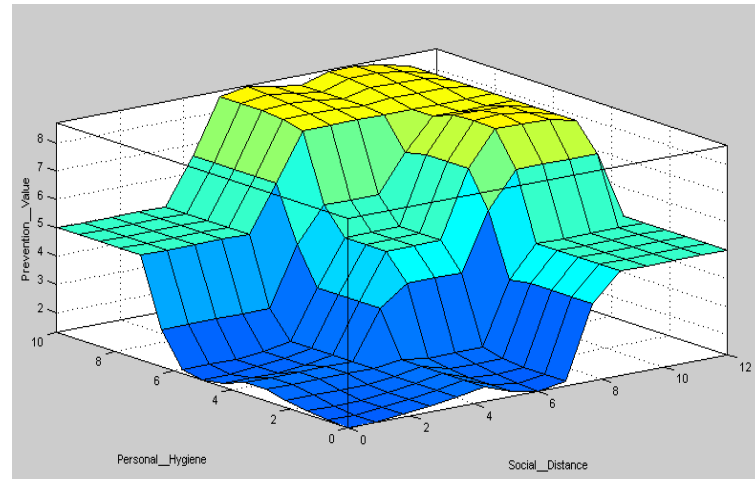


Fig. 11: 3-D Surface View of the FES2 Results

Surface view of FES2 results is represented in Fig. 11 which represents prevention value for the 2 input parameters viz. social_distancing and personal_hygiene.

5.3 Outcome of Susceptibility Index

The susceptibility index is evaluated from the optimal combination of risk factors and preventive factors. Fig. 12 demonstrates a particular case, when the risk value is 2.29 and prevention_value is 4.88, the susceptibility index becomes 1.59 (less susceptible).

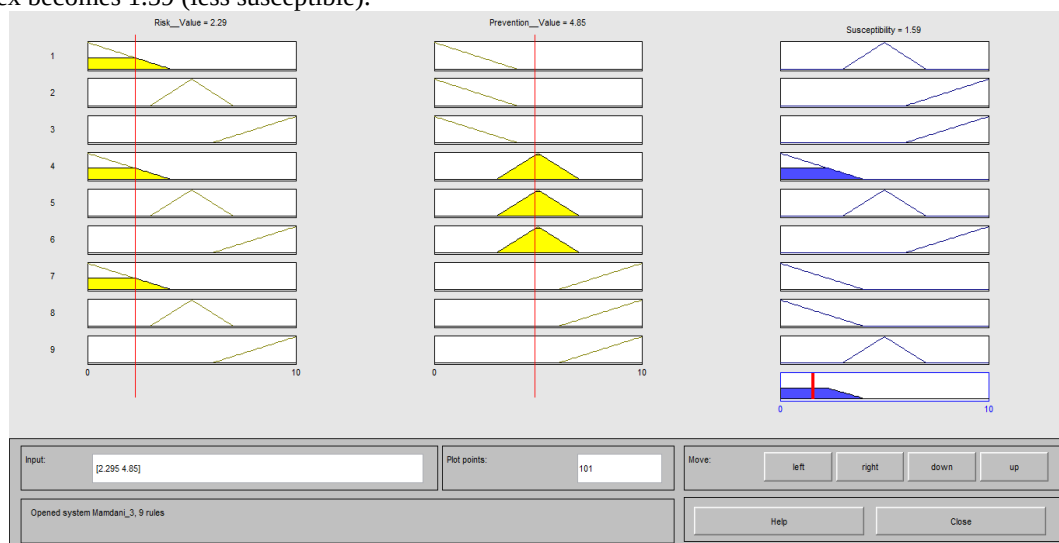


Fig 12: Snapshot of rule base for FES3

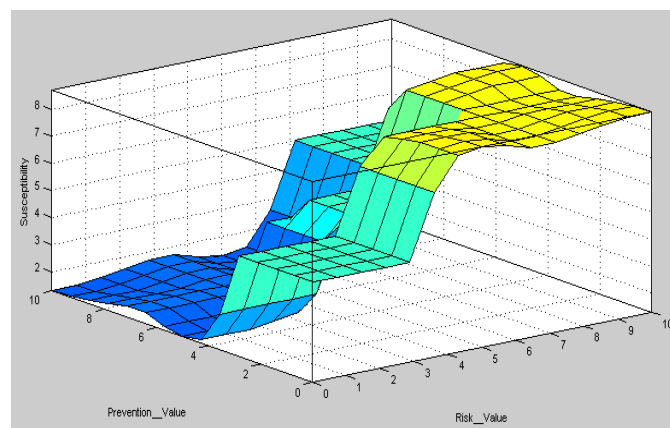


Fig.13: 3-D Surface View of the FES3 Results

Fig. 13 demonstrates the surface view of FES3 which represent susceptibility_index for 2 input parameters viz. risk_value and prevention_value. 3D demonstration of the results depicts that as the value of risk factors goes higher (i.e. above 4 units) and the preventive values, stay lower (i.e. below 4 units), susceptibility index reaches the value of >8 units. An optimal combination of risk and preventive factors will help us to keep the susceptibility index to minimum.

6. Conclusion

The manuscript proposes a Mamdani based Fuzzy inference model to determine and analyse the impact of various clinical and epidemiological factors on the susceptibility index of COVID-19. In the proposed model, the factors are broadly classified as risk and preventive factors. Risk factors evaluated are virulency, immunity, temperature, population density, and ventilation. Similarly, preventive factors like social distancing and personal hygiene are considered. Finally, susceptibility index is evaluated from the inferences of risk and preventive values. From the results, it is evident that if a host has higher risk_value, it should maintain highest prevention (social distancing and hygiene) to have lower susceptibility. The model gives promising results and thus can be successfully implemented to evaluate impact of various factors on the spread of COVID-19. The proposed model can also be extended further to incorporate additional stochastic factors.

