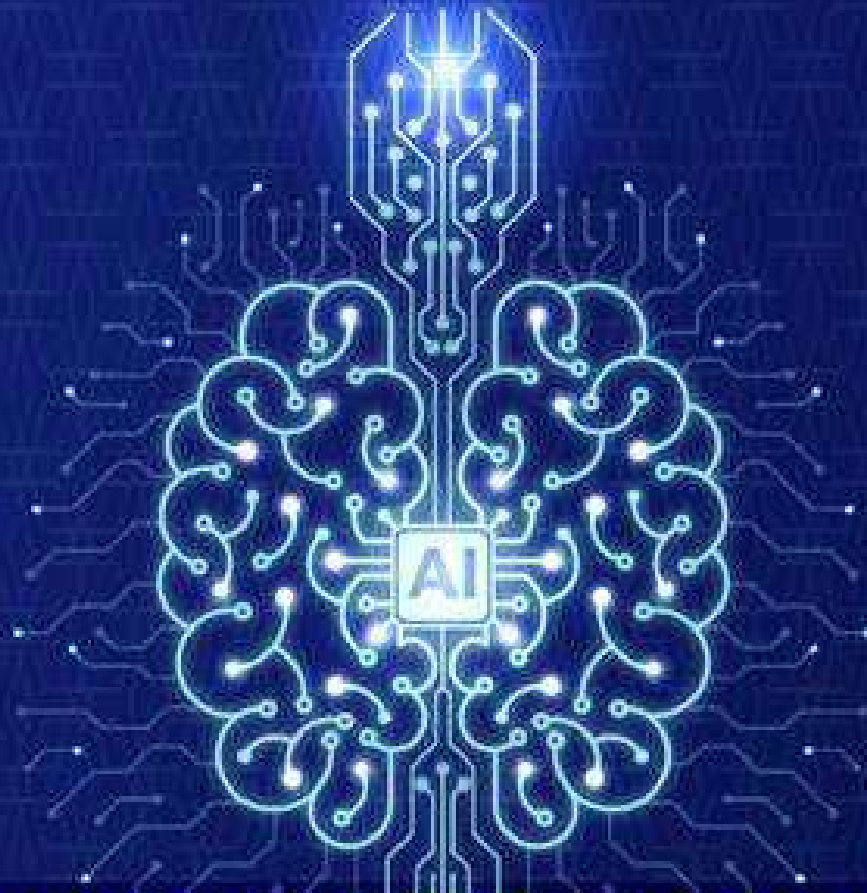




SRM INSTITUTE OF SCIENCE AND TECHNOLOGY
Ramapuram Campus

COLLEGE OF SCIENCE AND HUMANITIES
(A Place for Transformation)

DEPARTMENT OF COMPUTER APPLICATIONS (BCA)



**1ST INTERNATIONAL CONFERENCE ON
ARTIFICIAL INTELLIGENCE AND
DATA SCIENCE (ICAIDS-2022)**

Date :04.03.2022 and 05.03.2022

Editor

Dr. Agusthiyar R

Professor and Head, Department of Computer Applications, SRMIST, Ramapuram Campus.

Conference Secretaries

Mrs. S. Suriya, Asst. Professor / Mrs. J. Shyamala Devi, Asst. Professor



978-93-9620-078-3



**PROCEEDINGS OF THE
1ST INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND DATA
SCIENCE (ICAIDS-2022)**

DATE: 4 & 5th MARCH 2022

@SRM Institute of Science and Technology, Ramapuram, Chennai-89 August 2022. All rights reserved.

No part of the material protected by this Copyright notice may be reproduced or utilized in any form or by any means, electronic or mechanical including photocopying, recording or by any information storage and retrieval system, without prior written permission from the copyright owner. Statement and options in these proceedings or those of the contributors and the publisher assume no responsibility for them.

SRM Institute of Science and Technology-Ramapuram Campus

Bharathi Salai, Ramapuram,

Chennai – 600 089.

Website: <https://srmrmp.edu.in/about-srm/srm-in-focus/>

Mail-Id: hod.dca.rmp@srmist.edu.in



978-93-5620-075-3



**PROCEEDINGS OF THE
1ST INTERNATIONAL CONFERENCE ON ARTIFICIAL
INTELLIGENCE AND DATA SCIENCE (ICAIDS-2022)**

DATE: 4 & 5th MARCH 2022

EDITORS

Dr. Agusthiyar . R

Conference Secretary

Mrs. S.Suriya, Assistant Professor

Mrs. J.Shyamala Devi, Assistant Professor

Organized by

SRM Institute of Science and Technology,

Faculty of Science and Humanities

Department of Computer Applications (BCA)

Ramapuram, Chennai – 600 089.

Website :fsh.srmrmp.edu.in

ORGANIZING COMMITTEE

CHIEF PATRONS

Dr. T.R. Paarivendhar

Chancellor, SRM Institute of Science and Technology

Dr. R. Shivakumar

Chairman, SRM Group of Institutions

Ramapuram & Trichy Campus

PATRONS

Dr. N. Sethuraman, Registrar, SRM IST

Dr. V. Subbiah Bharathi, Director, Ramapuram Campus

Dr. C. Sundar, Dean (S&H), Ramapuram Campus

Dr. J. Dhilipan, VP – Admin (S&H)

Dr. N. Asokan, VP – Academic (S&H)

CONVENOR

Dr. R. Agusthiyar, Professor

Head - Computer Applications (BCA)

Conference Secretary

Mrs. S. Suriya, Assistant Professor

Mrs. J. Shyamala Devi, Assistant Professor

COMMITTEE MEMBERS

Dr. K. Pushpalatha, Associate Professor

Mrs. M. Divya, Assistant Professor

Mrs.P M Kavitha, Assistant Professor

Mrs.S.Sindhu, Assistant Professor

Dr.N.Vijayalakshmi, Assistant Professor

Mr.D.B.Shanmugam, Assistant Professor

Mrs.V.Devi, Assistant Professor

Dr.S.Jayachandran, Assistant Professor

Dr.T.S.Suganya, Assistant Professor



OUR ASSOCIATES



COMMITTEE MEMBERS

Mrs. Sandhya Palani

Mrs. Bragathalakshmi

Mrs.Sri kamani

Mr.Abhishek Sharma

Mr.Madhan Kartheesan

Ms.Deepthi



SRM Group of Educational Institutions

Dr. R. Shivakumar, M.D.,
Chairman: SRM Group of Institutions
Ramapuram & Trichy Campus



Dr. R. Shivakumar, M.D.,

CHAIRMAN'S MESSAGE

The most wonderful creation of God is Human being.
The most wonderful creation of Human being is Education.
The most wonderful creation of education is Morality.

I take this opportunity to reminisce over the glorious years of the existence and growth of College of Science and Humanities SRM Institute of Science and Technology and my heart fills with love, admiration and pride for the fertile soil which the college provides to the 17 years old, who enter its portals as secondary students, to leave as responsible human beings. From the aspect of infrastructure to the delivery of knowledge SRM Institute of Science and Technology now makes excellent facilities available to its students and teachers in impressive forms. We have had very good results so far, our past students are also doing well in lives. We have had full cooperation from the parents, teachers, staff and administration. Other educational institutions look upto us for leadership in the field of education and innovative approaches to teaching and guidance of the students.

I am extremely happy to learn that the Department of BCA is organizing the First International Conference on Artificial Intelligence and Data Science (ICAIDS-2022) in Association with Object Automation Software Solutions Pvt Ltd, IBM, OpenPower, Onstitute and X-Scale Solutions to be held on 4th and 5th of March, 2022.



SRM Group of Educational Institutions

Dr. R. Shivakumar, M.D.,
Chairman: SRM Group of Institutions
Ramapuram & Trichy Campus

I'm sure this would be a flat form be great scholars from institutions all over the world to unite and exhibit the glory of knowledge acquisition. I wish the department of Computer Application (BCA) and the team a grand success and the contributors best wishes to the of the proceeding. My best wishes to all who are associated with us in the growth of this institution.

R. Shivakumar

CHAIRMAN

SRM Ramapuram & Trichy Campuses



SRM Group of Educational Institutions



Mr.S.Niranjan, M.S.,

Co-Chairman's Message

Companies recognize innovations as a positive practice in their day to day business operations. The emerging technologies, evolving customer expectations and the globalized market place catalyze the need for companies to quickly identify and adapt these new opportunities to capitalize faster than ever.

Its indeed my pleasure to place on record the remarkable years of existence and growth of the SRMIST , College of Science and Humanities and I am immensely pleased to note that the Department of Computer Applications (BCA) for conducting its First International Conference on Artificial Intelligence and Data Science (ICAIDS-2022) in association with Object Automation Software Solutions Pvt ltd, IBM, OpenPower, Onstitute and X-Scale Solutions between 4th and 5th March 2022.

I am sure that this international platform brings together the brightest minds from various sectors of the industry for exchanging ideas to take the world next step forward. I appreciate the active participation of all the distinguished International & National Speakers, Faculty, Research Scholars and Students of this International Conference to make this event a memorable one .

With Best wishes

Co-Chairman

SRM Ramapuram & Trichy Campus



MESSAGE

I am extremely happy to note that the Department of Computer Applications SRMIST, Ramapuram campus is organizing an International Conference on Artificial Intelligence and Data Science (ICAIDS-2022) in Association with Object Automation Software Solutions Pvt. Ltd, IBM, Open Power, Institute and X-Scale Solutions on March 4 & 5, 2022.

Globally, Data Science and AI are currently reigning technologies that have conquered industries around the world due to the massive explosion in data and the increasing need for business to rely on data for various reasons. They are the key enablers for emerging technological revolutions that have reshaped our work and life style. Although data science techniques are maturing and becoming more openly used by the average user, its tools continue to evolve and does bring in a lot of challenges. It is here that the expertise of professionals who transform data into insights are required. It is the need of the hour to gain in depth knowledge in such areas as it will become part and parcel of future work life. I am sure that this conference will touch upon the much needed research culture among the participants and trigger interactions particularly on the recent advancements and future demands in AI and Data Science.

I wish the conference a grand success.


Prof. C. Muthamizhchelvan 2/3/22

Dr. S. Ponnusamy, Ph.D.,
Registrar



SRM
INSTITUTE OF SCIENCE & TECHNOLOGY
(Deemed to be University u/s 3 of UGC Act, 1956)



Greetings from SRM Institute of Science and Technology!!!

It is an honour and a pleasure to send you a welcome address, just to underline how the research initiatives across all the recent trends cover the way for the industrial world to strive forward with huge advancements. It is happy to note that the Department of Computer Applications (BCA) is Organizing **1st International Conference on Artificial Intelligence and Data Science (ICAIDS-2022) in Association with Object Automation Software Solutions Pvt Ltd, IBM, Open Power, Onstitute and X-Scale Solutions on March 4th and 5th, 2022 .**

The quest for knowledge has been from the beginning of time but knowledge only becomes valuable when it is disseminated and applied to benefit humankind. It is hoped that this conference will be a platform to gather and disseminate the latest knowledge in recent advancements in emerging areas of diversified research fields covered during this conference.

It is envisaged that the intellectual discourse will result in future collaborations between Universities, research institutions and industry both locally and internationally.

I congratulate the conveners, coordinators and organizing committee members for their dedication and hard work, who have been working tirelessly for organizing this conference in a befitting manner.

I wish them for their endeavors to spread knowledge.

A handwritten signature in green ink, appearing to be 'S. Ponnusamy'.

Registrar
Registrar

SRM Institute of Science and Technology
SRM Nagar, Kattankulathur - 603 203
Chengalpattu Dist, Tamilnadu, India.



SRM Group of Educational Institutions

CHIEF DIRECTOR MESSAGE



Science and Technology has become a part and parcel of man's life. Man can never be isolated from the influence of Science and Technology. Imparting value based education to the common masses is a sacred act which our educational institutions aim at and are imparting the same meticulously down the years. Immense care is taken to impart quality education through innovative methods, so that the students of our institution become eminent scholars in their particular field and at the same time it is ensured that avenues are provided for the faculty members to update themselves to stand the test of time.

The First International Conference on Artificial Intelligence and Data Science (ICAIDS-2022) in Association with Object Automation Software Solutions Pvt Ltd, IBM, OpenPower, Onstitute and X-Scale Solutions on March 4th and 5th, 2022 will be an enriching experience of knowledge accumulation and assimilation.

My best wishes to the department of Computer Applications (BCA) of the College of Science and Humanities and the entire organizing team.

A handwritten signature in red ink, appearing to read 'N. Sethuraman', positioned above the printed name.

Prof. N. Sethuraman

Chief Director.

Dr. V. Subbiah Bharathi, Director



I am indeed proud that SRM Institute of Science and Technology is in the journey of enriching lives and lifestyles of thousands and thousands of students and staff. We can humbly claim to have added professionalism to impart knowledge in these years and will continue to do so in the years ahead. It is quite inspiring to watch and witness the research potential of our faculty and students at various stages and situations each day. The management, administration and the faculty members have been supportive of the various activities that were undertaken by in view of helping researchers reach the summit of perfection and professionalism in whatever task they took on.

I take this opportunity to congratulate the Dean and his team of committed faculty members for their commitment towards value based quality teaching and unstinted efforts put in organizing the First International Conference on Artificial Intelligence and Data Science (ICAIDS-2022) in association with Object Automation Software Solutions Pvt Ltd, IBM, OpenPower, Onstitute and X-Scale Solutions on 4th and 5th March 2022 with all its objectives being fulfilled. My hearty congratulations and best wishes to the entire team of College of Science and Humanities. I thank the management and the college administration for standing with us for progress in all that we endeavour.

SRM Institute of Science and Technology

Ramapuram Campus, Bharathi Salai, Ramapuram, Chennai 600 089. Tamil Nadu, India

Phone: +91 44 4392 3133|E-Mail: office.admin@srmist.edu.in|Website: <https://srmrmp.edu.in>

Dr.C.Sundar
Dean
College of Science and Humanities



Many people feel as if they're adrift in the world. They work hard, but they don't seem to get anywhere worthwhile. A key reason that they feel this way is that they haven't spent enough time thinking about what they want from life, and haven't set themselves formal goals. After all, would one set out on a major journey with no real idea of their destination? Probably not! Goal setting is a powerful process for thinking about one's ideal future, and for motivating oneself to turn one's vision of this future into reality. The process of setting goals helps one choose where he/she wants to go in life. By knowing precisely what one wants to achieve, one knows where he/she has to concentrate their efforts.

College of Science and Humanities of SRM Institute of Science and Technology strives very hard in focusing Faculty members, students and the researchers towards definite goals and finds viable means to enable them to achieve the goals fixed. I wish to congratulate the entire teaching and non-teaching faculty, the students, on their commendable efforts in organizing the First International Conference on Artificial Intelligence and Data Science (ICAIDS-2022) in association with Object Automation Software Solutions Pvt. Ltd, IBM, OpenPower, Onstitute and X-Scale Solutions on 4th and 5th March 2022. It is a matter of great pride that the College has made consistent progress towards development in providing opportunities for the teaching and research communities to enrich themselves with the latest trends and updates, year on year, in academic and co-curricular activities.

The college management and administration is on their heels day in and day out to turn dreams into realities, imaginations into real incidents and wishes into glowing actions.

I have great pleasure in conveying my best wishes to all those who have been responsible in organizing conference of this sort.

Dr. J. Dhilipan, VP Admin



Dear Professors and Researchers,

It is my privilege and honor to welcome you all to the 1st International Conference on Artificial Intelligence and Data Science (ICAIDS-2022) in Association with Object Automation Software Solutions Pvt Ltd, IBM, OpenPower, Onstitute and X-Scale Solutions on March 4th and 5th 2022 at SRM Institute of Science and Technology , Ramapuram Campus.

The main goal of organizing this conference is to share and enhance the knowledge of each and every individual in this fast-moving Information Era. We have given a good opportunity for those who have a thirst in knowing the present technological developments and also share their ideas. Additionally, this Conference will also facilitate the participants to expose and share various novel ideas. The Conference aims to bridge the researchers working in academia and other professionals through research presentations and keynote addresses in current technological trends. It reflects the growing importance of Artificial Intelligence and Data Science as a field of research and practice for contribution and better opportunities in the IT industry. You will get ample opportunities to widen your knowledge and network.

I want to thank in advance the conference committee for extending their valuable time in organizing the program and all the authors, reviewers, and other contributors for their sparkling efforts and their belief in the excellence of ICAIDS - 2022.

I cordially invite all the enthusiasts to participate with full vigor in this celebrated event which can give immense exposure and global opportunities to all.

My Best Wishes to you all.

Friday, 4th March, 2022

Dr.N.Asokan
Vice Principal Academic
College of Science and Humanities
SRM IST Ramapuram Campus
Chennai, India



Message

Very Warm and Very Happy Greetings to all. I am delighted that Department of Computer Applications (BCA) is Organizing 1st International Conference on “Artificial Intelligence and Data Science (ICAIDS-2022) in Association with Object Automation Software Solutions Pvt Ltd, IBM, OpenPower, Onstitute and X-Scale Solutions on March 4th and 5th, 2022”.

The quest for knowledge has been from the beginning of time but knowledge only becomes valuable when it is disseminated and applied to benefit humankind. It is hoped that ICAIDS-2022 will be a platform to gather and disseminate the latest knowledge in recent advancements in emerging areas of diversified research fields covered during this conference.

This conference is a tumultuous opportunity for Academicians, Scientist, Researchers and Students to demonstrate their research skills, to share their experience and research findings and applications of science and engineering. This conference will be one for us to share our thoughts and exchange ideas on how to chart our journey forward to reach new heights.

My heartiest congratulations to all the faculty members of BCA department and all the participants from all over the world for their enthusiastic participation and invaluable time invested as part of their Life Long Self Learning.

I am privileged to say that this conference will definitely bring out the new knowledge to solve the day after tomorrow’s social problems through technology.

I wish the conference a grand success.

Dr.N.Asokan

Bharathi Salai, Ramapuram, Chennai 600 089. Tamil Nadu, India

Phone: 0 44 4392 3133,044 4393 3042, 3145

E-Mail: office.fsh.rmp@srmist.edu.in Website: www.srmist.edu.in



Dr. Agusthiyar R

Professor & Head

Greetings to all. This institution stands tall for its successful, developed holistic system where the formal education has been intricately woven with moral, spiritual and social education. The dynamism of the young talent blooming in our garden is being tapped; the skills and the potentialities of its students and faculty members are being mined out and chiseled. I would like to congratulate the staff, the students of the Department of Computer Applications (BCA) and all the others directly and indirectly associated in organizing the First International Conference on Artificial Intelligence and Data Science (ICAIDS-2022) in association with Object Automation Software Solutions Pvt Ltd, IBM, OpenPower, Onstitute and X-Scale Solutions on 4th and 5th March 2022. The mill started rolling long back and the efforts par excellence of the committed faculty members has made all the difference.

We have an outstanding legacy of magnanimity, benevolence, astute foresight, and remarkable leadership. We are but duty-bound to build upon this legacy to provide enriching holistic experience of education for the faulty members, researchers all over the world.

I firmly commit to sustain the engagement and pace set by a vibrant and motivated faculty members through the years and to raise the bar even higher for the ultimate benefit of the research community and for the development of the nation.

I am very happy to wish the conference a grand success.

Sandhya Palani
Director of Business Development
Object Automation Software Solutions Pvt Ltd



On behalf of Object Automation, I am delighted in acknowledging the 1st International Conference on Artificial Intelligence and Data Science (ICAIDS - 2022) organized by the Department of Computer Application (BCA) of SRM Institute of Science and Technology in association with Object Automation Software Solutions Pvt Ltd.

We are a leading Software Development and Training organization specialized in Artificial Intelligence providing immersive consumer experiences and developing the next generation of an intelligent system to make people discover the Power of AI and say Hello to the future.

We are honoured to be an elite partner with leading organizations like IBM, HPE, Microsoft Azure, Dell, Lenovo, Open POWER, Red hat, etc.

We have joined hands with universities, start-ups, and industries to bring the latest AI technology, allowing students, researchers, faculties and other trainees to use AI supercomputers, to learn and optimize AI applications in a distributed environment which helps establish their overall understanding of Artificial Intelligence from the perspective of their future.

We train students in Python programming, Data Science, Machine Learning, Deep learning, Bio python, AI in Cybersecurity, NLP, Kubernetes, Open stack, Quantum Computing, Data engineering on Microsoft Azure, and some analytic tools like Microsoft PowerBI and Tableau.

We also contribute to the Centre of Excellence where IBM is teaming up with universities, start-ups, ISV, and industries to help develop further the impact of AI solutions for real-world opportunities in association with the power servers. The IBM enterprise lab will play a major role in researching and developing emerging AI technologies.

We are honored to be part of this brilliant event and appreciate the organizing committee for showing a keen interest in organizing a successful Conference and contributing new ideas and research findings.

For Object Automation Software Solutions Pvt Ltd



Authorized Signature

INDEX

S.No	Title	Page. No
1	CLASSIFICATION OF PRODUCT REVIEWS USING SENTIMENT ANALYSIS ON DATASET GENERATED USING WEB SCRAPING Prof. Ashwinee Pulumkar , Kasturi Joshi, Riya Joshi, Rohan Joshi, Sumedh Joshi	01 - 05
2	DEVELOPMENT APPROACH: ELECTRICITY CONSUMPTION AND PREDICTION USING MACHINE LEARNING TECHNIQUES Palak Sharma, Brajlata Chauhan , Amrindra Pal	06 - 10
3	DEVELOPMENTS IN MACHINE LEARNING MODELS FOR THE PREDICTION OF RENEWABLE ENERGY Rashida Tabassum, Brajlata Chauhan, Amrindra Pal	11-15
4	SECURE DATA SHARING USING BLOCKCHAIN IN VEHICULAR SOCIAL NETWORKS Mamidala Sruthi, Mucha Swetha , Dadi Ramesh	16-19
5	WILL CSK WIN IPL 2022? APPLICATION OF MACHINE LEARNING ALGORITHMS IN SPORTS Uma Maheswari B, Selva Gomathy R, Kavitha D, Sujatha R	20-24
6	STUDENT ATTENDANCE AUTOMATION SYSTEM USING FACIAL RECOGNITION R. Vaibhav , D.Sudhagar	25-32
7	ANALYZING DATA MINING TECHNIQUES IN EDUCATIONAL SYSTEMS Ms. Nisha Raveendran , Dr. N. Vijayalakshmi,	33-36
8	GENERATIVE ADVERSARIAL NETWORKS BASED APPROACH ON REAL WORLD : THEORY AND APPLICATIONS Ashly Ann Jo , Ebin Deni Raj	37-44
9	POINT OF VIEW ON ENTERPRISE-WIDE TEXT SUMMARIZATION APPROACHES Sabari Rajan, Balasubramanian Vijayasankar, Selvakuberan karuapasamy , Subhashini Lakshminarayanan	45-48

10	ENHANCED CRYPTOGRAPHIC SOLUTIONS TO PROTECT MESSAGES DURING COMMUNICATION IN VEHICULAR CLOUD COMPUTING A. Sheela Rini, Dr C. Meena	49-56
11	REVIEW ON INTRUSION DETECTION SYSTEM IN WIRELESS SENSOR NETWORK Jyoti Srivastava, Jay Prakash, Anu Raj	57-64
12	ANIMAL DETECTION BY YOLO COCO MODEL USING IMAGES G.Elaiyaraja, T.K.Kalaiarasan and C.Manikanta	65-76
13	FACE MASK DETECTION USING TENSORFLOW, RESNET, AND MACHINE LEARNING ALGORITHM Mrs. S. Suriya, Agusthiyar R, Shyamala Devi J, S. ABISHEK, R.HARISH, ANTOS MARIA KARUNAI	77-80
14	IDENTIFYING IMAGES IN MULTIFRAME SEGMENTATION IMAGE CLASSIFICATION USING SVM MODEL IMPLEMENTS WITH OPENCV Mrs.S.Sindhu, Mrs.J.Shyamala Devi, Mr.B N Swaminathan,Mr.Sanjay kumar, Mr.M Sanjai	81-84
15	AN ENSEMBLE APPROACH FOR DOCUMENT SUMMARIZATION Vetrivel Panneerselvam, Srushti Gajbhiye, Selvakuberan Karuppasamy, Subhashini Lakshminarayanan	85-89
16	EMOTION AND SENTIMENT CLASSIFICATION USING TRANSFORMER MODELS Gem Rose Kuriakose, Prabhitha Nagarajan, Selvakuberan Karuppasamy and Subhashini Lakshminarayanan	90-97
17	SPAM DETECTION BASED ON DEEP LEARNING TECHNIQUE K Ranjith Reddy, Dr. Sanjay Chaudhary	98-102
18	TOOLS AND TECHNIQUES FOR DETECTING SARCASM – A SURVEY Ayesha Shakith, Dr. M. Kriushanth, Dr. L. Arockiam	103-111
19	A REVIEW ON MACHINE LEARNING APPROACH OF VIRTUAL SCREENING AND DRUG TARGET IDENTIFICATION IN DRUG DISCOVERY G.Hemalatha, Dr.P.Sasikala, Dr.R. Reka,	112-118

20	SURVEY ON AN EFFICIENT CREDIT CARD FRAUD DETECTION USING BIG DATA ANALYTICS WITH MACHINE LEARNING APPROACHES Ms V.Suganthi, Dr.J.Jebathangam	119-123
21	A SYSTEMATIC STUDY ON APPLICATIONS OF DIGITAL IMAGE PROCESSING Dr.D.Sasirekha, Dr.A.Ambeth Raja	124-130
22	QUESTION-ANSWERING WITH PERSONALIZED RESPONSE RESTRUCTURING Srushiti Gajbhiye, Anshuman Mahapatra, Selvakuberan Karuppasamy ,Subhashini Lakshminarayanan	131-134
23	MULTILINGUAL OPEN DOMAIN ENTITY EXTRACTION Priyank Bhardwaj, Venkatesan Paramasivam, Balasubramanian Vijayasankar, Selvakuberan Karuppasamy ,Subhashini Lakshminarayanan	135-138
24	MEDICAL IMAGE SEGMENTATION AND CLASSIFICATION USING NEURAL NETWORKS Dr.D.Suganthi, Mr.C.Jeyaganthan	139-142
25	FUZZY K-MEANS CLUSTERING TECHNIQUE FOR ENERGY EFFICIENT ROUTING IN WIRELESS SENSOR NETWORKS Dr.S.Lavanya, Dr.P.Calista Bebe, Dr.C.Sudha	143-149
26	FAKE NEWS DETECTION USING MACHINE LEARNING Dr. T. S. Suganya, Deepthi Jayadevan, Nethra R, Praveen Kumar	150-152
27	AUTOMATIC NUMBER PLATE RECOGNITION SYSTEM USING MORPHOLOGICAL ALGORITHM WITH OCR AND OSTU Dr.S.Jayachandran, Mrs.V.Devi, Mr.J.Sanurag Nair	153-155
28	PIXEL HIGH DENSITY NOISE FILTER METHOD FOR DENOISING IMAGES USING IMAGE PROCESSING TECHNIQUES Suriya Priyadharsini M, Dr. J. G. R. Sathiaseelan	156-161

CLASSIFICATION OF PRODUCT REVIEWS USING SENTIMENT ANALYSIS ON DATASET GENERATED USING WEB SCRAPING

Prof. Ashwinee Puluojkar
Department of Electronics and
Telecommunication
Vishwakarma Institute of Technology
Pune, India
ashwinee.barbadekar@vit.edu

Rohan Joshi
Department of Electronics and
Telecommunication
Vishwakarma Institute of Technology
Pune, India
rohan.joshi19@vit.edu

Kasturi Joshi
Department of Electronics and
Telecommunication
Vishwakarma Institute of Technology
Pune, India
kasturi.joshi19@vit.edu

Sumedh Joshi
Department of Electronics and
Telecommunication
Vishwakarma Institute of Technology
Pune, India
sumedh.joshi19@vit.edu

Riya Joshi
Department of Electronics and
Telecommunication
Vishwakarma Institute of Technology
Pune, India
riya.joshi19@vit.edu

Abstract: - The rise in E-commerce websites has widened the utility of various products or services. Thus, companies want real time reviews on what the consumers think about their product. Reviews are the primary source of commercial feedback that the companies can use to improve as well as promote their services. In this paper, we have proposed a model which is optimized in predicting the Sentiment of a text, primarily product reviews from E-commerce websites and classifying them as positive or negative using Natural Language processing and machine learning algorithms like Logistic regression, Decision Tree, Random Forest, SVM and KNN. The most important prerequisite to solving such a problem is the dataset which needs to be of honest and insightful reviews which can help in making accurate predictions in analysing the sentiment behind it. Thus, we have used Web-scraping to extract real and latest updated dataset consisting of reviews of a product from an e-commerce website.

Keywords — *Machine Learning, Natural Language Processing, Sentiment Analysis, Text pre-processing, TFIDF vectorization, Web Scraping.*

I. INTRODUCTION

The increasing use of E-commerce websites has given its users a wide range of products and services to choose from. While buying products online, the customers rely heavily on the reviews and feedback of the other customers. These reviews provide a better and a deeper understanding of the quality and the useability of the product and help in narrowing down the search for the user. This improves the user experience as the user can make an informed decision by reading these reviews. The end product of our model, which is the analysis of the feedback, will also help the E-commerce companies to improve their product quality and/or service quality on their part. Sentiment Analysis helps in processing large amounts of reviews and feedback for the E-commerce companies to cater to the demands or grievances of their customers and contributes to a better and smoother interaction between the company and the users.

This project focuses on Analysing the Sentiment behind the Product Reviews that are available for products on E-commerce websites by using Natural Language

Processing. It takes factors such as the amount of positive and negative words used in the sentence, the ratings or the number of stars the product has received from a particular user and the average rating of that product. While implementing the project we have taken into consideration the latest product reviews that were available to us. The datasets on product reviews of various E-commerce websites that were available to us were not up to date and hence performing any kind of analysis on those reviews would be redundant at this time. To overcome this, we performed web scraping using Python on a website, through which we independently created our own dataset of the latest reviews. We used the information such as the product review and rating of the latest comments and created a data frame out of it to perform our analysis and apply algorithms to get the sentiment behind these products. For better classification of reviews, the dataset has been processed using text pre-processing and TF-IDF Vectorization. Decision Tree, Logistic Regression and SVM models had the highest accuracy. These analysed customer reviews can be greatly insightful and of immense help in case of product recommendation as well as the owners knowing about the actual feedback on their product.