References

- [1] "https://www.who.int/emergencies/diseases/novel-coronavirus-2019."
- [2] M. Cascella, M. Rajnik, A. Cuomo, S. C. Dulebohn, and R. Di Napoli, "Features, evaluation and treatment coronavirus (COVID-19)," in *Statpearls [internet]*, StatPearls Publishing, 2020.
- [3] L. Li *et al.*, "Propagation analysis and prediction of the COVID-19," *Infect. Dis. Model.*, vol. 5, pp. 282–292, 2020.
- [4] Z. Yang *et al.*, "Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions," *J. Thorac. Dis.*, vol. 12, no. 3, pp. 165–174, 2020.
- [5] H. Al-Najjar and N. Al-Rousan, "A classifier prediction model to predict the status of Coronavirus CoVID-19 patients in South Korea," *Eur. Rev. Med. Pharmacol. Sci.*, vol. 24, no. 6, pp. 3400–3403, 2020.
- [6] Y. Li *et al.*, "Mathematical Modeling and Epidemic Prediction of COVID-19 and Its Significance to Epidemic Prevention and," 2020.
- [7] J. Hellewell *et al.*, "Feasibility of controlling COVID-19 outbreaks by isolation of cases and contacts," *Lancet Glob. Heal.*, vol. 8, no. 4, pp. e488–e496, 2020.
- [8] V. Volpert, M. Banerjee, and S. Petrovskii, "On a quarantine model of coronavirus infection and data analysis," *Math. Model. Nat. Phenom.*, vol. 15, no. March, 2020.
- [9] E. Models, "P REDICTIONS FOR COVID-19 O UTBREAK IN I NDIA USING," 2020.
- [10] A. C. Melanie Bannister-Tyrrell, Anne Meyer, Céline Faverjon, "Preliminary evidence that higher temperatures are associated with lower incidence of COVID-19, for cases reported globally up to 29th February 2020," *medRxiv* (2020), 2020.
- [11] J. Lu *et al.*, "COVID-19 Outbreak Associated with Air Conditioning in Restaurant, Guangzhou, China, 2020," *Emerg. Infect. Dis.*, vol. 26, no. 7, pp. 3–5, 2020.
- [12] J. Yang *et al.*, "Prevalence of comorbidities and its effects in coronavirus disease 2019 patients: A systematic review and meta-analysis," *Int. J. Infect. Dis.*, vol. 94, pp. 91–95, 2020.
- [13] Y. Chen, Q. Liu, and D. Guo, "Emerging coronaviruses: Genome structure, replication, and pathogenesis," *J. Med. Virol.*, vol. 92, no. 4, pp. 418–423, 2020.
- [14] J. Cui, F. Li, and Z.-L. Shi, "Origin and evolution of pathogenic coronaviruses," *Nat. Rev. Microbiol.*, vol. 17, no. 3, pp. 181–192, 2019.
- [15] F. A. Rabi, M. S. Al Zoubi, A. D. Al-Nasser, G. A. Kasasbeh, and D. M. Salameh, "Sars-cov-2 and coronavirus disease 2019: What we know so far," *Pathogens*, vol. 9, no. 3, pp. 1–14, 2020.
- [16] "http://www.nhc.gov.cn/xcs/zhengcwj/202001/4294563ed35b43209b31739bd0785e67/files/7a9309111267475a99d4306962c8bf78.pdf." .
- [17] "http://www.nhc.gov.cn/xcs/zhengcwj/202001/4294563ed35b43209b31739bd0785e67/files/7a9309111267475a99d4306962c8bf78.pdf. Accessed 1 Feb 2020." .
- [18] T. P. Velavan and C. G. Meyer, "The COVID-19 epidemic," *Trop. Med. Int. Heal.*, vol. 25, no. 3, pp. 278–280, 2020.
- [19] Z. Xu, L. Shi, Y. Wang, and others, "Pathological findings of COVID-19 associated with acute respiratory distress syndrome [published online ahead of print February 18, 2020]," *Lancet Respir Med.*
- [20] Y. Shi *et al.*, "COVID-19 infection: the perspectives on immune responses," *Cell Death Differ.*, vol. 27, no. 5, pp. 1451–1454, 2020.
- [21] G. Li *et al.*, "Coronavirus infections and immune responses," *J. Med. Virol.*, vol. 92, no. 4, pp. 424–432, 2020.
- [22] "https://www.businessinsider.in/india/news/ a-new-research-on-coronavirus-claims-warm-and-humid-weather-can-contain-the-spread-but-thats-not-enough/articleshow/74672778.cms]." .

- [23] K. H. Chan, J. S. M. Peiris, S. Y. Lam, L. L. M. Poon, K. Y. Yuen, and W. H. Seto, "The effects of temperature and relative humidity on the viability of the SARS coronavirus," *Adv. Virol.*, vol. 2011, 2011.
- [24] J. Wang, K. Tang, K. Feng, and W. Lv, "High temperature and high humidity reduce the transmission of COVID-19," *Available SSRN 3551767*, 2020.
- [25] G. Correia, L. Rodrigues, M. Gameiro da Silva, and T. Gonçalves, "Airborne route and bad use of ventilation systems as non-negligible factors in SARS-CoV-2 transmission," *Med. Hypotheses*, vol. 141, no. April, p. 109781, 2020.
- [26] E. Mathai, B. Allegranzi, C. Kilpatrick, and D. Pittet, "Prevention and control of health care-associated infections through improved hand hygiene," *Indian J. Med. Microbiol.*, vol. 28, no. 2, pp. 100–106, 2010.
- [27] M. East, "Infection prevention and control during health care when novel coronavirus (nCoV) infection is suspected: interim guidance. (2020). Geneva: World Health Organization," *Available [https://www.who.int/publications-detail/infection-prevention-and-control-during-health-care-when-novel-coronavirus-\(ncov\)-infection-is-suspected-20200125](https://www.who.int/publications-detail/infection-prevention-and-control-during-health-care-when-novel-coronavirus-(ncov)-infection-is-suspected-20200125), accessed March 9th, 2020*, no. January, pp. 1–5, 2020.
- [28] "[WHO Guidelines on Hand Hygiene in Health Care. (World Health Organization, Geneva) 2009. Available from: http://whqlibdoc.who.int/publications/2009/9789241597906_eng.pdf." .
- [29] P. Mittal, C. K. Nagpal, S. Gupta, and K. Garg, "A Fuzzy Logic based Efficient Routing Strategy for Ad hoc Cognitive Radio Network," *Int. J. Futur. Gener. Commun. Netw.*, vol. 10, no. 10, pp. 1–22, 2017.

Parametric Optimization of Friction Stir Welding for Electrical Conductivity of Aluminium Joints Using ANN Approach

Maneesh Tewari¹, Devaki Nandan²

^{1,2}GB Pant University of Agriculture & Technology, Pant Nagar 263145, Uttarakhand, India

Abstract: Friction Stir Welding is a newer replacement of conventional fusion welding for joining low melting point materials. The solid state nature of this unique process makes it almost defect-free welding process for critical joints. The FSW set up for the experimentation was completed by developing necessary tool and fixture and joints were made on Universal Milling machine. The present research work investigates the effect of some desired input parameters for Friction Stir Welding on the electrical conductivity of AA1350 aluminium butt joints. Rotational speed of tool, Tool Traverse speed (welding speed), and tool tilt angle were opted as input parameters. After experimentation with selected parameters, investigation by ANN is done. Back propagation (BP) algorithm is utilised for developing the network. After ANN modelling, optimization of FSW parameters by full factorial method is also investigated. The ANN result outcome and the experiment data was found to be in good agreement. The optimum value of electrical conductivity of welded joints obtained was 52.7 IACS.

Keywords: Friction Stir Welding, Solid State Welding, Full factorial method, Artificial Neural Network.

1. INTRODUCTION

The Welding Institute (TWI), Cambridge (UK) was the first ever organization to introduce Friction Stir Welding as a commercial solid state joining method in 1991 [1]. Later this process became popular for its application in Aluminium alloys [2]. FSW is recognised as the key development in the defect-free metal joining of low melting point alloys. This metal joining technique is simple and green process. This process was later adopted commercially in aerospace aluminium alloys welding as well as other metallic alloys that faced problems to be welded with traditional welding [3, 5]. FSW is basically a mechanical way of joining low-melting point alloys in which a rotating tool having shoulder, mostly with threaded pin, is made to insert into the gap of joining plates. The tool is given a dwell time for a few seconds before the material softens into a paste, after which it is run at constant speed to make the FSW joint. Thus, an FSW tool heats the joining pieces as well as retains the plastic state material below its shoulder. The weld resulted by FSW is of fine grain which provides good mechanical properties [4, 5]. FSW can be understood as a mechanical way of joining using process of extrusion by a rotating tool [6].

Optimization techniques is the statistical tool to solve the engineering problems [7-9]. A broad variety of non-linear modelling implementations and optimization issues were used for ANNs (Artificial Neural Networks) in the past [10-14]. All these studies find that ANN can be utilized effectively for modelling of complex relationship among many inputs as well as outputs to predict precisely the behaviour of outputs (response variables) once the input parameter values are varying within the study range. The current research work explores the relationship between electrical conductivity and electrical conductivity as input parameter of thick AA1350 grade aluminium plates butt welded with the help of FSW.