The objectives of the project are:

1. Generating a dataset of Product Reviews by scraping the web pages of a website using Python
2. Performing Sentiment Analysis on the reviews
3. Identifying the reviews as positive or negative.
4. Predicting the category of new reviews

II. LITERATURE REVIEW

Sentiment Analysis is a way of understanding and identifying the tone or emotion behind a text or a paragraph which can be further used to make some inferences. Natural Language Processing is most commonly used for this purpose. It uses a set of in-built functions that are applied on the target text.

[1] This paper talks about detection of comment types, especially hate speech found on Twitter data. The benchmark dataset is almost 16,000 annotated tweets which are experimented with different classifiers like Random Forest, Gradient Boosted Decision Trees(GBDTs), Logistic

Regression, SVMs and Deep Neural Network. Furthermore, the three main Deep learning architectures used for training are FastText, CNN and Long Short-Term Memory Networks.

[2] This paper talks about identifying the scope of negation in newspaper articles by evaluating different sentiment analysis methods. The data was collected through articles on web portals of two leading news sources, The Hindu and NDTV. This data was then used to predict which political party would win the Lok Sabha Election. Sentiment analysis methods like Naive Bayes, Support Vector Machine and SentiWordNet resulted in the accurate detection of negative news. Few words found in the text which influenced Sentiments in favour of the BJP and correctly predicted them to be the winners of elections.

[3] In this paper, the authors put forward a survey and comparative study of existing techniques for sentiment analysis by using machine learning and lexicon-based approaches which are then paired with evaluation metrics and some cross-lingual methods. Target dataset used here is Twitter data. The final results show SVM and Naive Bayes to give the highest accuracy and lexicon-based methods to be the most effective.

[4] The paper talks about the Internet and the wide amount of data it contains, B2C and B2B systems and big data. Further paper talks about how web scraping can be useful and how it can be used to automate and collect large amounts of data which is quite a tedious task for us to do. The data can be used for ML models and business intelligence and growth.

[5] This paper talks about the sentiment analysis method of short texts in micro blogs. In these kinds of blogs there is a limit of 140 words so that the text content is fragmented, irregular and other characteristics. Using sentiment analysis, these texts can be characterized, summarized and analysed.[6] Uses a similar approach to analyse the sentiments of the reviews of two leading smartphone brands based on the data obtained from social media sites like twitter and Facebook. The system then uses Naive Bayes algorithm to improve the accuracy.

A Study on Sentiment Analysis of Product Reviews [7] demonstrates a comparative analysis of various algorithms used for sentiment analysis. It talks about the use of SVM, Logistic Regression, KNN and other classification and the algorithm that gave the best accuracy. T. K. Shivaprasad et al. in [8] have also used SVM and Naive Bayes for performing sentiment analysis. They have also implemented a Binary approach for classifying the reviews as positive or negative and a Multiclass approach for categorizing the reviews based on the product ratings. M. Kanakaraj et al. in [9] have used ensemble classifiers to predict the sentiments and have worked on increasing the accuracy by adding similar words in the dataset.

The authors in [10] and [11] have used web scraping for implementing Natural Language Processing and Deep Learning algorithms to gain insights from the dataset collected. C. Kaur and A. Sharma in [12] have used twitter scrapers to scrap the tweets on certain social issues to perform analysis on the data obtained. The paper [13] has also proposed a method to scrape data from sites such as google and bing for implementing optimization of image searches with an accuracy up to 90%.

Text sentiment analysis [14] and Opinion mining [15] talk about analysing the data on social media sites to predict the sentiment. They also talk about how pre-processing of the sentiments using NLP plays an important role to get the best results. C. Chauhan and S. Sehgal in [16] also talk about the importance of understanding the needs of the customers based on their reviews for the E-commerce websites to improve their services and product qualities. They have emphasized on the use of Sentiment Analysis techniques to analyse the large number of customer reviews.

III. METHODOLOGY/EXPERIMENTAL

In this project we have generated our own dataset of product reviews by using Web Scraping which was further used to understand the sentiments and feedback of the customers and to classify the reviews as positive or negative. The implementation of the system was carried out in a sequential manner as shown in Fig. 1

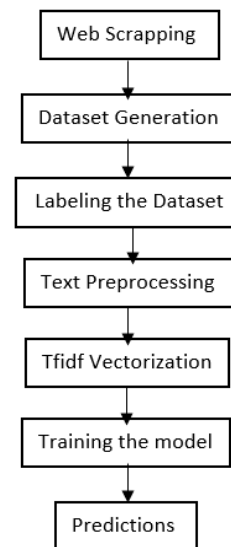


Fig. 1: Implementation workflow

i. Dataset generation using Web scraping:

Web Scraping is a method in which data from the web is collected automatically instead of manually doing the data collection task. The term Web Scraping is also called Screen Scraping, Data Extraction or Web Harvesting which essentially mean the same thing, gathering data from the internet Generally data is gathered from websites by looking at it is HTML (Hyper-Text Markup Language) code and storing the necessary data in a structured format. It is done using something called Web Scraper which can be a raw script or a sophisticated computer program.

For the same purpose, it uses the URL of the website.

It's achieved with HTTP requests, where the program makes a HTTP get request to the server at given URL and gets HTML page as a response from the server, further HTTP programming, HTML parsers like lxml, DOM parsing, etc

are used for analysing the data. This serves the main purpose of Web Scraper which is to gather data from unorganised or unstructured manner and store it in one place in a structured format like CSV, JSON, XML which are most used formats. Also, this is very fast as compared to manual operations.

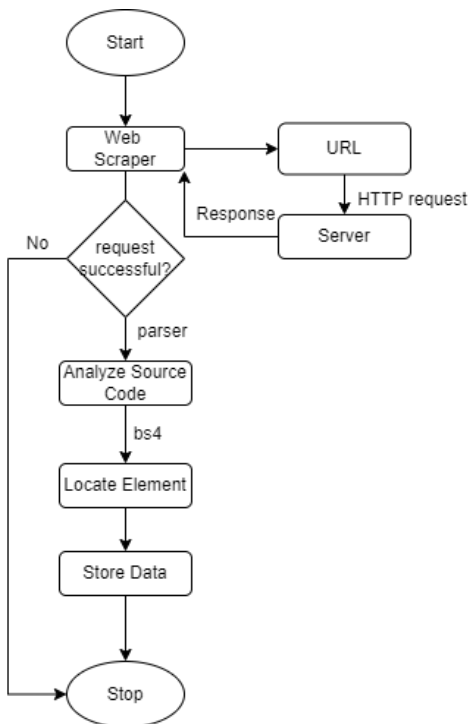


Fig. 2: Flowchart showing general Web Scraping process

ii. Pre-processing on the Dataset

The dataset generated using web scraping had the reviews of the product along with the corresponding rating of that particular product. The first step of pre-processing was to map these ratings into binary labels by setting the threshold rating value to 3. Any product with a rating greater than and equal to 3 is identified as a positive review and any product with a rating less than 3 is considered as a negative review. We assigned the label 1 to the positive reviews and 0 to the negative reviews.

The next step was to perform pre-processing on the reviews. Since the reviews were in the form of text, cleaning the dataset by removing all the unnecessary characters and spaces plays an important role. Any redundant text or numbers in the reviews were removed to reduce the complexity of computation. We used the nltk library for the same for removing the stop words, punctuation marks trailing and leading spaces and other special characters. The pre-processed sentences were then used as the final dataset. These sentences were then used for applying TFIDF Vectorisation.

Term Frequency Inverse Document frequency (TF-IDF) is a type of algorithm which helps in transforming text into a meaningful representation of numbers. This numerical representation can be used to fit a machine learning model. It basically is the amount of originality of a word, which is found out by comparing the number of times the word appears in the document with the number of documents the word appears in.

In our project, we use this method to find the frequency of words in the reviews to find out which high frequency words determine sentiment prediction to be positive or negative. Based on the frequency of words we even eliminate the very low frequency words which have no impact on the sentiment behind the reviews. Such low frequency and low impact words are called Stop words.

By applying TFIDF Vectorisation, every word in the review is assigned to a unique numerical value, like a weight, which tells us how important that word is in order to decide whether the review is positive or negative. A word with a greater weight will be considered more important while determining whether the sentence is positive or negative, as compared to the words having a lower weight. Similar meaning words will hold approximately the same values.

iii. Exploratory Data Analysis

To analyse the data and get a better understanding and insights from the reviews, we implemented some data visualization techniques.

We first grouped the reviews with respect to the ratings associated with them and plotted a histogram to check the number of reviews for each type of rating.

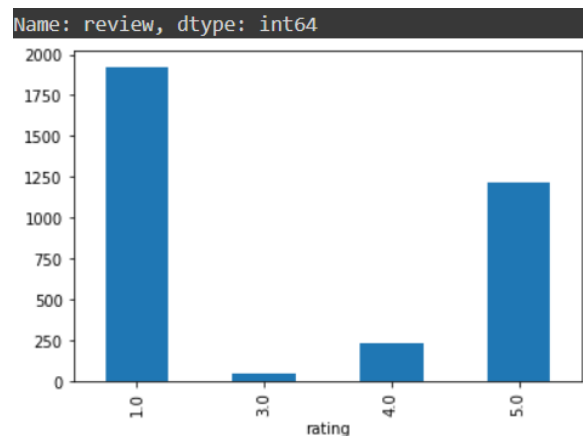


Fig. 2: No. of reviews with respect to ratings

As shown in Fig. 2, it was found that there were 1923 reviews of rating 1, 48 reviews with rating 3, 228 with rating 4 and 1219 with rating 5. From this it was evident that a majority of the reviews had a lower rating. The reviews were further grouped with respect to the Labels assigned to them. The histogram thus plotted, gave the number of positive and negative reviews that were present in our dataset.

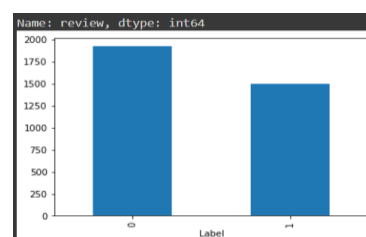


Fig. 3: No. of reviews with respect to labels

From the histogram shown in Fig. 3, we found that the number of positive and negative reviews were almost equal, with a slight bias to the negative reviews, which are 1923. The remaining 1495 reviews are positive.

iv. Training the Machine Learning Model

The processed sentences are then fed into the Supervised Machine Learning models, to predict whether they are positive or negative reviews. The dataset was split into training and testing parts, where 30 percent of the data will be used only for testing the performance of the machine learning model and the model will be actually trained on the remaining 70 percent of the data. Since the aim of the model is to classify the data into binary classes, Logistic regression, Decision Tree, Random Forest and KNN and SVM Algorithms were used for the same.

Logistic regression is used to depict the relationship between a binary value which is the dependent variable, and one or more independent variables which can be nominal, interval, ratio-level and ordinal.

KNN is a supervised learning algorithm which is used to group similar kinds of data. The algorithm stores the grouped similar data points and classifies a new data point based on its similarity with the available group of datapoints.

For the project, we use this method for the binary type of classification that it provides which is needed in this case to predict whether the review is positive= 1 or negative=0. We set the value of K as 5.

For training the model, we created a pipeline using the Pipeline module from sklearn. This module is used to carry out sequential transformation on data. CountVectorizer () was initially passed in the pipeline to convert the word documents into vectors. The next parameter was TfidfTransformer () which assigns a numerical weight to the vector tokens that were created from the original text documents. Following this transformation, the text is ready to be fed into any Machine Learning classification algorithms. At the end, we also printed the confusion matrix for every algorithm to observe the true positive and true negative values.

v Predicting the Sentiment of new reviews

To test the performance of our models, we tried to input some reviews to the model using the predict() function. It was found that the models predicted the positive and negative sentences accurately. For reviews with multiple sentences which were partly positive and partly negative, the model classified the review by considering the weights of the highly positive as well as the highly negative words that were calculated using TFIDF Vectorization.

IV. RESULTS

The pre-processed dataset was trained on 5 Machine learning models, namely Logistic Regression, Decision Tree, SVM, K-Nearest Neighbours and Random Forest. The accuracy of each model was recorded and the performance

of each model was evaluated with the help of a confusion matrix.

Algorithm Name	Accuracy
Logistic Regression	99.71%
Decision Tree	99.71%
Support Vector Classifier	99.71%
Random Forest	99.56%
K-Nearest Neighbours	98.93%

Table 1: Accuracy of models

Table 1 represents the accuracy that was obtained on training the dataset on the 5 supervised machine learning models. The overall accuracy of all the models is high with Logistic Regression, Decision Tree and SVM (Support Vector machine).

The confusion matrix obtained for classifiers Logistic Regression, Decision Tree and Support Vector Machine (shown in Fig. 4) showed similar results. It has correctly classified 366 reviews as negative and 310 reviews as positive. The remaining 8 reviews were falsely classified as negative reviews.

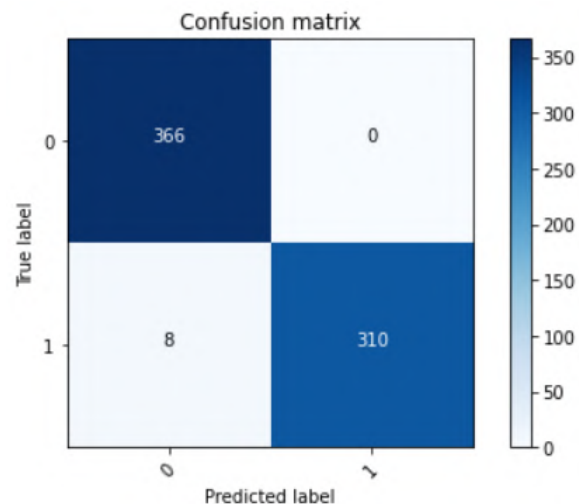


Fig. 4: Confusion matrix for Logistic Regression

V. CONCLUSION

This project envisions to provide a tool for text classification which performs analysis on an incoming review and tells whether the underlying sentiment is Positive or Negative with 99.56% accuracy. The project can help users like business and product owners who want to know the public or consumer opinions and their emotions about the services and products. The project also will help potential customers who want to know the opinions and emotions of the existing users before using a particular service or purchasing a product.

The project can be further improved by adding more reviews for a variety of products that are available and popular on the E-commerce website. The insights obtained from the system can also be used for implementing search

engine optimization. This model can be improved and deployed on the website for the customers to analyse the reviews efficiently.

REFERENCES

- [1] P.Barjatya, S.Gupta, M.Gupta, V.Verma, "Deep Learning for Hate Speech Detection in Tweets", 26th International Conference on World Wide Web Companion, April 2017
- [2] S. Padmaja, "Evaluating Sentiment Analysis Methods and Identifying Scope of Negation in Newspaper Articles", International Journal of Advanced Research in Artificial Intelligence, 2014
- [3] V.Kharde, S.S.Sonawane, "Sentiment Analysis of Twitter Data: A Survey of Techniques", International Journal of Computer Applications, April 2016
- [4] J. Li and L. Qiu, "A Sentiment Analysis Method of Short Texts in Microblog," 2017 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC), 2017, pp. 776-779, doi: 10.1109/CSE-EUC.2017.153.
- [5] P.Gupta, R.Tiwari,N.Robert, "Sentiment analysis and text summarization of online reviews: A survey", 2016 International Conference on Communication and Signal Processing (ICCSPP)
- [6]N. Srivats Athindran, S. Manikandaraj and R. Kamaleshwar, "Comparative Analysis of Customer Sentiments on Competing Brands using Hybrid Model Approach," 2018 3rd International Conference on Inventive Computation Technologies (ICICT), 2018, pp. 348-353, doi: 10.1109/ICICT43934.2018.9034283.
- [7] A. S. Parihar and Bhagyanidhi, "A Study on Sentiment Analysis of Product Reviews," 2018 International Conference on Soft-computing and Network Security (ICSNS), 2018, pp. 1-5, doi: 10.1109/ICSNS.2018.8573681.
- [8] T. K. Shivaprasad and J. Shetty, "Sentiment analysis of product reviews: A review," 2017 International Conference on Inventive Communication and Computational Technologies (ICICCT), 2017, pp. 298-301, doi: 10.1109/ICICCT.2017.7975207.
- [9] M. Kanakaraj and R. M. R. Guddeti, "NLP based sentiment analysis on Twitter data using ensemble classifiers," 2015 3rd International Conference on Signal Processing, Communication and Networking (ICSCN), 2015, pp. 1-5, doi: 10.1109/ICSCN.2015.7219856.
- [10] S. Lunn, J. Zhu and M. Ross, "Utilizing Web Scraping and Natural Language Processing to Better Inform Pedagogical Practice," 2020 IEEE Frontiers in Education Conference (FIE), 2020, pp. 1-9, doi: 10.1109/FIE44824.2020.9274270.
- [11] F. Ertam, "Deep learning-based text classification with Web Scraping methods," 2018 International Conference on Artificial Intelligence and Data Processing (IDAP), 2018, pp. 1-4, doi: 10.1109/IDAP.2018.8620790.
- [12] C. Kaur and A. Sharma, "Social Issues Sentiment Analysis using Python," 2020 5th International Conference on Computing, Communication and Security (ICCCS), 2020, pp. 1-6, doi: 10.1109/ICCCS49678.2020.9277251.
- [13]Ridwang, A. A. Ilham, I. Nurtanio and Syafaruddin, "Image search optimization with web scraping, text processing and cosine similarity algorithms," 2020 IEEE International Conference on Communication, Networks and Satellite (Commnetsat), 2020, pp. 346-350, doi: 10.1109/Commnetsat50391.2020.9328982.
- [14] R. Hu, L. Rui, P. Zeng, L. Chen and X. Fan, "Text Sentiment Analysis: A Review," 2018 IEEE 4th International Conference on Computer and Communications (ICCC), 2018, pp. 2283-2288, doi: 10.1109/CompComm.2018.8780909.
- [15] R. K. Bakshi, N. Kaur, R. Kaur and G. Kaur, "Opinion mining and sentiment analysis," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 2016, pp. 452-455.
- [16] C. Chauhan and S. Sehgal, "Sentiment analysis on product reviews," 2017 International Conference on Computing, Communication and Automation (ICCCA), 2017, pp. 26-31, doi: 10.1109/CCAA.2017.8229825.

Development Approach: Electricity Consumption and Prediction using Machine Learning Techniques

Palak Sharma
DIT University, Dehradun
palakbhayal@gmail.com

Brajlata Chauhan
DIT University, Dehradun
braj.lata@rediffmail.com

Amrindra Pal
DIT University, Dehradun
amrindra.pal@gmail.com

Abstract—Power is a fundamental component which determines country's economic progress and a person's quality of life. There are many challenges in power sector such as Insufficient Electricity Generation, Poor Management, Lack of Investments, Poor Infrastructure, Raw Material Shortage. Therefore, prediction of power consumption or demand is important in Power Industry, as it helps in power system planning and operation. Electrical Utilities makes predictions for the short, medium, and long periods. The electricity consumption depends both on meteorological and socio-economic factors where conventional methods seems to be insufficient in terms of prediction and accuracy thus, we use Machine Learning models for better accuracy and efficiency. This paper describes various Machine Learning Algorithm for electricity consumption prediction such as ANN, SVM, MLP, LR, SVM, ANFIS, KNN, XG BOOST, ELM, DNN etc. These can be used as individual model or hybrid models for higher accuracy and efficiency. Mainly, hybrid models and ensemble models are used to overcome the challenges.

Keywords— *electricity consumption, electricity prediction, ml, ANN, SVM, LR, ANFIS, KNN, XG BOOST, ELM, DNN*

I. INTRODUCTION

Global electricity consumption has risen at a quicker rate than global energy use. Annual energy consumption per capita is a fundamental indication of a country's electric power development progress. In general, power consumption rises quicker as the industrialization process progresses quickly and falls rapidly when the process is completed or near completion. Power consumption has risen considerably in recent years due to the massive growth of the world's population. It has a significant impact on socio-economic development, because without electricity, no habitation can exist [1]. The four categories of electricity consumption prediction are:

1. Fairly long strategy (>1 year).
2. Moderate (1 week – 1 year).
3. Short period (1 hour - 1 week).
4. Ultrashort-term (< 1 hour).

To define acceptable techniques to reduce the electricity usage during periods of higher power prices and build schemes for the construction & repairing electrical components, as well as for energy market planning and retrofitting, medium- to long-term predictions are utilized. Short-period predictions are employed in everyday planning, each week planning, and fairly long energy storing systems. For security surveillance and emergency management, ultrashort-term energy prediction is used. Furthermore, because non-conventional energy and power usage are both uncertain, ultra-small period building power usage forecasts are crucial in decentralized power management planning and requirement real-time response. [2]. The primary concern of this paper is to go through the most common machine learning models for predicting

power usage. ML Models utilize past dataset to generate future results.

- A variety of prediction models are utilized for determining total power use.

Arithmetical Models & Artificial Intelligence-based methods are the two types of methodologies used. Because of the proliferation of Machine Learning algorithms, the latter has grown in popularity and efficacy. The latter were largely data-driven approaches to decision-making that were adaptive, autonomous, and intelligent [3]. Prediction depends on DSM, and environmental variables like sunshine availability, temp, rainfall patterns, moisture, and cloudage.

The following is the outline of this paper:

The literature reviews are represented in Section II, the kinds of machine learning (ML) are represented in Section III, and the different Machine Learning Models (ML) for Prediction are described in Point IV. Section V states that there are a variety of machine learning algorithms with varying prediction accuracy.

II. LITERATURE REVIEW

In [2], Che Liu et al. suggested a combination of two approaches for ultra-short period home power consumption estimation, combining the HW approach and the ELM network. When compared to the Holt-Winters, Extreme Learning Machine, and long- short-term memory models, the suggested approach performed better with 50 days data, and the RMS error value decreased to 87.98 percent, 64.89 percent, and 53.39 percent respectively. In [3], Ghassen Ben Brahmin et al. presented a Machine Learning (ML)-based technique for estimating energy consumption based on weather data and 350 days of equipment data. They got the results using Random Forest with a correlation coefficient of 75.7%.

In [4], Shalika Walker et al. analyzed and tested various Machine Learning (ML) algorithms i.e., as BT, RF, Support vector machine, Artificial Neural Network for prediction of electricity demand of particular or group of buildings. The hourly data was collected from different 47 buildings of two years. BT, RF and Artificial Neural Network gave better accuracy. In [9], Tobias Haring et al. used three models of Machine Learning such as LR, LSTM & NN for prediction of load of smart cities depending upon Estonian power usage dataset where hourly consumption data of 2019 was used for prediction. Linear Regression gave better accuracy of all.

In [10], Alfonso Gonzalez-Briones et al. reviewed some ML approaches for prediction of energy usage forecast. & daily consumption data was chosen for prediction in which Linear Regression and Support Vector Regression have better accuracy of 85.7%. In [11], Risul Islam Rasel et al. proposed two machine learning methods i.e., SVR and ANN with cross validation strategy were used to solve data

dimensionality problems for predicting power usage of a low-energy house, and also F-test, Correlation analysis, and Principal component analysis. The BP-ANN method performed better than SVR. The proposed approach has a 98 percent accuracy rate in predicting electric energy usage.

In [12], Roya Ahmadihangar et al. focused on using machine learning-based regression models to create load patterns in order to estimate residential customers' potential flexibility and enhance both technical and economic smart grid operations. The proposed method may be employed in a variety of control approaches, in online and real-time methods. In [13], Ahmed Ghareeb et al. predicted load demand of Kirkuk city using three machine learning algorithms i.e GLM, ANN and RF then compared individual method capabilities with Ensemble Models. The RF model outperforms the other models.

In [14], HAN-YUN CHEN et al. The author focused on two methods i.e., ANFIS and GRA for power usage forecasting of buildings. Their research was based on natural elements such as temperature, sunlight period, radiation from the sun, working day and school day data, as well as electrical use in a school library. ANFIS performs better as it is a faster method for prediction. In [15], Meng Shen et al. focused to create better SVR approach to forecast residential electricity usage, uses Akaike Information Criterion for variable selection approach and A Monte Carlo methodology is utilized for examining the relation in character, finest interference choices, & most energy reduction feasible. The results showed a monthly power usage drop of 12.1% on mean.

In [16], Seungwon Jung et al. The researcher uses Multilayer Perceptrons to provide a missing-value estimation technique for electricity usage data. They compared these methods with machine learning algorithms to get the greater efficiency. In [17], Prince Waqas Khan et al. for load consumption prediction, author employed EML approach with XGBoost, SVR, and KNN regressor method, as well as the GA for load consumption forecasting. For the three-month test data, they obtained a MAPE of 3.35 percent, using this hybrid model that has been proposed for prediction of energy consumption.

In [18], Fath U Min Ullah et al. suggested a combination of two approaches which includes neural network with a multi-layer bi-directional long-short term memory for prediction of power consumption and attain low RMS Error for particular house using 10-fold cross validation & hold-out approach.

In [19], Anh-Duc Pham et al. The author proposed Random Forests (RF) for predicting brief power usage in several buildings at an hourly frequency. On comparison to the RT model, the RF model has an accuracy of 49.21%, and 46.93% in mean absolute percentage error (MAPE). In [20], Ritika Tilwalia et al. proposed Grey Wolf Algorithm, Support Vector Machines, PSO and Cuckoo algorithm and smart house dataset was obtained from Kaggle, and the characteristics were chosen using the pearson coefficient. The accuracy is predicted by using SVM and there is a slight difference of about 1% in accuracy of PSO and Cuckoo algorithm.

Table 1. Summary of sources, models and data collected for prediction

Author	Year	Techniques	Source	Data Collected
Alfonso Gonzalez-Briones	2019	Linear Regression and Support Vector Regression	Electricity	Shoe store
Zhe Wang	2020	Shallow learning, Deep Learning, Heuristic methods, XGBoost and LSTM.	Thermal Load	Building
Tobias Häring	2020	Linear regression, LSTM and NN.	Electricity	Smart cities
Ritika Tilwalia	2021	Grey Wolf Algorithm, SVM, PSO and Cuckoo algorithm	Electricity	Smart homes
Shubing Shan	2019	Ensemble Model	Electricity	Five-star hotel building, Office building.
Shalika Walker	2019	Boosted-tree, RF, SVM & ANN	Electricity	Commercial Buildings
Seungwon Jung	2020	Deep Learning; Multilayer Perceptron; Ensemble Learning	Electricity	Smart meters
Roya Ahmadihangar	2019	Regression Model	Electricity	Estonian household
Risul Islam Rasel	2019	SVR and BP-ANN	Electricity	Low Energy House
Prince Waqas Khan	2020	XGBoost, SVR, and KNN Regressor Algorithms, Genetic Algorithm.	Electricity	Jeju island, Korea
Mohammad Munem	2020	Multivariate Bayesian optimization, Long Short-Term Memory Neural Network, CNN, ANN and SVM.	Electricity	Residential house
Meng Shen	2020	GA-RBF-SVR model	Electricity	House
Minglei Shao	2020	Support Vector Machine (SVM)	Electricity	Hotel building
X.J. Luo	2020	Genetic Algorithm,	Electricity	Campus

		Deep Feed Forward Neural Network	building
Ran Wang	2020	Stacking model, RF, Gradient Boosted Decision Tree, Extreme Gradient Boosting, SVM, and KNN.	Electricity Building

III. TYPES OF MACHINE LEARNING

Machine learning (ML) is a subset of artificial intelligence (AI), system's ability to imitate human behavior. In ML, the data is used to explain what happened, predictive in which the data is used to predict what will happen; or prescriptive, in which the data is used to offer recommendations on what action to take.

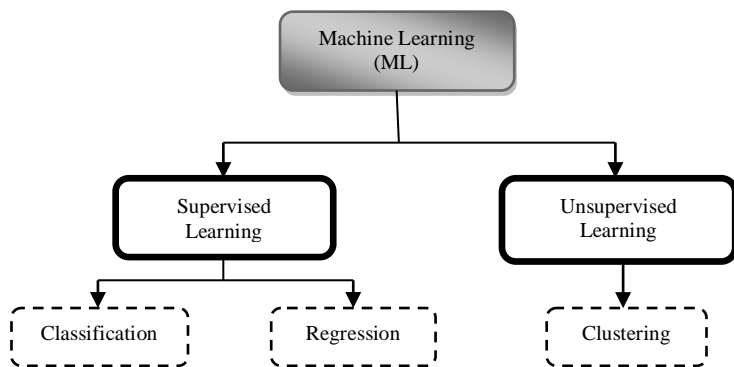


Fig.1. Machine Learning (ML)Types

- **SUPERVISED LEARNING**

These models are trained with labeled data sets, which enables the models to learn and continue to get better. It is the most common type used. The process of giving input and output data to the machine learning model is known as supervised learning. With the help of supervised learning, the model can forecast the output on the basis of historical data. We can use supervised learning model to resolve a variety of issues, such as fraud detection, spam filtering etc.

- **UNSUPERVISED LEARNING**

In this, Models are trained on unlabeled data sets before being left to function on its own. It cannot be applied directly to regression or classification as unsupervised learning has only input data. **Unsupervised Learning Algorithms** let's users to do more complicated jobs than supervised learning. Unsupervised learning, on the other hand, is more unpredictable than other learning approaches. It includes clustering, anomaly detection, neural networks, etc.

IV. MACHINE LEARNING MODELS FOR PREDICTION

- **SUPPORT VECTOR MACHINE(SVM)**

SVM is used to handle non-linear problems, as well as to predict and classify data. It uses a maximum margin to distinguish the classes. When SVM is used to forecast a time series or a set of values, it is referred to as Support Vector Regression. SVR and SVM are based on the same ideas [4]. SVMs are used for structural risk minimization based on statistical learning theory. It outperforms alternative approaches in pattern recognition, classification, and regression analysis [5].

- **K-NEAREST NEIGHBORS**

The K-Nearest Neighbors algorithm is type of supervised ML method for solving both classification & regression. It is simple to learn & implement. It has a disadvantage that it slows down as the size of the data increases [7]. The algorithm trains data stored & similarity between test data & training set records is calculated for forecasting the test data. Then data sets is given, and the machine is trained therefore we can figure out why our energy consumption is increasing [6].

- **XG BOOST**

XG Boost is ELM, a predictor is created from a collection of tiny predictors. Combining numerous predictors improves models accuracy & capacity. Rather than using strong predictors, XG Boost uses weak predictors to solve this problem [8].

- **LINEAR REGRESSION(LR)**

Linear regression is a type of supervised ML algorithm. It is mostly utilized in forecasting and determining the relations between variables. Different regression models differ in terms of relationship they examine as dependent & independent variables and the number of independent variables they utilize. LR determines relation between energy use and the return or other variables. It forecast energy demand based on one or more independent variables [7].