2 Materials & Methods

Experimental set up was installed using a traditional Universal Milling machine of HMT make with a 7.5 HP motor (Fig. 1). Development of FSW tool was done by machining HSS material of shoulder diameter 25mm and straight threaded pin diameter 12mm. Strong fixtures along with backing plates were made in the shop floor for defect-free welding. The straight threaded pin was necessitated because of larger heat input for thick plate joining. AA1350 Aluminium alloy of length 100mm, breadth 100mm and thickness 12.7mm cut from rolled plate and two such samples joined together by FSW to make a welded sample size of 200x200x12.7. Samples were cut as per specification. As no previous literature data was available for FSW of thick Al plates, exhaustive and careful trial runs were required on the milling machine. First of all, the machine's rpm range for satisfactory welding were achieved through a number of trials. At the same time, a suitable travel speed for welding was also determined. Three parameters Tool Travel Speed, Tool rpm and Tool tilt angle were selected as input parameters for their effect on electrical conductivity as an output response parameter. Experimentation was done by adopting full factorial orthogonal array. An experimentation of 27 runs was conducted on

different settings as per full factorial table. Three levels of each parameter were decided (Table1). With the help of experimental results, ANN was applied to model the input and output parameters. Design of Experiments (DOE) is used to find the best input parameters setting of the process to optimize performance characteristics.

Electrical Conductivity tests (Fig. 2) were performed under laboratory conditions at BHEL- Rudrapur plant. During the test, the room temperature was maintained continuously at 20°C as per standard. The below equation was used for compensation of any temperature difference error:

$S_1 = S_2 / [1 + a \times (T_1 - T_2)]$, Where S_2 = Conductivity at observed temperature

S_1 = Conductivity at 20°C

a = Temp. Coefficient of Al alloy = 0.0043

T_1 = 20°C

T_2 = readings Temperature (taken as 23°C)



Fig. 1. Universal Milling Machine



Fig. 2. Electrical Conductivity Machine (Courtesy: BHEL-Rudrapur)

Table 1. Input Parameter with Levels

Sr No	Process Parameter	Units	Level 1	Level 2	Level 3
1.	Welding Speed	mm/min	80	100	125
2.	Tool rpm	rpm	560	710	900
3.	Tool Tilt Angle	degree	0	1	2

Table 2. Full factorial experimental data and results

Sr No	Welding (mm/min)	Speed	Rotational Speed (rpm)	Tool Tilt Angle (degree)	Electrical Conductivity
1.	80		560	0	50.7
2.	80		560	1	48.6
3.	80		560	2	46.1
4.	80		710	0	48.8
5.	80		710	1	47.5
6.	80		710	2	46.6
7.	80		900	0	47.4
8.	80		900	1	46.9
9.	80		900	2	46.7
10.	100		560	0	52.6
11.	100		560	1	49.5
12.	100		560	2	47.9
13.	100		710	0	50
14.	100		710	1	48.8
15.	100		710	2	47.9
16.	100		900	0	48.2
17.	100		900	1	48.3
18.	100		900	2	48.4
19.	125		560	0	54.1
20.	125		560	1	51.4
21.	125		560	2	50.2
22.	125		710	0	51.2
23.	125		710	1	50.9
24.	125		710	2	50.2
25.	125		900	0	49.2
26.	125		900	1	49.8
27.	125		900	2	50.4

3 Modelling Framework

This research study has aimed to establish an ANN Network to estimate the electrical conductivity (output) of thick AA1350 aluminium plates during FSW according to Kechagias et al. [15]. The three input parameters (Welding speed, Tool tilt angle, and Rotational Speed) with 3 levels each of the ANN were selected (Table 1). The experimental data of Table 2, used in modelling procedure, was linked to selected input parameters as well as the output parameter. The experimental results obtained by 27 runs were classified into 3 categories, ie, training, validation and the test samples during the frame of the ANN modelling process. These are shown to the network throughout training, and the network is modified by its errors. For testing the generalization of the network, validation samples are used till generalization stops improving. Test samples do not influence training and thus have an independent network output assessment both during and after training called confirmation runs. Seventeen (17) samples (60%) have been utilized for training, five samples (20 percent) for testing purposes, and five (5) samples (20 percent) for validation.

There are several architectural styles available for ANN. For the prediction of electrical conductivity, this work was performed using the FFBP (feed-forward with back propagation) learning. These network types have an X-input layer, hidden layers of one or more multiple neurons along with a Y-output layer. The hyperbolic tangent sigmoid hidden layer function is used in the chosen ANN whereas the function of linear transfer has been used for the output layer. Table 1 includes the three procedure parameters of the input vector. The output layer contains the efficiency measure. For the initial training process, several experiments and failures have been carried out in order to measure the best number of hidden layers and neurons. A minimum error between the ANN estimates the output with one hidden layer of 10 neurons (Fig.3) of experimental data was observed. Back propagation ANNs are prone to overtraining issues which may restrict the ANN's generalization ability. Overtraining typically takes place in ANNs with plenty of flexibility and after a number of training cycles, in that, the output of the training data set is increased whereas validation data set performance is decreased. Therefore, the network size, in the present case, is small in respect of the data set training and as shown in Fig. 4, the validation performance is sufficient if ANN training ends.

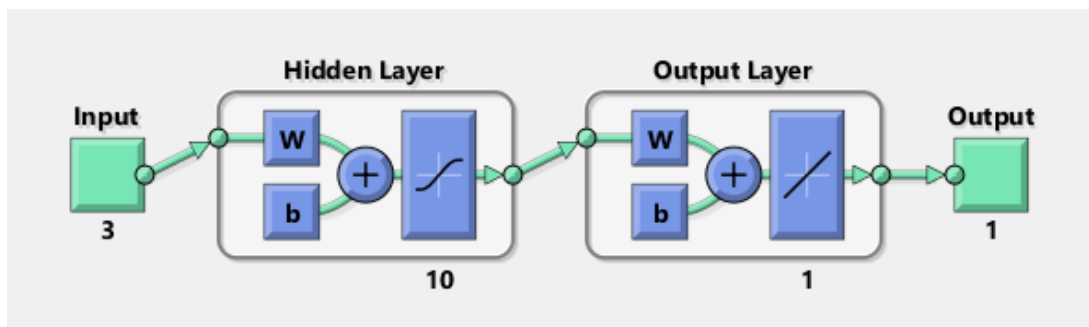


Fig.3.The ANN Architecture

The mean square error (MSE) of predicted results is used to calculate network efficiency with regard to the value of the experimental data. MSE is hereby defined as the average square difference between network outputs and experimental values. Lower the MSE values, the better it is. Null indicates no error. Regression (R) is another indicator of the performance of network efficiency. Relation between response values and the experimental values is indicated by Regression value. An R-value of one implies a close relationship (Fig. 4), 0 implies a random relationship. The following formula shows the mathematical relationship between the input and output parameters [16].

$$y = \text{purelin}(\sum_{i=1}^S w2_i * \text{tansig}(\sum_{j=1}^N (w1_{ij} * x_j + b1_i)) + b2)(1)$$

Where:

y: Output response

S: No of hidden neurons

N: No of input variables

w1: Vector of weights between the input and the hidden layer.

w2: weight vector from the hidden layer to the output.

b1: Vector of biases of the neurons in the hidden layer.

b2: Bias of the output neuron.

purelin: Linear transfer function and $\text{purelin}(z) = z$

tansig: Hyperbolic tangent sigmoid function and

$$\text{tansig}(z) = \frac{2}{1+e^{-2z}} - 1(2)$$

The experimental data is hereby used exclusively for calculation of weights and biases of Eq. 1 during the learning phase of the ANN. (Table2).