- **ARTIFICIAL NEURAL NETWORK**

ANN are frameworks that allow various machine learning algorithms to interpret complex data inputs. ANNs can be used for forecasting, regression, and curve fitting, among other things. A neuron that uses a transfer function for output formulation is referred to as an artificial neural network fundamental unit [5]. Computer systems are programmed to act as if they were interconnected brain cells to build ANNs. ANN can be used to categorise data, forecast results &

collection of data in a variety of ways. The networks may categorise data in preset class as they process and learn from the input. It can be trained to predict outputs from a given input and to identify and classify data using a specific feature. ANN allows computers to grasp the environment around them in a human-like manner [7].

- **RANDOM FOREST(RF)**

A random forest is a machine learning technique used for solving classification and regression problems. It makes use of ensemble learning, which solves complicated problems by combining several classifiers. It provides a higher level of accuracy. By assuming missing data, the Random Forest Algorithm operates fast and effectively in larger databases, giving high accurate predictions. Random Forest, or RF, is an effective way for training models in development process. The main disadvantage is that it takes longer to develop. Only different feature such as types, numeric, binary, and categorical should be considered. Random Forest is recognized as a quick, easy, and adaptable tool [7].

- **EXTREME LEARNING MACHINE(ELM)**

Feedforward neural networks are extreme learning machines. For classification, regression, clustering, sparse approximation, compression, and feature learning, it has a single layer or many layers of hidden neurons. These hidden nodes can be given any name they want and never updated, or they can be anchored in their predecessor and never modified. It has a minimal computing cost during the training procedure. ELM was created with the intention of being used in time series prediction. ELM is a fast-computational model for electricity price forecasting. ELM can also be used to predict wind power density [7].

- **DEEP NEURAL NETWORK(DNN)**

The Deep Neural Network is more complicated and creative than a traditional neural network. Deep Neural Network algorithms are able to recognize sound and voice instructions, making predictions, thinking creatively, and analyzing data. They function similarly to the human brain. Networks with at least one hidden layer between the input and output layers are known as deep neural networks., each layer accomplishes particular types of sorting and ordering. It deals with unlabeled or unstructured input.

- **ANFIS**

An adaptive neuro-fuzzy inference system (ANFIS) is a sort of artificial neural network that uses the Takagi–Sugeno fuzzy inference system as its

cornerstone. It allows you to combine the advantages of neural networks and fuzzy logic concepts in a single model. ANFIS is regarded as a universal estimator.

- **K-MEANS CLUSTERING**

Clustering is an unsupervised learning strategy that uses an unlabeled test dataset. To establish a hierarchy, hierarchical clustering uses two types of clustering techniques. They are both aggregative and divisive in nature. In a bottom-up method, agglomerative clustering is combined to form a large cluster. In a top-down method, Divisive Clustering divides large cluster into smaller clusters. Every items in collection is classified as a cluster. Dataset is divided into K-small clusters using K-Means Clustering.

V. CONCLUSION

Electricity consumption is increasing every day as a result of industrialization, rising population, and a variety of other variables, making it necessary to forecast consumption. Machine learning is gaining importance in today's world for prediction during the last decade and there are various ML models such as ANN, SVM, MLP, LR, SVM, ANFIS, KNN, XG BOOST, ELM, DNN etc. These can be used as individual model or hybrid models for higher accuracy and efficiency. In most problems, hybrid models are used.

REFERENCES

- [1] M. Munem, T. M. Rubaith Bashar, M. H. Roni, M. Shahriar, T. B. Shawkat, and H. Rahaman, "Electric power load forecasting based on multivariate LSTM neural network using bayesian optimization," *2020 IEEE Electr. Power Energy Conf. EPEC 2020*, vol. 3, 2020, doi: 10.1109/EPEC48502.2020.9320123.
- [2] C. Liu, B. Sun, C. Zhang, and F. Li, "A hybrid prediction model for residential electricity consumption using holt-winters and extreme learning machine," *Appl. Energy*, vol. 275, no. February, p. 115383, 2020, doi: 10.1016/j.apenergy.2020.115383.
- [3] G. Ben Brahim, "Weather Conditions Impact on Electricity Consumption in Smart Homes: Machine Learning Based Prediction Model," *2021 8th Int. Conf. Electr. Electron. Eng. ICEEE 2021*, no. 1, pp. 93–98, 2021, doi: 10.1109/ICEEE52452.2021.9415917.
- [4] S. Walker, W. Khan, K. Katic, W. Maassen, and W. Zeiler, "Accuracy of different machine learning algorithms and added-value of predicting aggregated-level energy performance of commercial buildings," *Energy Build.*, vol. 209, p. 109705, 2020, doi: 10.1016/j.enbuild.2019.109705.
- [5] A. Mosavi, M. Salimi, S. F. Ardabili, T. Rabczuk, S. Shamshirband, and A. R. Varkonyi-Koczy, "State of the art of machine learning models in energy systems, a systematic review," *Energies*, vol. 12, no. 7, 2019, doi: 10.3390/en12071301.
- [6] A. Balachandran, Ramalakshmi, Venkatesan, M. Lakshmi, K. Jahnavi, and V. Jothi, "Energy consumption analysis and load management for smart home," *Proc. Int. Conf. Trends Electron. Informatics, ICOEI 2019*, vol. 1, no. Icoei, pp. 46–49, 2019, doi: 10.1109/ICOEI.2019.8862734.

- [7] T. Ranjan Jena, S. Sucharita Barik Assistant Professor, and S. Kumari Nayak Assistant Professor, "Electricity Consumption & Prediction using Machine Learning Models," *Acta Tech. Corviniensis - Bull. Eng.*, vol. 14, no. 1, pp. 61–68, 2021, [Online]. Available: <https://lavasallibrary.remotexs.in/scholarly-journals/electricity-consumption-amp-prediction-using/docview/2513326744/se-2?accountid=38885>.
- [8] Z. Wang, T. Hong, and M. A. Piette, "Building thermal load prediction through shallow machine learning and deep learning," *Appl. Energy*, vol. 263, no. November 2019, p. 114683, 2020, doi: 10.1016/j.apenergy.2020.114683.
- [9] T. Haring, R. Ahmadiyahangar, A. Rosin, T. Korotko, and H. Biechl, "Accuracy Analysis of Selected Time Series and Machine Learning Methods for Smart Cities based on Estonian Electricity Consumption Forecast," *Proc. - 2020 IEEE 14th Int. Conf. Compat. Power Electron. Power Eng. CPE-POWERENG 2020*, pp. 425–428, 2020, doi: 10.1109/CPE-POWERENG48600.2020.9161690.
- [10] A. González-Briones, G. Hernandez, J. M. Corchado, S. Omatu, and M. S. Mohamad, "Machine Learning Models for Electricity Consumption Forecasting: A Review," *2nd Int. Conf. Comput. Appl. Inf. Secur. ICCAIS 2019*, 2019, doi: 10.1109/CAIS.2019.8769508.
- [11] R. I. Rasel, N. Sultana, S. Akther, and A. Haroon, "Predicting Electric Energy Use of a Low Energy House: A Machine Learning Approach," *2nd Int. Conf. Electr. Comput. Commun. Eng. ECCE 2019*, pp. 1–6, 2019, doi: 10.1109/ECACE.2019.8679479.
- [12] R. Ahmadiyahangar, T. Häring, A. Rosin, T. Korotko, and J. Martins, "Residential Load Forecasting for Flexibility Prediction Using Machine Learning-Based Regression Model," *Proc. - 2019 IEEE Int. Conf. Environ. Electr. Eng. 2019 IEEE Ind. Commer. Power Syst. Eur. EEEIC/I CPS Eur. 2019*, 2019, doi: 10.1109/EEEIC.2019.8783634.
- [13] A. Ghareeb, H. Al-Bayaty, Q. Haseeb, and M. Zeinalabideen, "Ensemble learning models for short-term electricity demand forecasting," *2020 Int. Conf. Data Anal. Bus. Ind. W. Towar. a Sustain. Econ. ICDABI 2020*, 2020, doi: 10.1109/ICDABI51230.2020.9325623.
- [14] H. Y. Chen, C. H. Le, and B. M. Huang, "Electricity Consumption Forecasting of Buildings Using Hierarchical ANFIS and GRA," *Proc. - Int. Conf. Mach. Learn. Cybern.*, vol. 2019-July, pp. 1–7, 2019, doi: 10.1109/ICMLC48188.2019.8949177.
- [15] M. Shen, Y. Lu, K. H. Wei, and Q. Cui, "Prediction of household electricity consumption and effectiveness of concerted intervention strategies based on occupant behaviour and personality traits," *Renew. Sustain. Energy Rev.*, vol. 127, no. March, p. 109839, 2020, doi: 10.1016/j.rser.2020.109839.
- [16] S. Jung, J. Moon, S. Park, S. Rho, S. W. Baik, and E. Hwang, "Bagging ensemble of multilayer perceptrons for missing electricity consumption data imputation," *Sensors (Switzerland)*, vol. 20, no. 6, pp. 1–16, 2020, doi: 10.3390/s20061772.
- [17] P. W. Khan and Y. C. Byun, "Genetic algorithm based optimized feature engineering and hybrid machine learning for effective energy consumption prediction," *IEEE Access*, vol. 8, pp. 196274–196286, 2020, doi: 10.1109/ACCESS.2020.3034101.
- [18] F. U. M. Ullah, A. Ullah, I. U. Haq, S. Rho, and S. W. Baik, "Short-Term Prediction of Residential Power Energy Consumption via CNN and Multi-Layer Bi-Directional LSTM Networks," *IEEE Access*, vol. 8, pp. 123369–123380, 2020, doi: 10.1109/ACCESS.2019.2963045.
- [19] A. D. Pham, N. T. Ngo, T. T. Ha Truong, N. T. Huynh, and N. S. Truong, "Predicting energy consumption in multiple buildings using machine learning for improving energy efficiency and sustainability," *J. Clean. Prod.*, vol. 260, p. 121082, 2020, doi: 10.1016/j.jclepro.2020.121082.
- [20] R. Tilwalia, A. Jain, and D. Gupta, "Optimization of Electricity Consumption using Grey Wolf Algorithm," *2020 IEEE 5th Int. Conf. Comput. Commun. Autom. ICCCA 2020*, pp. 401–407, 2020, doi: 10.1109/ICCCA49541.2020.9250899.
- [21] S. Shan, B. Cao, and Z. Wu, "23 Forecasting the Short-Term Electricity," no. 1, pp. 1–15.
- [22] 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, no. September, p. 110287, 2020, doi: 10.1016/j.rser.2020.110287.
- [23] M. S. Ibrahim, W. Dong, and Q. Yang, "Machine learning driven smart electric power systems: Current trends and new perspectives," *Appl. Energy*, vol. 272, no. May, p. 115237, 2020, doi: 10.1016/j.apenergy.2020.115237.
- [24] R. Wang, S. Lu, and W. Feng, "A novel improved model for building energy consumption prediction based on model integration," *Appl. Energy*, vol. 262, no. January, p. 114561, 2020, doi: 10.1016/j.apenergy.2020.114561.
- [25] X. J. Luo *et al.*, "Genetic algorithm-determined deep feedforward neural network architecture for predicting electricity consumption in real buildings," *Energy AI*, vol. 2, p. 100015, 2020, doi: 10.1016/j.egyai.2020.100015.
- [26] M. Shao, X. Wang, Z. Bu, X. Chen, and Y. Wang, "Prediction of energy consumption in hotel buildings via support vector machines," *Sustain. Cities Soc.*, vol. 57, no. December 2019, p. 102128, 2020, doi: 10.1016/j.scs.2020.102128.
- [27] V. Uher, R. Burget, M. K. Dutta, and P. Mlynek, "Forecasting electricity consumption in Czech Republic," *2015 38th Int. Conf. Telecommun. Signal Process. TSP 2015*, pp. 262–265, 2015, doi: 10.1109/TSP.2015.7296264.
- [28] I. Ullah, R. Ahmad, and D. H. Kim, "A prediction mechanism of energy consumption in residential buildings using hidden markov model," *Energies*, vol. 11, no. 2, pp. 1–20, 2018, doi: 10.3390/en11020358.
- [29] Y. Shen, R. Wei, and L. Xu, "Energy consumption prediction of a greenhouse and optimization of daily average temperature," *Energies*, vol. 11, no. 1, 2018, doi: 10.3390/en11010065.
- [30] M. K. M. Shapi, N. A. Ramli, and L. J. Awal, "Energy consumption prediction by using machine learning for smart building: Case study in Malaysia," *Dev. Built Environ.*, vol. 5, no. December 2020, p. 100037, 2021, doi: 10.1016/j.dibe.2020.100037.

DEVELOPMENTS IN MACHINE LEARNING MODELS FOR THE PREDICTION OF RENEWABLE ENERGY

Rashida Tabassum
DIT University, Dehradun
rashidatabassum1@gmail.com

Brajlata Chauhan
DIT University, Dehradun
braj.lata@rediffmail.com

Amrindra Pal
DIT University, Dehradun
amrindra.pal@gmail.com

Abstract— Renewable energy is increasingly being used to mitigate the consequences of environmental changes and overheating. All Renewable Energy Sources (RES) have one thing in common: they all rely on the surrounding, which creates considerable management and planning issues. Various prediction strategies have been developed in an attempt to increase renewable energy (RE) prediction ability. The ML models for -example ANN, Random Forest, DT, Support Vector Machine, ANFIS, etc. are used for the prediction of Renewable energy generation This paper aims to give an evaluation of ML techniques used to forecast RE. Such machine learning models give high accuracy and precision. This paper also comes to a conclusion about the efficacy of ML techniques for the forecasting of Renewable energy.

Keywords— *RE; Forecasting; ML models; ANN; DT; SVM*

I. INTRODUCTION

Renewable energy is seen as the most encouraging replacement of fossil fuel since it is clean, green, and regenerated over a large geographic region; yet, it also introduces unplanned uncertainty, endangering energy reliability and stability, particularly when it comes to large scale renewable energy integration [1]. In the last few decades, renewable generating integration has increased over the world to satisfy ever-increasing power demand and emissions targets. Energy prediction utilizing RESs is gradually gaining popularity in several sectors due to the merits it gives to the modernistic surrounding. Using RESs to predict energy consumption not only aids in the reduction of carbon discharge but also aids in the conservation of energy for future use. Several approaches for predicting RE have been grown over time, all of which have emphasized on the efficiency of estimation techniques with no or small concern for the environmental conditions.

Renewable energy can efficiently cut fossil energy use, minimize pollution, and promote the healthy growth of the social economy [2]. In the latest decades, the electricity market has moved its attention to renewable energy sources to minimize its greenhouse discharge during power production. Solar and wind have regularly been used as a combination due to their RE variety and reliability. Hybrid power plants have been studied due to their versatility, efficacy, and accuracy. Due to the variable character of power production from solar and wind, the operators must control and maintain the power plant properly. For short and long-term power transmission scheduling, the electricity production of photovoltaic- wind must be forecasted.[3]

To anticipate renewable energy generation, numerous ML approaches have been developed [4]. Machine learning (ML) approaches are now widely used in a variety of renewable energy-related applications, including the growth of energy and unification, utilization, and forecasting [5].

These approaches right away understand the knowledge from the content rather than the technique depending on preset assumptions. It's a type of AI that uses statistical approaches to teach machines how to study and develop over time. It draws out design from datasets, i.e. machines can modify the developments in the world as well as establish the guidelines for excellent responses. There are various techniques used in machine learning such as MLFNN, WNN, SVM, BP, MLR, RF, ANN, DT. Energy demand growth, population density growth, demand response as parameters, and natural characteristics such as sunshine duration, temperature, precipitation, humidity, and cloud cover are used to make the prediction. The remaining part of this study is presented as follows: A literature review is conducted in Part 2. Classification of Machine Learning Model is given in Part 3. At last, Part 4 brings the paper to an end.

II. LITERATURE REVIEW

In [1] Huaizhi Wang et al. give a broad analysis of prediction of RE depending on deep neural network techniques and these techniques are divided among 5 categories: Deep Convolution Neural Network, Deep Recurrent Neural Network, Deep Belief Network, Stacked Auto Encoder, and additional techniques for investigating its efficacy and applications. In [2], Ling-Ling Li et al. proposed a combination of improved multi-universe optimizer and support vector machine for solar power forecasting, which has shown to be accurate under a variety of weather conditions. The HIMVO-SVM prediction model has been shown to have improved prediction accuracy and stability.

In [3], Zakria Qadir et al. proposed feature selection techniques (ANN) for the prediction of energy and power and having MSE, MAE, R^2 and computation of 0.00000104, 0.00083, 99.6%, and 0.02 s respectively by using the LR model, and this show that the proposed scheme has an ability to improve accuracy and the predict the renewable energy sources. In [4], Rahul et al. proposed a DT regressor model in order to predict solar power generation. The outcomes of this study were good. In [5], Md Mijanur Rahman et al. used Artificial Neural Network approaches such as MLP, RNN, CNN, and LSTM for the accurate prediction of Renewable energy, and MLP produces better results than linear techniques in the works examined. In [8], Mohammad Hassan Fathollahzadeh et al. used the bottom-up method for developing an outline of the power load of the whole community that is happening every hour and incorporates it with technical analysis of hybrid renewable energy, as well as site selection by using Energyplus, SAM, and Wind toolkit.

In [9], Xiaomin Chang et al. proposed extreme grading boosting and a self-organizing map to forecast Photovoltaic energy. In [10], Chao Zhang et al. predicted the wind power by using the LLE-IELM model based on

CEEMD-LCZ, and by comparing with ELM, LLE-IELM gives better accuracy for the prediction of wind power.

In [11], Lin Qiao et al. interpret wind speed data and forecast power generation by using NN technique and backpropagation, and in this work, accuracy is improved by 3.1% when compared with SVM. In [12], Prince Waqas Khan et al. proposed a hybrid technique to forecast RE and non-RESs. MLP and Category boosting are all utilized in the proposed model. To forecast energy consumption, aggregated the information of demand from RE and non-RESs.

In [13], Min Xia et al. presented an upgraded GRU-RNN to predict Renewable Energy generation and electrical load. In this study, two experiments were used to validate the proposed method: wind power generation prediction using numerous meteorological characteristics and electrical load prediction using historical energy usage data and outcomes of the proposed model gives better prediction of Renewable Energy. In [14], Jamal Faraji et al. presented an FF-ANN depending on the weather forecast for the PV and Wind power output. To test the effectiveness of the suggested method, the output of forecasting and real power of Renewable energy is used, and to solve the specified increment issue happening in a day, a Mixed integer linear programming technique is used.

In [15], Seul-Gi Kim et al. proposed a two-step technique to predict the PV generation. The empirical data reveal that this strategy outperforms a baseline approach by a significant margin. Independent of sorts of ML approach used in test data, the RF regression technique gives the best value of R^2 . In [16], Manish Kumar Thukral et al. proposed an ML-FNN technique for photovoltaic energy forecasting. The proposed neural network gives better accuracy when compared with multiple linear regression. In [17], Steven de Jongh et al. proposed Spatio-temporal methods such as ARMA and Long short-term memory NN for the prediction of PV generation by gathering data of weather.

Table 1. Summary Of Models and Sources Used In prediction

Author	Year	Techniques	Source
Yan Cao	2021	Deep Recurrent type-three Fuzzy system	RE
Min Xia	2021	GRU-RNN	Wind, Electricity
Seul-Gi Kim	2019	Random Forest Regression Algorithm	Solar Power
Rabin K. Jana	2020	MODWT, LSTM	Residential, Commercial, industrial, Transport sectors
Zakria Qadir	2021	Artificial Neural Network (ANN), Recursive Feature Technique, Linear regression	Solar, Wind
Md Mijanur Rahman	2021	Artificial Neural Networks (ANNs), Multi-layer Perception, CNN, LSTM, Backpropagation	Solar, Wind, and Hydropower

Algorithm			
Nivethitha Somu	2020	Clustering, CNN, LSTM	Electricity
Prince Waqas Khan	2020	Category Boosting, MLP, SVR, soft-Computing	Wind Energy, Solar Energy, Non-Renewable Energy
Steven de Jongh	2020	Spatio- Temporal Methods, LSTM, Auto Regressive-Moving Average (ARMA), Quantile Regression (QR)	Solar Energy
Manish K. Thukral	2020	MLFNN, MLR, NN	Solar Power
Yitao Long	2020	Ensemble Learning Algorithm, support vector regression,	Wind Energy
Tanveer Ahmad	2020	Artificial Neural Networks (ANN), Ensemble Models	Wind, Solar, and Geothermal Energy, and Electricity
Divyaang Agarwal	2020	Regression Model	Solar Power
Zhen Wang	2020	XGBoost Model, LightGBM, and LSTM Model	Solar Power
Md Amimul Ehsan	2020	Deep structured Network, CNN, LSTM	Wind Power
Lin Qiao	2019	Neural Network, BP and NIMF	Wind Energy

III. CLASSIFICATION OF MACHINE LEARNING MODELS USED IN PREDICTION

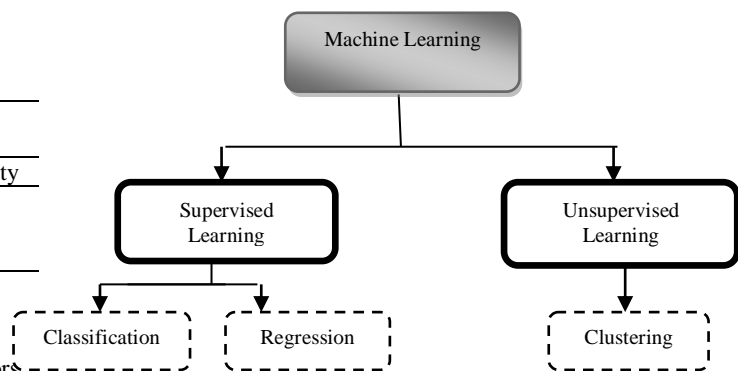


Fig.1 Types of Machine Learning

1. **Supervised Machine Learning:** It is a type of ML algorithm which consists of input variables and anticipated outcomes. Supervised ML design a model that produces predictions depending upon confirmation. It makes realistic forecasting for the reaction to the output by using data of input.

There are two types of SML:

- **Classification techniques** are used for the prediction of discrete responses, such as whether an email is actual or spam. Medical imaging, speech recognition, and credit scoring are some of the applications of classification techniques. Further types of classification algorithms are as:
 - **Support Vector Machine:** It is a type of SML, used for solving the issues of discrete and continuous classes. SVM is a well-known method that was created to solve binary classification problems [6]. Its aim is to observe the optimal track which means the supplementary information may be easily put down into an accurate section.
 - **k-Nearest Neighbor:** It is a simple ML technique. This approach records the current information and categorizes new information i.e the new information can be immediately categorized into a certain class. Also known as a non-parametric technique and typically employed for classification instead of binary.
 - **Logistic Regression:** It is the most popular and famous approach of SML. It predicts a definite constrained quantity from an unconstrained quantity. LR allocates new data using distinct as well as continual indices. LR is applied to designate results based on distinct data and can rapidly point out the most functional components for classification.
 - **Naive Bayes:** It is a Bayes formula-dependent SML. It is sometimes often utilized in script categorization job that requires a big index. It produces predictions depending upon the possibility of an object, so-called a probabilistic classifier.
- **Regression Techniques** are used for the prediction of continuous responses, such as modification in temperature. Electricity load forecasting and algorithmic trading are the applications of regression techniques. Further types of regression algorithms are as:
 - **Neural Network:** Also called ANN. To increase the accuracy over time, a neural network used the training data. These algorithms allow us to speedily colligate and gather data as tuned for accuracy. ANNs can be used for forecasting, regression, and curve fitting.
 - **ANFIS Learning:** An ANFIS, also known as an adaptive network-based fuzzy inference system, is a type of artificial neural network based on the Takagi–Sugeno fuzzy inference system. The capabilities of both fuzzy logic and neural networks are combined in these techniques [7]. [1]
 - **Stepwise Regression:** A model in which the independent variables to be utilized in the final model are chosen step by step. It entails incrementally adding or eliminating potential explanatory factors, with each iteration requiring statistical significance assessment. [2]

2.. Unsupervised Machine Learning: This technique reveals unseen design in reports. It's used to draw inferences from indices with no tagged solutions. In unsupervised machine learning, training data sets do not have expected outputs. It detects patterns based on the characteristics of input data. The most frequent unsupervised learning technique is clustering. It is employed in investigating data analysis in order to uncover unseen designs in reports. Further types of unsupervised machine learning are as:

- **Hidden Markov Model:** HMM is a new version of the basic Markov model that is designed for situations where the system is not directly visible, i.e. hidden, but the output is visible depending upon the internal states [6].
- **K-Means Clustering:** It's used to figure out if there's a difficulty with clustering. It offers a straightforward technique to gather information into multiple classes and a rapid way to determine the classes in unlabelled indices without requiring any instruction.
- **Hierarchical Clustering:** It is an approach to arranging similar items into groups. The terminal is a series of groups where each one is separate from everyone else but has broadly comparable objects within it.
- **Apriori Algorithm:** In the year 1994, R. Agrawal and Srikant presented this algorithm. The Apriori algorithm generates association rules by using repeated item sets, and it is designed to deal with databases that contain transactions. To efficiently calculate the itemset relationships, this approach uses a breadth-first search and a Hash Tree. It is an iterative method for locating common item sets in a large dataset.

IV. CONCLUSION

The growth of renewable energy depends on the extension of Machine Learning techniques. Because of the recent popularity of RE as a result of climate change and environmental concerns, gathering power from the wind, solar, and other RESs is becoming more acceptable. All RES have one thing in common: they all rely on the surrounding, which creates considerable management and planning issues. Moreover, due to the enlargement of the power supply system, power generation and demand must be predicted. As a result, machine learning models have become useful in such energy systems. Various machine learning models such as ANN, RF, DT, MLP, etc. are used for the prediction of Renewable energy generation. The above techniques can be used to get better results and achieve accuracy.

REFERENCES

- [1] H. Wang, Z. Lei, X. Zhang, B. Zhou, and J. Peng, "A review of deep learning for renewable energy forecasting," *Energy Conversion and Management*, vol. 198. Elsevier Ltd, Oct. 15, 2019. doi: 10.1016/j.enconman.2019.111799.
- [2] L. L. Li, S. Y. Wen, M. L. Tseng, and C. S. Wang, "Renewable energy prediction: A novel short-term

- prediction model of photovoltaic output power,” *Journal of Cleaner Production*, vol. 228, pp. 359–375, Aug. 2019, doi: 10.1016/j.jclepro.2019.04.331.
- [3] Z. Qadir et al., “Predicting the energy output of hybrid PV–wind renewable energy system using feature selection technique for smart grids,” *Energy Reports*, vol. 7, pp. 8465–8475, Nov. 2021, doi: 10.1016/j.egy.2021.01.018.
- [4] Rahul, A. Gupta, A. Bansal, and K. Roy, “Solar energy prediction using decision tree regressor,” in *Proceedings - 5th International Conference on Intelligent Computing and Control Systems, ICICCS 2021*, May 2021, pp. 489–495. doi: 10.1109/ICICCS51141.2021.9432322.
- [5] M. M. Rahman et al., “Prospective methodologies in hybrid renewable energy systems for energy prediction using artificial neural networks,” *Sustainability (Switzerland)*, vol. 13, no. 4. MDPI AG, pp. 1–28, Feb. 02, 2021. doi: 10.3390/su13042393.
- [6] I. Ullah, R. Ahmad, and D. H. Kim, “A prediction mechanism of energy consumption in residential buildings using hidden markov model,” *Energies*, vol. 11, no. 2, Feb. 2018, doi: 10.3390/en11020358.
- [7] A. Mosavi, M. Salimi, S. F. Ardabili, T. Rabczuk, S. Shamshirband, and A. R. Varkonyi-Koczy, “State of the art of machine learning models in energy systems, a systematic review,” *Energies*, vol. 12, no. 7. MDPI AG, 2019. doi: 10.3390/en12071301.
- [8] M. H. Fathollahzadeh, A. Speake, P. C. Tabares-Velasco, Z. Khademiyan, and L. L. Fight, “Renewable energy analysis in indigenous communities using bottom-up demand prediction,” *Sustainable Cities and Society*, vol. 71, Aug. 2021, doi: 10.1016/j.scs.2021.102932.
- [9] X. Chang, W. Li, J. Ma, T. Yang, and A. Y. Zomaya, “Interpretable Machine Learning In Sustainable Edge Computing: A Case Study of Short-Term Photovoltaic Power Output Prediction,” *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020pp. 8981–8985, doi: 10.1109/ICASSP40776.2020.9054088.
- [10] C. Zhang et al., “An Improved ELM Model Based on CEEMD-LZC and Manifold Learning for Short-Term Wind Power Prediction,” in *IEEE Access*, vol. 7, pp. 121472–121481, 2019, doi: 10.1109/ACCESS.2019.2936828.
- [11] L. Qiao et al., “Wind power generation forecasting and data quality improvement based on big data with multiple temporal-spatial scale,” *2019 IEEE International Conference on Energy Internet (ICEI)*, 2019, pp. 554–559, doi: 10.1109/ICEI.2019.00104.
- [12] Khan, Y. C. Byun et al., “Machine learning-based approach to predict energy consumption of renewable and nonrenewable power sources,” *Energies*, vol. 13, no. 18, Sep. 2020, doi: 10.3390/en13184870.
- [13] M Xia, H. Shao et al., “A Stacked GRU- RNN-Based Approach for Predicting Renewable Energy and Electricity Load for Smart Grid Operation,” in *IEEE Transactions on Industrial Informatics*, vol. 17, no. 10, pp. 7050–7059, Oct. 2021, doi: 10.1109/TII.2021.3056867.
- [14] J. Faraji, A. Abazari, M. Babaei, S. M. Muyeen, and M. Benbouzid, “Day-ahead optimization of prosumer considering battery depreciation and weather prediction for renewable energy sources,” *Applied Sciences (Switzerland)*, vol. 10, no. 8, Apr. 2020, doi: 10.3390/APP10082774.
- [15] S. G. Kim, J. Y. Jung, and M. K. Sim, “A two-step approach to solar power generation prediction based on weather data using machine learning,” *Sustainability (Switzerland)*, vol. 11, no. 5, 2019, doi: 10.3390/SU11051501.
- [16] K. Thukral, “Solar power output prediction using multilayered feedforward neural network: A case study of Jaipur,” Dec. 2020. doi: 10.1109/iSSSC50941.2020.9358821.
- [17] S. de Jongh, et al, “Spatio-Temporal Short Term Photovoltaic Generation Forecasting with Uncertainty Estimates using Machine Learning Methods,” *2020 55th International Universities Power Engineering Conference (UPEC)*, 2020, pp. 1–6, doi: 10.1109/UPEC49904.2020.9209764.
- [18] M. AlKandari and I. Ahmad, “Solar power generation forecasting using ensemble approach based on deep learning and statistical methods,” *Applied Computing and Informatics. Elsevier B.V.*, 2019. doi: 10.1016/j.aci.2019.11.002.
- [19] P. Jia, H. Zhang, X. Liu, and X. Gong, “Short-Term Photovoltaic Power Forecasting Based on VMD and ISSA-GRU,” *IEEE Access*, vol. 9, pp. 105939–105950, 2021, doi: 10.1109/ACCESS.2021.3099169.
- [20] Y. Long and R. Zhang, “Short-term Wind Speed Prediction with Ensemble Algorithm,” in *Proceedings - 2020 Chinese Automation Congress, CAC 2020*, Nov. 2020, pp. 6192–6196. doi: 10.1109/CAC51589.2020.9326917.
- [21] G. Mohy-Ud-Din, K. M. Muttaqi, and D. Sutanto, “Adaptive and Predictive Energy Management Strategy for Online Optimal Power Dispatch from VPPs with Renewable Energy and Energy Storage,” Oct. 2020. doi: 10.1109/IAS44978.2020.9334772.
- [22] K.B. Navas Raja Mohamed and S. Prakash, “A Systematic Review on Wind Energy Resources Forecasting by Neural Network,” Dec. 2020. doi: 10.1109/ICRAIE51050.2020.9358370.
- [23] Lennard Visser, Tarek AlSkaif, and Wilfried van Sark, “Benchmark analysis of day-ahead solar power forecasting techniques using weather predictions”. DOI:10.1109/PVSC40753.2019.8980899
- [24] Z. P. Ncane and A. K. Saha, “Forecasting Solar Power Generation Using Fuzzy Logic and Artificial Neural Network,” *2019 Southern African Universities Power Engineering Conference/Robotics and Mechatronics/Pattern Recognition Association of South Africa (SAUPEC/RobMech/PRASA)*, 2019, pp. 518523, doi: 10.1109/RoboMech.2019.8704737.
- [25] A. Khalyasmaa et al., “Prediction of Solar Power Generation Based on Random Forest Regressor Model,” *2019 International Multi-Conference on Engineering, Computer and Information Sciences (SIBIRCON)*, 2019, pp. 0780–0785, doi: 10.1109/SIBIRCON48586.2019.8958063.
- [26] K. J. Nam, S. Hwangbo, and C. K. Yoo, “A deep learning-based forecasting model for renewable energy scenarios to guide sustainable energy policy: A case study of Korea,” *Renewable and Sustainable Energy Reviews*, vol. 122, Apr. 2020, doi: 10.1016/j.rser.2020.109725.
- [27] N. Somu, G. Raman M R, and K. Ramamritham, “A deep learning framework for building energy consumption forecast,” *Renewable and Sustainable Energy*