4 Results & Discussion

The electrical conductivity is predicted by using the trained ANN model within the defined range. Statistical tools and techniques are utilized for comparison of results produced by the network. The root-mean-square (RMS), mean error percentage and absolute fraction of variance (R^2) values are used to describe errors that occur during the learning and testing periods. This was carried out by checking the behaviour of the response variable with variations in the travel speed values, Rotational Speed, and Tool tilt angle.

Fig. 4 depicts the nature of regression to be linear as regards network outputs versus goals. In present case, the objective for training, testing, and validation was matched with the performance of the model, with R-values of 0.92, 0.96, and 0.88, respectively. The R-value is approximately 0.93 for the total response. The final MSE is also shown in Fig. 5. The MSE estimated prediction error is extremely low (i.e., 0.25231). The set error for both testing and validation has similar characteristics. Best validation performance was found to be at iteration 11. No significant over fitting was observed at this point. Fig. 4 explains this validation performance. Therefore, once the Welding speed, rotational speed, and tool tilt angle have been provided values within the valid range, the ANN will accurately estimate output response.

The electrical conductivity of the welding joints is analysed to study the effect of the FSW process parameters using full factorial method. The main effect plots are drawn in Fig.6. Evidently, the travel speed increase the cause of rise in electrical conductivity of weld nugget zone, and high rotational speed as well as tool tilt angle are the reasons for decrease in electrical conductivity. The best combination of input variables for maximum electrical conductivity are 125mm/min,

560 rpm, 0° and 52.7 IACS. The comparative effect of input parameters on output parameter were Tool rotation > Travel speed > Tool tilt angle.

Table 3 shows the optimum results and the confirmation test values. This proves that estimation is quite correct and prediction of percentage error is precise. The error, as a result of confirmation test, is 2.58 % by the proposed method under acceptable range.