- Reviews, vol. 137, Mar. 2021, doi: 10.1016/j.rser.2020.110591.
- [28] D. Pintossi, C. Simões, M. Saakes, Z. Borneman, and K. Nijmeijer, "Predicting reverse electro dialysis performance in the presence of divalent ions for renewable energy generation," *Energy Conversion and Management*, vol. 243, Sep. 2021, doi: 10.1016/j.enconman.2021.114369.
- [29] R. K. Jana et al., "A granular deep learning approach for predicting energy consumption," *Applied Soft Computing Journal*, vol. 89, Apr. 2020, doi: 10.1016/j.asoc.2020.106091.
- [30] J. Wang, S. Chung, A. AlShelahi, R. Kontar, E. Byon, and R. Saigal, "Look-ahead decision making for renewable energy: A dynamic 'predict and store' approach," *Applied Energy*, vol. 296, Aug. 2021, doi: 10.1016/j.apenergy.2021.117068.
- [31] Y. Cao, A. Raise et al., "Deep learned recurrent type-3 fuzzy system: Application for renewable energy modeling/prediction," *Energy Reports*, vol. 7, pp. 8115–8127, Nov. 2021, doi: 10.1016/j.egy.2021.07.004.
- [32] T. Ahmad, H. Zhang, and B. Yan, "A review on renewable energy and electricity requirement forecasting models for smart grid and buildings," *Sustainable Cities and Society*, vol. 55. Elsevier Ltd, Apr. 01, 2020. doi: 10.1016/j.scs.2020.102052.
- [33] M. A. Zamee and D. Won, "Novel mode adaptive artificial neural network for dynamic learning: Application in renewable energy sources power generation prediction," *Energies*, vol. 13, no. 23, Dec. 2020, doi: 10.3390/en13236405.
- [34] Habib Hadj-Mabrouk, "Analysis and prediction of railway accident risks using machine learning". <https://doi.org/10.3934/ElectrEng.2020.1.19>
- [35] C. Liu, W. Z. Wu, W. Xie, and J. Zhang, "Application of a novel fractional grey prediction model with time power term to predict the electricity consumption of India and China," *Chaos, Solitons and Fractals*, vol. 141, Dec. 2020, doi: 10.1016/j.chaos.2020.110429.

SECURE DATA SHARING USING BLOCKCHAIN INVEHICULAR SOCIAL NETWORKS

Mamidala Sruthi¹, Ranjith Kumar Marrikkukala², Dadi Ramesh³

^{1,2} Department of Computer Science and Engineering, Sumathi Reddy Institute of Technology for Women, Warangal, Telangana, India

³ School of Computer Science & Artificial Intelligence, SR University, Telangana, India
sruthi.m527@gmail.com

Abstract — There are many other types of associations that the VSNs stay on top of. A few examples include: street blossoming and information sharing. A dynamic affiliation structure and stuttering growth provide new security challenges in any situation. Securing transmission of information has risen to the top of the list of challenges. To recall one-to-different details concerning VSNs, ciphertext-system quality-based encryption (CP-ABE) may be adopted. The access system is handled with and supplied by the could in typical CP-ABE designs, which requires authenticity due to centralization because it is centralized. To address the aforementioned problem, we present in this study a well-defined one-to-different information sharing expectation. As part of our system, we leverage blockchain to record how clients perceive self-demand while also ensuring cloud non-renunciation. We propose a convincing certification strategy based on the vehicle client's figures at the farthest reaches. We offer a technique hiding arrangement in the meantime, contemplating the unstable data associated with the philosophy of the way. The design we've come up with also accommodates information denouncement when a vehicle client no longer requires sharing information in VSNs.

Keywords— CP-ABE, Vehicular Social Network, data revocation.

I. INTRODUCTION

VANET intends to organize the information flow among vehicles distant from their capability when the climate changes. As a result of VANET, cars and pilgrims in the area will be more connected than ever before. Autonomous networks and loose associations like traffic pioneers, street thriving, and information sharing are grouped together under the banner of VSNs. Temporary social gatherings of people with similar interests, proclivities, or needs are referred to as VSNs. The location, parking place, and/or national identification number of clients in Types of structural may be disclosed by them. Because of VSN's dynamic affiliation structure, fairly clear and multi-jump propagation are crucial. Despite your best efforts, information can still be compromised throughout either of these two cycles, regardless of your level of caution. VSN safety affirmation is crucial because of this.

Check to see if any data has been scrambled before sharing. Data can be scrambled with the general public car keys for one-to-one information exchange. When it came to transmitting information, access management is also a big challenge [6-11]. Male cab drivers nearing the age of 30 should receive confidential voice messages from taxi groups,

for example. He'll have to demonstrate how an ability to keep track system is working, basically. As long as the client is a [male][taxi driver], over 30 years old], they have access to the data. CP-ABE (ciphertext approach brand identity based encryption) is one of the more perplexing encryption degrees of advancement when it comes to data exchange and consent control. Each ciphertext is segregated using an entry control mechanism in CP-ABE, and the secret keys of the each client are assigned to unique attributes. A ciphertext can be deciphered by a client whose credits match the section control structure.

To facilitate the ongoing development of CP-benefit ABEs, customers' consent is withdrawn to the cloud, which means that customers' consent is withdrawn from the cloud. With a definite aim in sight, conventional CP-ABE strategies are useless and inaccurate in their current state since the cloud is considered as untouchable.

To address the CP-ABE security issue, we favored distributed acceptance control. Bitcoin and other latest bank structures are based on a new decentralized system known as Blockchain, which is attracting ideas from a variety of organizations. Blockchain technology is characterized by decentralization, simplicity, ego, and immutability. Only the record of interactions between two social occasions is missing from this picture. The difficulty with conventional CP-ABE can be solved using blockchain because once the trades are recorded, they can't be changed. The crypto currency section control method can reveal client's self and cloud non-renunciation. In addition, we can store the information's hash value to guard against a data-altering assault.

To summarize, the use of CP-ABE, a blockchain-based protocol, allows VSNs to securely and economically exchange data.

A VSN information exchange approach built on top of CP-ABE and the blockchain is proposed in this project. To maintain track with one information sharing, we use CP-ABE software. We can guarantee client's self while simultaneously preventing cloud non-denial by using crypto currency as an interim measure. Our certification methods are also effective due to the fact that we're dealing with VSNs' focus breaking points

II. BACKGROUND WORK

A. VSNs

Institute Of technology was the first to explore the possibility of vehicle social ties in 2006. Along with this, the teachers additionally powered an engineering structure called Flosster, which was used to divide knowledge across car partners. GM and BMW, for example, have already implemented social sharing modules in their vehicles. Whatever the case may be, everything works except for the problem of how to create a simple affiliation in VANETs. IP Multimedia Sub - system and Machines and Machines restrictions were presented by Lequerica et al. as an approach for fostering social affiliation. It was discussed in Abbani et al. paper 's how to cope with the trust issue in VANETs of laid-back organizations. Local engagement and interpersonal similarity were used to inform the information-sending strategy developed by Li and colleagues (LASS). For example, Oliveria et al. advocated for the use of approvals to facilitate the transfer of encrypted material in small-scale networks, such as social gatherings. In this way, customers in the affiliation build trust, and notoriety can turn into a reward for customers who give information in an appropriate manner. An affirmation display that ensures security among changeable targets was proposed by Xu and colleagues in 2018. As of 2019, Cheng et al. recommended using Three-Valued Subjective Logic to evaluate VSN user trust. Most existing plans are based on PKI and cannot handle one-to-one information exchange or finely-grained authorization controls. This is a problem.

B. CP-ABE

When Sahai and Waters presented feathered character based encryption in 2004, it was a radical idea (FIBE). The owner of the data may decide to divide it among clients who have a specific set of attributes in mind. Data owners could implement access control by setting up access techniques with Bethencourt et al. crucial's CP-ABE contrive, which allowed them to do so. One of the ideas put out by Melissa was a multi-expert ABE plan. A variety of topic matter specialists were credited in his research, allowing him to address the issue of a single sign of disappointments. Customers' problematic keys were divided into characteristic private keys and unwinding secret keys, according to a plan proposed by Green et al.

We outsourced all of the unscrambling overhead to a third-party cloud labor provider. When an attribute was renounced, the data owner simply reactivated the variation numbers contained in the quality secret keys and distributed them to the legal consumers, as Yang et al. proposed. Researchers in the field of CP-ABE have made incredible development in the last few years. In the future, CP-ABE is expected to be the most suitable advancement for recognizing fine-grained induction control. Given the sensitivity of the tunnel system's data, a few ideas for information concealment were floated. [38] Nishide et

colleagues suggested an ABE plot with 'AND' doorway authentication scheme that kept the methods stowing indefinitely to some extent. It was Lai et al. who initially came up with the idea of using LSSS to disguise an ABE scheme, but that plan didn't take account of consumer rejection. For a distributed capacity, Zhong and colleagues suggested a technique called Multi-authority quality encryption - based access control system with a procedure hidden. However, this scheme had an extravagant calculation cost. However, data owners believed that the passing mechanism would be changed by a production before encoding after Fan et al. successful ABE plot.

C. Blockchain

When Sahai and Waters presented feathered character based encryption in 2004, it was a radical idea (FIBE). The owner of the data may decide to divide it among clients who have a specific set of attributes in mind. Data owners could implement access control by establishing access techniques with Bethencourt et al. crucial's CP-ABE technology, which allowed them to do so. One of the ideas put out by Melissa was a multi-expert ABE plan. A variety of topic matter specialists were credited in his research, allowing him to address the issue of a single sign of disappointments. Customers' problematic keys were divided into characteristic private keys and unwinding secret keys, according to a plan proposed by Green et al.

We outsourced all of the unscrambling expenses to a third-party cloud labor provider. When a characteristic was renounced, the data owner simply reactivated the variation numbers contained in the excellent secret keys and distributed them to the legal consumers, as Yang et al. proposed. Researchers in the field of CP-ABE have made incredible development in the last few years. In the future, CP-ABE is expected to be the most suitable advancement for recognizing fine-grained inducement control. Given the sensitivity of the tunnel system's data, a few ideas for information concealment were floated. [38] Nishide et colleagues suggested an ABE plot with 'AND' doorway authentication scheme that kept the methods stowing indefinitely to some extent. It was Lai et al. who initially came up with the idea of using LSSS to disguise an ABE scheme, but that plan didn't take account of consumer rejection. For a distributed capacity, Zhong and colleagues suggested a technique called Multi-authority quality data encryption physical access system with a procedure hidden. However, this scheme had an extravagant calculation cost. However, data owners believed that the passing mechanism would be changed by a production before decoding after Fan et al. successful ABE plot.

III. PROPOSED WORK

The concept of feathery character based encryption, first

described by Sahai and Waters in 2004, was considered innovative (FIBE). There is the possibility that the data's owner will opt to distribute it among clients who have a specific set of characteristics in mind. Data owners could use Bethencourt et al. crucial's CP-ABE technology to build access strategies for access control. Melissa proposed a multi-expert ABE plan as an option. His research included contributions from a wide range of subject matter experts, which allowed him to focus on the problem of a single sign of dissatisfaction. Green et al. advocated dividing customers' troublesome keys into typical private keys and unwinding secret keys.

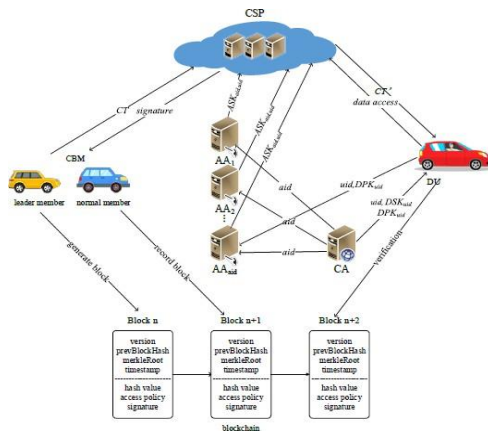


Fig. 1 System Model

We hired a third-party cloud labor provider to handle all of the unscrambling costs. Because of this, the data owner merely reactivated variant numbers in excellent secret keys and distributed them to lawful consumers, as Yang et al. proposed when a characteristic was renounced in their work. A lot of progress has been made recently in the realm of CP-ABE research. Fine-grained inducement control will soon be recognized using CP-ABE, according to future predictions. Many solutions for information hiding were floated due to the importance of the tunneling system's data. [38] A BE plot with 'AND' doorway verification was proposed by Nishide and colleagues, and it retained the techniques stowing forever to some extent. In the beginning, Lai et al. thought of employing LSSS to mask an ABE system, however they failed to account for consumer rejection. Zhong and coworkers proposed a technique metric I_s used to give quality data encrypting access control system with a concealed procedure for distributing computing power. However, the computational cost of this plan was astronomical. In the wake of Fan et al. ABE plot, data owners hoped that a production would alter the transfer mechanism prior to decoding.