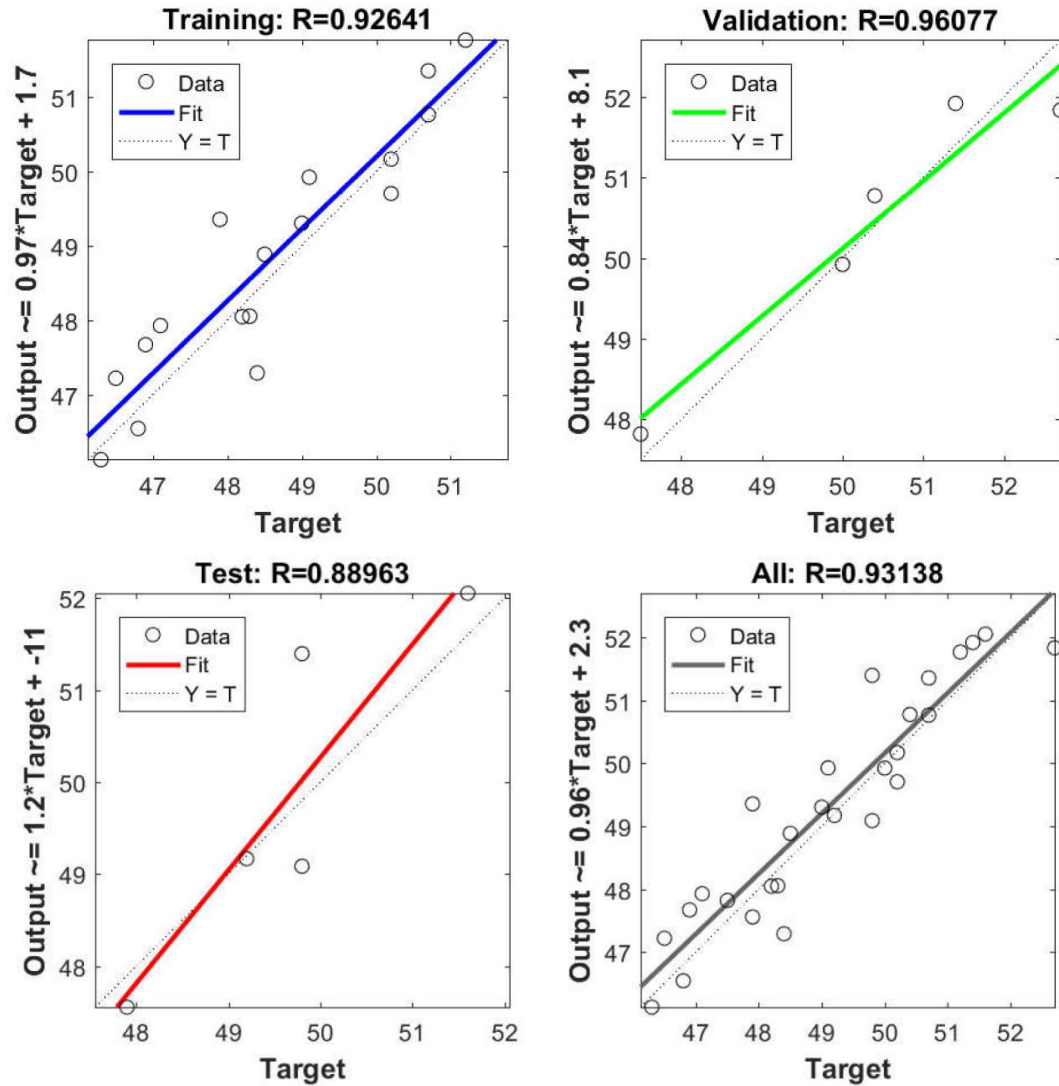


Fig.4. Correlation curves for Different ANN Stages

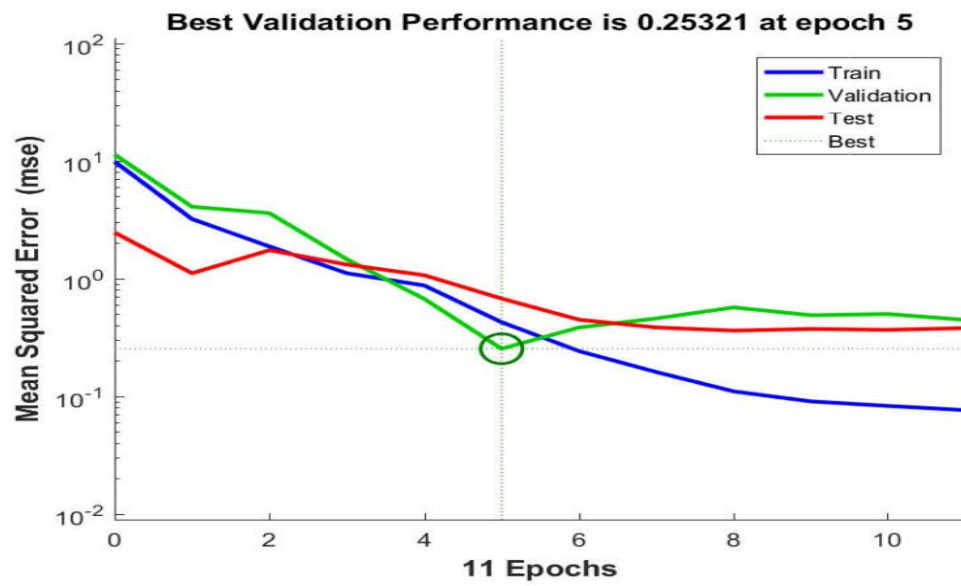
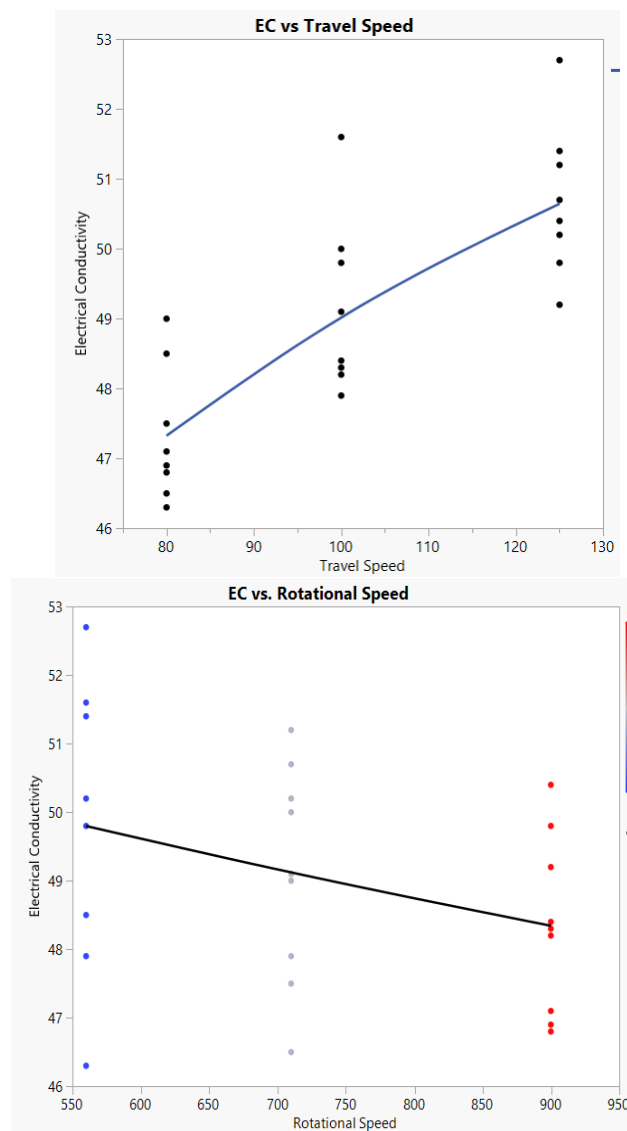


Fig. 5: Performance Curves for Various ANN Stages



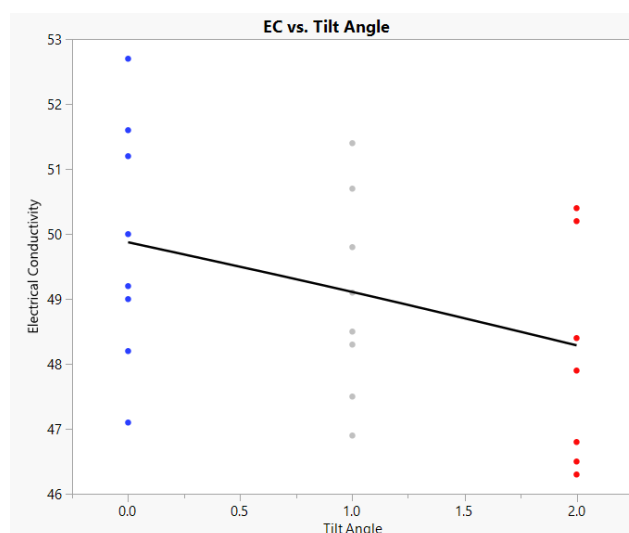


Fig.6. Main effect plot for means

Table 3. Optimum Results and Confirmation Test

Optimum parameter			Response	Actual	% Error
TS	RS	TA	parameter	Value	
125	560	0	52.7	54.1	2.58

5 Conclusion

The experimental data utilized for the ANN modelling process has been obtained through a set of properly chosen experiments, design of experiments as per the full factorial method. Precise prediction for Electrical Conductivity under the given conditions can be made with the developed neural network. Experiment results are in close agreement with networks prediction. The present technique was able to generate best combination of input parameters for the optimum level of Electrical Conductivity at the stir zone of FSW joint. Results show that high welding speed, low rotational speed, and low tilt angle are favourable for better conductivity at the joints. To be on the safer side, Friction Stir welded joint, the heat input and the material mixing in the stir zone have to be properly combined. Thus, Parameters that provide either lower input heat or higher input heat contribute negatively for sound FSW welds. Hence too high a traverse speed combined with low tool rpm needs to be avoided for conditions leading to weaker joint. Similarly, too low a travel speed combined with high tool rpm are detrimental for better joint.

References

- [1] Thomas, W.M.: Friction stir butt welding. Int. Patent No. PCT/GB92/02203,(1991).
- [2] Gibson, B.T., Lammlein, D.H., Prater, T.J., Longhurst, W.R., Cox, C.D., Ballun, M.C., Dharmaraj, K.J., Cook, G.E., Strauss, A.M.: Friction stir welding: Process, automation, and control. Journal of Manufacturing Processes 16(1), 56-73 (2014).
- [3] Mishra, R.S., Ma,Z.Y.: Friction stir welding and processing. Materials science and engineering: R: reports 50(1-2), 1-78 (2005).
- [4] Lakshminarayanan, A.K., Balasubramanian, V.: Process parameters optimization for friction stir welding of RDE-40 aluminium alloy using Taguchi technique. Transactions of Nonferrous Metals Society of China 18(3), 548-54 (2008).
- [5] Tewari, M., Jadoun, R. S., Kumar, S.: Effect of process parameters on tensile strength of Friction stir welded butt joints of thick AA1350 aluminum plates using Taguchi experimental design. Materials Today: Proceedings(2021).
- [6] Threadgill, P.L., Leonard A.J., Shercliff, H.R., Withers,P.J.: Friction stir welding of aluminium alloys. International Materials Reviews54(2), 49-93 (2009).
- [7] Sharma, D. N., Kumar, J. R.: Optimization of dross formation rate in plasma arc cutting process by response surface method. Materials Today: Proceedings 32, 354-357 (2020).
- [8] Shrama, D. N., Tewari, M.: Optimization of Friction Stir Welding Parameters using Combined Taguchi L9 and Genetic Algorithm. International Journal of Spatio-Temporal Data Science (2021).

- [9] Kapoor, V., Nandan, D.: Optimization of Physico-Chemical Parameters for the Production of Endoxylanase Using Combined Response Surface Method and Genetic Algorithm. In: V. Laha et al. (eds.), Optimization, Variational Analysis and Applications, Springer Proceedings in Mathematics & Statistics 355 (2021).
- [10] Kechagias, J., Iakovakis, V.: A neural network solution for LOM process performance. The International Journal of Advanced Manufacturing Technology 43(11-12), 1214-22 (2009).
- [11] Chrysosolouris, G., Lee, M., Ramsey, A.: Confidence interval prediction for neural network models. IEEE Transactions on neural networks 7(1), 229-32 (1996).
- [12] Özel, T., Karpaz, Y.: Predictive modeling of surface roughness and tool wear in hard turning using regression and neural networks. International journal of machine tools and manufacture 45(4-5), 467-79 (2005).
- [13] Benardos, P.G., Vosniakos, G.C.: Prediction of surface roughness in CNC face milling using neural networks and Taguchi's design of experiments. Robotics and Computer-Integrated Manufacturing 18(5-6), 343-54 (2002).
- [14] Munoz-Escalona, P., Maropoulos, P.G.: Artificial neural networks for surface roughness prediction when face milling Al 7075-T7351. Journal of Materials Engineering and Performance 19(2), 185-93 (2010).
- [15] Kechagias, J., Pappas, M., Karagiannis, S., Petropoulos, G., Iakovakis, V., Maropoulos, S.: An ANN approach on the optimization of the cutting parameters during CNC plasma-arc cutting. In Engineering Systems Design and Analysis 49187, 643-649(2010).
- [16] Ghetiya, N. D., Patel, K. M.: Prediction of tensile strength in friction stir welded aluminium alloy using artificial neural network. Procedia Technology 14, 274-281(2014).

Effect of Thickness on Laser Induced Nonlinear Optical Properties of Erbium Doped ZnO Thin Films

Vinay Kumari¹, Mukesh Kumar²

¹Department of Applied Sciences and Humanities (Physics), Panipat Institute of Engineering and Technology, Samalkha, India, 132102

²Department of Information Technology, Panipat Institute of Engineering and Technology, Samalkha, India, 132102

vinaykhatri76@gmail.com

Abstract: Erbium doped zinc oxide (EZO-doping concentration of 0.4 at. %) thin films of variable thickness were deposited on glass substrate by sol-gel method using spin coating technique. The film thickness was found to vary from 92-375 nm. EZO(0.4 at. %) films were characterized using X-ray diffraction (XRD), Scanning Electron Microscopy (SEM), UV-VIS-NIR transmission and single beam z scan method under illumination of frequency doubled(532 nm) Nd:YAG laser. The deposited films were found to have a preferential growth orientation along the ZnO (002) plane. All the films exhibited reverse saturable absorption (RSA) attributed to the two photon absorption.

Keywords: EZO, z scan, RSA, Two photon absorption

1.1 Introduction

The nonlinear refractive index and nonlinear absorption are two important optical phenomena which are extensively used in optical switching, optical limiting and labeling. In recent years, wide band gap semiconductors have been subjected to extensive studies because of the rising interest in the development of new nonlinear optical materials for potential applications in integrated optics. ZnO thin films have shown strong nonlinear properties by exhibiting second order, $\chi^{(2)}$ and third order susceptibility, $\chi^{(3)}$ [1-3]. The properties of ZnO thin films depend on the deposition and post deposition treatment conditions as these properties change significantly with (i) the nature of chosen doping element, (ii) the absorption of oxygen that takes place during film deposition, (iii) film deposition thickness and (iv) desorption during annealing treatment in a reducing atmosphere[4]. M.Sharma et al. has shown that surface roughness decreases up to a particular thickness (200nm) and then increased again depicting an increase in electrical properties at 200nm[5]. Doping is also highly beneficial in ZnO for getting high nonlinearities in it. Karthikeyan et.al has synthesized Na doped ZnO nanoparticles of various shapes through a simple wet chemical method and Dy³⁺ doped ZnO nanoparticles and got high value of nonlinear absorption coefficient than undoped ZnO[6]. K.Bahedi et al. has also reported effects of zirconium as doping agent and effect of surface roughness on the nonlinear optical properties resulting in the best value of nonlinear optical susceptibility $\chi^{(3)}$ with least roughness[7].

Previously we have optimized the value of doping concentration of erbium which is 0.4 at. %[8]. In the present work, we have been investigating the effect of thickness on the structural and optical (linear and nonlinear) properties of erbium (0.4 at. %) doped zinc oxide (EZO) thin films deposited by sol gel route as discussed in our previous articles[8]. These films have also been annealed for 1 hour in a muffle furnace [model]. The film thickness was found to vary from 92-375 nm using surface profilometer (Sloan Dektak 3-ST). The films have been characterized to study their structural, optical (linear and nonlinear) properties. X-ray diffraction (Rigaku, (RINT 2000) and scanning electron microscopy (Zeiss EVO-40 EP) have been used to study the crystallinity and growth of the films. The purpose of this work is to estimate an optimum thickness for EZO (0.4 at. %) films to be used as optical limiter in sensor protection. This application is related to the nonlinear properties of the material and for estimating the nonlinear optical properties viz. nonlinear absorption, nonlinear refraction, and nonlinear susceptibility of the films we have used single beam z-scan technique as described by the pioneer worker Sheik Bahae et al.[9]. Here for z-scan we have used a frequency doubled Nd:YAG laser (Quanta System, HYL-101) at 532 nm of energy per pulse of 200 mJ and repetition rate of 10 Hz with nanosecond laser pulses(5 ns). The beam waist was measured to be ~24.3 μ m. The transmitted light was recorded by a photodiode detector (Thorlabs, DET-110) and then was read by an oscilloscope (Tektronics, TDS-2024).

1.2 Structural Properties

1.2.1 XRD

It is evident from fig.1 that with increase in the film thickness up to 280 nm, intensity of (002) peak increases and FWHM decreases whereas, with film of thickness beyond 280 nm, the intensity of (002) peaks diminishes with a broadened FWHM. This indicates deterioration in crystalline quality of the films of thickness greater than 280 nm.

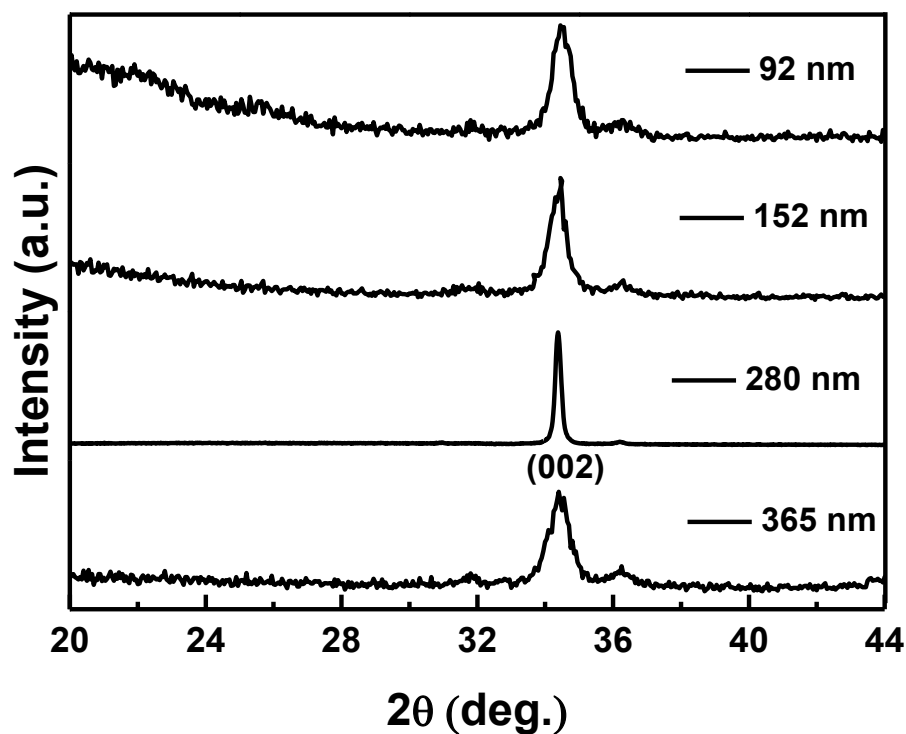


Fig.1: X-ray diffraction patterns for EZO(0.4 at.%) films with different thickness

Kinetics of crystal growth theorem implies that [10-11] the growing faces of a crystal are part of the free surface of the film. These crystal faces correspond to the crystal shape at equilibrium and are determined by the orientation of the crystal. A growth competition can start among the neighboring crystals according to their growth types (i.e., to their orientation). The faster growing crystals will grow over the slower growing ones. This competition is terminated when only crystals exhibiting the same type of crystal faces proceed to the free surface. This competitive crystal growth represents an orientation selection among the crystals resulting in the competitive growth texture [10]. This was probably why the crystallinity was increased when EZO (0.4 at.%) film became thicker.

It is clear from the Fig. 2 that as the film thickness increases from 92 to 280 nm, the intensity of (002) peak increases and the width of the peak narrows. This suggests that as the film becomes thicker, the crystalline quality of the films improves. The film quality is seen to deteriorate as the thickness goes up from 280 nm. With further increase in thickness of the films up to 365 nm, the intensity of (002) peak fall down .

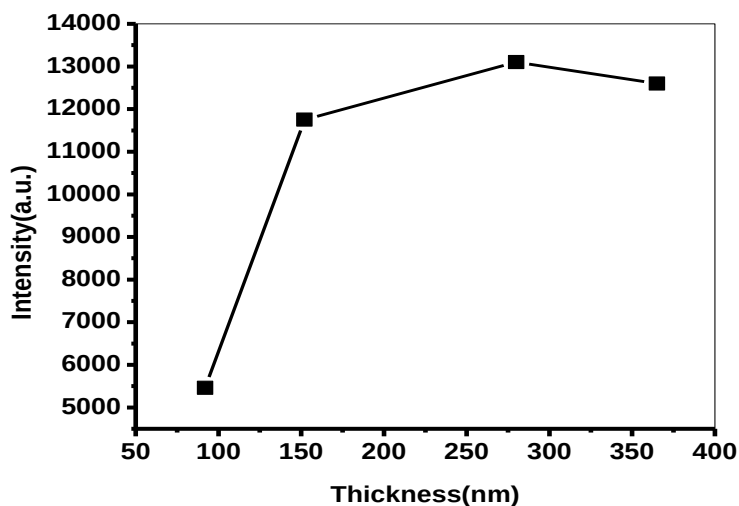


Fig.2: Variation of intensity of (002) peak with thickness of EZO (0.4 at.%) thin films

1.2.2 Crystallite Size

The crystallite size of EZO (0.4 at.%) films with different thickness were calculated using Scherrer's formula [12]. As depicted in Fig. 3, the crystallite size increased with increase of film thickness up to 280 nm and then decreased. Also, the FWHM decreased with thickness up to 280 nm and then showed an increase in conformity with the variation of crystallite size.

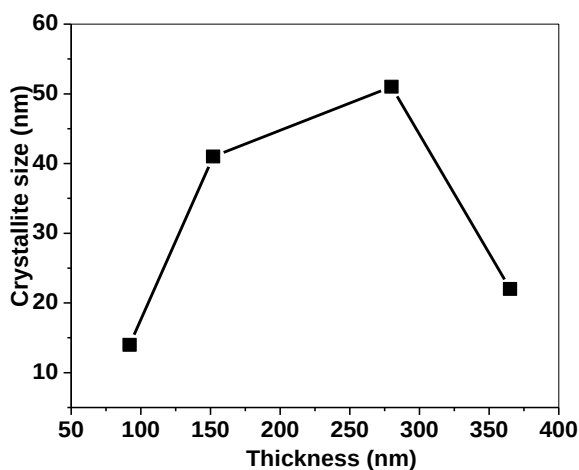


Fig.3: Variation of crystallite size of EZO (0.4 at.%) films with thickness

1.3 Nonlinear Optical Properties

Open Aperture Z-scan

Typical results of the open aperture z-scan measurements of all the samples are shown in Fig.4 corresponding to the far-field normalized transmittance, $T(z)$, as a function of the distance from the lens focus.

In the fig.4 it is clearly depicted that EZO (0.4 at.%) at thickness of 280 nm has deepest valley at focus indicating high value of nonlinear absorption in the sample. This observation is in conformity with the crystallinity and better transmission in the sample. Here, the theory of TPA process fitted well with the experimental curve shown in Fig.4 establishing TPA as the basic mechanism. But, there is also a possibility of higher order nonlinear processes such as free carrier absorption. Thus, one can infer that the observed nonlinearity is due to TPA followed by free carrier absorption in the thin film sample[13-15] .

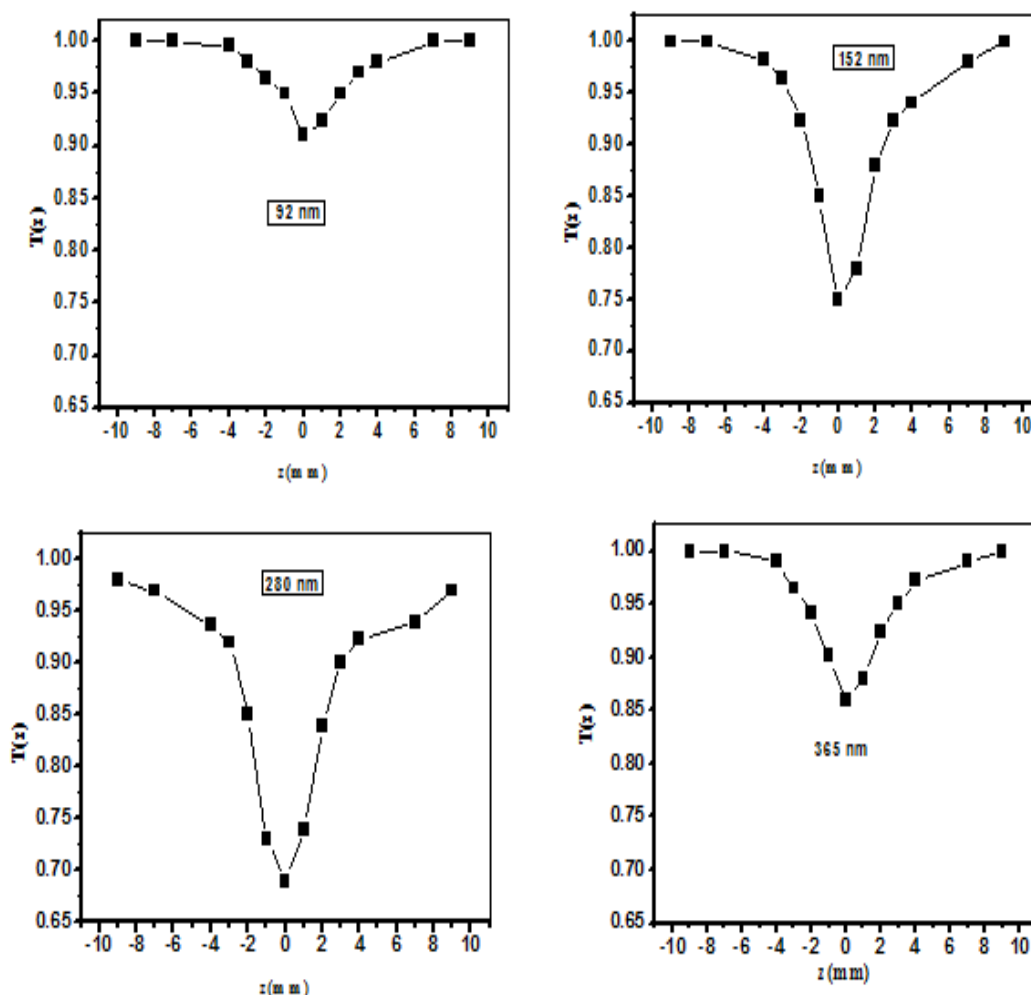


Fig.4: Open aperture z-scan measurements for EZO (0.4 at.%) thin films of variable thickness at 532 nm

The open aperture curve exhibits a normalized transmitted valley, indicating the presence of reverse saturable absorption (RSA) in the films [16]. Fig. 6 depicts that EZO (0.4 at.%) film with thickness 280 nm gives the optimized results for open aperture z-scan. EZO (0.4 at.%) at a thickness of 280 nm has the highest value of β (8.90×10^{-6} cm/W). The high values of nonlinear absorption coefficient (β) indicate that EZO (0.4 at.%) thin film samples have good nonlinear response .

The measured value of β of ZnO thin films are shown in table 1.

Table 1 -Measured values of absorption coefficient(β) of EZO(0.4 at.%) films of different thickness at 532 nm

Thickness (nm)	Nonlinear absorption coefficient (β) (cm/W)
92	2.56×10^{-6}
152	4.85×10^{-6}
280	8.90×10^{-6}
365	3.20×10^{-6}

Among EZO (0.4 at.%) samples of variable thickness the sample having thickness 280 nm exhibits the best nonlinear behaviour.

Conclusion :

EZO (0.4 at.%) thin film samples with variable thickness showed self defocusing nonlinearity and good nonlinear absorption behavior at off-resonant wavelengths. The nonlinear absorption increased with increasing thickness up to 280 nm at 532 nm. The observed nonlinear absorption is explained through TPA followed by weak free carrier absorption. It indicates that the material under investigation has very good nonlinear optical behavior and is a future material for the generation of third harmonic. EZO (0.4 at.%) sample at 280 nm thickness was found to be good candidate for optical limiting in comparison to all other sample. Hence the EZO (0.4 at. %) thin films can be used for sensor protection for avoiding them from being damaged under high irradiations.

References

- [1] G. I. Petrov, V. cheslavskiy and V. V. Yakovlev I. Ozerov, E. Chelnokov, and W. Marine, Near-resonant high order nonlinear absorption of ZnO thin films, Appl. Phys. Lett., 83 (2003) 3993.
- [2] H. Cao, J.Y. Wu, H.C. Ong, J.Y. Dai, R.P.H. Change, Second harmonic generation in laser ablated zinc oxide thin films, Appl. Phys. Lett. 73 (1998) 572.
- [3] A. Mitra, R.K. Thareja, Dependence of second harmonic generation on size of nanocrystallites of ZnO, Mod. Phys. Lett. B 15 (1999) 515.
- [4] J. D. Bhawalkar, G. S. He and P. N. Prasad, Nonlinear multiphoton processes in organic and polymeric materials, Rep. Prog. Phys. 59 (1996) 1041.
- [5] Mamta Sharma, R.M. Mehra, Effect of thickness on structural, electrical, optical and magnetic properties of Co and Al doped ZnO films deposited by sol-gel route, Applied Surface Science 255 (2008) 2527.
- [6] B. Karthikeyan, C. S. Suchand Sandeep, T. Pandiyarajan, P. Venkatesan and R. Philip, Optical and nonlinear absorption properties of Na doped ZnO nanoparticle dispersions, Appl. Phys. Lett. 95 (2009) 023118.
- [7] K. Bahedi, M. Addou, M. El Jouad, Z. Sofiani, M. Alaoui Lamrani, T. El Habbani, N. Fellahi, S. Bayoud, L. Dghoughi, B. Sahraoui, Z. Essaidi, Diagnostic study of the roughness surface effect of zirconium on the third-order nonlinear-optical properties of thin films based on zinc oxide nanomaterials, Applied Surface Science 255 (2009) 4693.
- [8] Vinay Kumari, Vinod Kumar, B. P. Malik, R. M. Mehra and Devendra Mohan. "Nonlinear Optical Properties of Erbium Doped Zinc Oxide (EZO) Thin Films" Optics Communication (Elsevier), 285, 2182-2188 (2012).
- [9] M S Bahae, A A Said, T H Wei, D J Hagan and E W Van Stryland, Sensitive measurement of optical nonlinearities using a single beam, IEEE J. Quantum Electron. 14 (1990) 760.
- [10] G. Knuyt, C. Quaeys, J. D. Haen, L. M. Stals, A quantitative model for the evolution from random orientation to a unique texture in PVD thin film growth, Thin Solid Films 258 (1995) 159.
- [11] P. B. Barna, M. Adamik, Fundamental structure forming phenomena of polycrystalline films and the structure zone models, Thin Solid Films 317 (1998) 27.
- [12] L. V. Azaroff, Elements of X-ray Crystallography (McGraw Hill, New York, 1968).
- [13] X. Zhang, H. Fang, S. Tang, and W. Ji, Determination of two-photon-generated free-carrier lifetime in semiconductors by a single-beam Z-scan technique, Appl. Phys. B: Lasers Opt. 65 (1997) 549.
- [14] M. Y. Han, W. Huang, C. H. Chew, L. M. Gan, X. J. Zhang, and W. Ji, Large Nonlinear Absorption in Coated Ag₂S/CdS Nanoparticles by Inverse Microemulsion, J. Phys. Chem. B 102 (1998) 1884.
- [15] L. Irimpan, D. Ambika, V. Kumar, V.P.N. Nampoori, P. Radhakrishnan, Effect of annealing on the spectral and nonlinear optical characteristics of thin films of nano-ZnO, Journal of Applied Physics 104 (2008) 033118.

- [16] F. M. Quereshi, S. J. Martin, X. Long, D. D. C. Bradley, F. Z. Heneri, W. J. Balu, E. C. Smith, C. H. Wang, A. K. Kar, and H. L. Anderson, Optical limiting properties of a zinc porphyrin polymer and its dimer and monomer model compounds, *Chem. Phys.* 231 (1998) 87.

NOVEMBER 19-20, 2021

ABOUT THE CONFERENCE

IRCET-2021



The International Conference on Innovation, Research and Challenges in Emerging Technologies (IRCET-2021) provides a global platform for researchers, scientists, academicians, policymakers, and industry experts to share experiences, discuss research findings and acquire desired knowledge in the subject from around the world with many networking opportunities. Authors from around the globe submitted their research papers on various aspects of emerging technologies, including artificial intelligence, machine learning, deep learning, data science, data analytics, virtual reality, augmented reality, blockchain, internet of things, cybersecurity, cryptography and privacy, cryptocurrencies, industry automation, telecommunications, surveillance technologies, neuroscience, military technologies image processing etc. Selected papers will be considered for publication in SCOPUS indexed journal International Journal of Information Technology (BJIT) published by Springer and SCOPUS indexed Springer series Communications in Computer Sciences.



Published by :
ABC Computers Printers

PIET Campus, 70, Milestone GT Road, Samalkha, Panipat, Haryana - 132102 India
Call: 1800 120 6884 (Toll Free), 0180-2569700, 2569799, 2569699