A. Implementation Module

1) *Cloud Service Provider*: The CBM sends the cipher -

text to the CSP, which decrypts it and sends back the ciphertext's signature. DUs' feature secret keys are stored in CSP, and these keys will be used to decrypt the encrypted before it is transmitted.

2) *Attribute Authority* : To recognize DUs and generate DU attribute private key within its administrative domain, the AA has been signed by a worldwide distinct character aid. This is followed by sending every attribute secret key and associated user identification uid to CSP. Our approach allows each AA to manage several attributes, however each AA can only control one characteristic at a time.

3) *Certificate Authority*: The Certificate Authority (CA) is a globally recognized and respected name in the IT industry. Each lawful AA and user can register with it, and it will provide a global unique identifier aid and uid to each of them. The encryption secret key is generated for each authenticated person while this is going on. It does not, however, take part in the maintenance of attributes or the development of secret keys for attributes.

4) *Blockchain*: The CSP is monitored using the blockchain. We employ a consortium blockchain in our system, and the participants are all licensed drivers. The ciphertext's CSP signature is included in each block's body, along with the hash value of the shared data and the related access policy. We use the PBFT consensus mechanism to guard against malicious attackers.

5) *Data User*: Globally unique identifiers uids certify data requesters known as DUs. By checking their qualities against the access rules defined in the blockchain, individuals can get to data before anybody else does. They can check the data for tampering after decoding the ciphertext. The encrypted message can only be decrypted by DU if its properties comply with the access policy. CBMs are also known as DUs in the system.

IV. CONCLUSIONS

In this venture, we have proposed a safe and certain information sharing plan in VSNs, which depends on both CP-ABE and blockchain. In our plan, we have created CP-ABE to acknowledge one-to- numerous information sharing. In the mean time, we have likewise evolved blockchain to record the entrance strategy of the information, acknowledging client self-affirmation and cloud non-disavowal. Considering the processing abilities of the VSNs hub, we have proposed a compelling plan for certificating. We have planned a strategy concealing plan to conceal the touchy data remembered for the entrance strategy. Our plan likewise upholds information renouncement when a vehicular client no longer needs to

share the information on the cloud. Later on, we will investigate on the most proficient method to diminish the hour of arriving at agreement.

REFERENCES

- [1] L. FAN, AND Y. WANG. "ROUTING IN VEHICULAR AD HOC NETWORKS: A SURVEY." *IEEE VEHICULAR TECHNOLOGY MAGAZINE*, VOL. 2, NO. 2, PP. 12-22, 2007.
- [2] J. WU ET AL., "FCSS: FOG COMPUTING BASED CONTENT-AWARE FILTERING FOR SECURITY SERVICES IN INFORMATION-CENTRIC SOCIAL NETWORK." 1-1, 2017.
- [3] K. ZHANG, X. LIANG, J. NI, K. YANG, AND X. SHEN. "EXPLOITING SOCIAL NETWORK TO ENHANCE HUMAN- TO-HUMAN INFECTION ANALYSIS WITHOUT PRIVACY LEAKAGE." *IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING*, VOL. 15, NO. 4, PP. 607-620, 2018.
- [4] H. REN, ET AL., "QUERYING IN INTERNET OF THINGS WITH PRIVACY PRESERVING: CHALLENGES, SOLUTIONS AND OPPORTUNITIES." *IEEE NETWORK*, VOL. 32, PP. 144-151, 2018.
- [5] L. GUO, ET AL., "A SECURE MECHANISM FOR BIG DATA COLLECTION IN LARGE SCALE INTERNET OF VEHICLE." *IEEE INTERNET OF THINGS JOURNAL*, VOL. 4, PP. 601-610, 2017.
- [6] K. FAN, ET AL., "CLOUD-BASED RFID MUTUAL AUTHENTICATION SCHEME FOR EFFICIENT PRIVACY PRESERVING IN IoV." *JOURNAL OF THE FRANKLIN INSTITUTE* (2019).
- [7] G. XU ET AL., "DATA SECURITY ISSUES IN DEEP LEARNING: ATTACKS, COUNTERMEASURES, AND OPPORTUNITIES." *IEEE COMMUNICATIONS MAGAZINE*, VOL. 57, NO. 11, PP. 116-122, 2019.
- [8] K. FAN, ET AL., "A LIGHTWEIGHT AUTHENTICATION SCHEME FOR CLOUD-BASED RFID HEALTHCARE SYSTEMS." *IEEE NETWORK*, 2019.
- [9] G. XU ET AL., "ENABLING EFFICIENT AND GEOMETRIC RANGE QUERY WITH ACCESS CONTROL OVER ENCRYPTED SPATIAL DATA." *IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY*, VOL. 14, NO. 4, PP. 870-885, 2019.
- [10] G. XU ET AL., "EFFICIENT AND PRIVACY-PRESERVING TRUTH DISCOVERY IN MOBILE CROWD SENSING SYSTEM." *IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY*, VOL. 68, NO. 4, PP. 3854-3865, 2019.

WILL CSK WIN IPL 2022? APPLICATION OF MACHINE LEARNING ALGORITHMS IN SPORTS

Uma Maheswari B
PSG Institute of Management
Coimbatore, India
0000-0002-7487-7783

Selva Gomathy R
PSG Institute of Management
Coimbatore, India
selvagomathyr@gmail.com

Sujatha R
PSG Institute of Management
Coimbatore, India
0000-0002-5622-8392

Kavitha D
PSG Institute of Management
Coimbatore, India
0000-0002-2040-6871

Abstract— The application of analytics in games has totally transformed the way sports have been played and viewed across different platforms. Analytics has gained traction in designing training programs, in-game strategies, health and fitness of the players, team selection, and game viewership. Machine Learning algorithms aid in tracking the performance of the players as well as the team, providing a clear outline of the game strategies, recognizing the risks involved, and forecasting the match outcome. The pace of the game of cricket has evolved over a period of time and one of the most popular formats of cricket is the T20. The Indian Premier League (IPL) is a popular T20 league which has millions of fans across the globe. Because of the popularity and the money involved in this format, the stakes for winning the game are extremely high. Application of analytics in this aspect would help in better decision-making for a sports team. This paper aims at the application of machine learning algorithms to predict the results of the IPL matches played by the Chennai Super Kings (CSK) franchise in 2022 and, furthermore, to identify the factors that impact the team's ability to win the game. The forecast is based on historical data sourced from the official website of CSK. Model performance metrics are applied to measure and compare the accuracy, sensitivity, and specificity of each model, and the results are discussed. The insights from the outcome can be utilized to formulate strategies to increase the team's chances of winning the game.

Keywords— *Predictive analytics, Machine learning algorithms, Prediction model, sports analytics, IPL component*

I. INTRODUCTION

Cricket is one of the most cherished sports with a huge fanbase of billions of people across the globe. The popularity of cricket increased with the introduction of the IPL (Indian Premier League), a new T20 tournament league introduced in the year 2007. The IPL is a short 3-hour format of the game of cricket with a limit of 20 overs, which has grabbed the attention of many cricket enthusiasts. The International Cricket Council (ICC) sanctioned the IPL format suggested by the Board of Control for Cricket in India (BCCI). It was founded with eight teams, namely Chennai Super Kings, Kolkata Knight Riders, Sunrisers Hyderabad, Mumbai Indians, Royal Challengers Bangalore, Delhi Capitals, Rajasthan Royals, and Punjab Kings, with each team representing a city in India. Cricket is a profit-making business with multiple stakeholders, including the athletes, sponsors, spectators, media, and corporate entities. Winning the tournament is extremely crucial for many of the stakeholders mentioned, more so because of the monetary implications for the winning franchise. This is where the application of analytics in sports comes into the fray.

The global sports analytics market is expected to reach \$4.6 billion by 2025, at a CAGR of 31.2 percent (source: Grand View Research Inc.). Each match of cricket generates a massive amount of data. Every single detail of the game, a dot ball, a missed catch, the ball's twist, the speed of the ball bowled, the batsman's shot, etc., is captured and analyzed using machine learning. Data is also collected off the field from the athletes with the aim of better understanding and optimizing their training, conditioning, and health status. Analytics has altered the game's course, and leveraging data science in sports would help with accurate decision-making and formulating game strategies, which are critical for success. Sports franchises use analytics to track player performance, decide on team compositions, devise game strategies and, in turn, create a sustainable competitive advantage.

The players' performance is measured using a variety of metrics, including the bowling speed of the player, the capacity to lift weights, and the nutrients they consume throughout the day. As much as the players must continue to focus on their individual performance, it is equally essential to work as a team for success in matches. Team coaches try different player combinations on the field to see if different line-ups result in better performance. The team's management concentrates on statistics about the opponent's players' batting and bowling styles. They track match-day weather conditions and players' strengths or weaknesses based on the prevalent conditions. They devise strategies such as how many games they must win in order to qualify for the playoffs or break past records. The franchise is able to develop marketing strategies and advertising campaigns with the help of data by determining how and when fans are likely to attend events or purchase team merchandise.

II. LITERATURE REVIEW

Studies have applied machine learning-based algorithms that predicts the cost at which a player can be purchased in the Indian Premier League Auction [8,3]. The players' selling prices were determined using past performance measures like runs, balls, innings, wickets, and games played. Multiple Random Forest Regression was used to predict the batsmen and bowlers' attributes in a given match, which would aid in the selection of players for a given tour [9]. Studies were also conducted to measure the popularity and people's feedback on IPL 2020 using machine learning algorithms. With the help of application interface, about 7 lakhs feedback were collected from the official IPL twitter website to analyze the emotions of the fans about the game [7]. The study on the trade-off between

accuracy and interpretability in machine learning was addressed by a mimic learning model that allowed sports analytics to combine the forecasting power of modern machine learning techniques with interpretations and actionable insights for sports consultants [14]. Machine learning in sports offers predictive and analytic solutions to track a player's wellness and safety. An attempt to create a model for predicting the performance of the Indian team in the Olympics was made by considering factors like GDP, population, wellbeing, education rate, etc. [11].

Artificial Intelligence models helped in determining the on-field playing positions of the football players, assessing injury risk and assessing fitness levels post injuries [10, 4]. The data-driven defensive strategies in basketball were studied to understand player decision making and game strategies by utilizing player and ball tracking information [16]. A study of predicting the number of runs the player might score and the number of wickets the bowler might take was carried out in order to help the team management select the best players for the match[5]. Studies also attempted to propose a model where the match result was anticipated ball by ball from the beginning of the second innings using various deep learning models like long-term memory (LSTM) and gated recurrent units (GRU) algorithms. The prediction of the total score using big data in the context of one-day international (ODI) cricket [2] was carried out. An investigation into machine learning technology was conducted to check if the algorithms could derive accurate predictive models for cricket matches and provide the best accuracy, precision, and recall evaluation measures [6,15]. The various aspects of a cricket match that can influence the outcome of a match by using several supervised learning techniques were identified. The strengths and shortcomings of the bowling and batting groups are used to improve the overall performance of the team and also to increase their winning probability [13,12]. The research was carried out using the KNIME tool and machine learning techniques like Naïve Bayes Classification and Euler's Strength Formula. A statistical modelling approach was used to predict the team players for each match using Hadoop and the Hive framework [1].

IPL has been a huge success in India and it is currently a billion-dollar industry attracting many investors. The 2021 IPL season set a massive viewership record with 380 million average impressions, and the 2022 IPL is all set to break new records. Chennai Super Kings (CSK) is one of the most consistent and successful franchises in the Indian Premier League (IPL). They have a huge fanbase not only within the country but also worldwide. Even though Mumbai Indians (MI) have won the highest number of IPL titles, CSK stands at the top with the highest winning percentage among its opponents, having made it to the playoff stage in 11 seasons and 9 times to the finals. CSK is a very formidable team and, therefore, was chosen to be the focus of our study. This paper attempts to create a machine learning model to predict the CSKs' win in the 2022 IPL matches. All match data from all matches played by CSK over the last 12 seasons was collated and the necessary preprocessing was done to prepare the data for model building.

III. METHODOLOGY

This research attempts to apply various machine learning algorithms in order to predict the result of CSK matches in the forthcoming IPL season (2022). The architecture for this study is presented in Fig. 1.

Fig.1 Architecture of the study

The dataset for prediction is gathered from the official website of CSK (www.chennaisuperkings.com), which consists of data from the past 12 seasons from 2008–2021.

The table below shows the dataset columns and their descriptions.

TABLE 1.DATASET DESCRIPTION

Column name	Description
city	The city where the match was played.
Season	The season (year) in which the match was played.
team1	Team 1 is taken as Chennai Super Kings (CSK).
team2	All the opponent teams of CSK are taken as team2.
toss_winner	The team that won the toss.
toss_decision	The toss winning team's decision of choosing batting or fielding.
dl_applied	Application of DL (Duckworth Lewis Method) method involved or not if the match got interrupted.
winner	The team that won the match.
win_by_runs	The number of runs by which the team won.
win_by_wickets	The number of wickets by which the team won.
player_of_match	The player who won the man of the match title.
venue	The stadium where the match took place.
umpire1	On field umpire1 name
umpire2	On field umpire2 name
extra_balls	Total number of extra balls given by CSK team.
wickets_taken	Total number of wickets taken by CSK team
4s	The number of fours scored by CSK.
6s	The number of sixes scored by CSK.
total_runs	Overall runs scored by the team against its opponent
Status (Dependent variable)	Won/Lost

The dataset contains information from 194 matches played by CSK. Data preprocessing involves cleaning the data by dealing with missing records and outliers. Feature selection was done by removing features from the dataset that did not have an impact on the outcome of the match result. The features that were so eliminated included the match id, the season and the date of the match. Some of the matches were abandoned without even a single ball being

bowled. Data pertaining to such matches was removed during the preprocessing stage. As a result, the dataset contains ten distinct teams. The data was then split into training and testing in the ratio of 75 to 25. Supervised machine learning algorithms like logistic regression, the k-NN algorithm, support vector machines, random forest, and decision trees are applied to create predictive models to predict match results. Table 2 presents the tuning parameters for each of the algorithms.

TABLE 2. MODEL TUNING PARAMETERS

Parameters	Explanation	Value
Decision tree model		
criterion	Measures the quality of the split and the criterion 'gini' gives the impurity value	gini
splitter	Used to choose the split at each node as best split or random split	best
max_depth	Maximum depth of the tree	3
random_state	Controls the randomness of the estimator	1234
Random Forest model		
n_estimators	No. of trees in the forest	50
K Nearest Neighbors Classification		
n_neighbors	No. of neighbors to be used	7
weights	Determines the closeness of points in each neighborhood	Uniform
metric	distance metric	Minkowski
p	p is the power parameter for minkowski metric	2

Model performance measures, including sensitivity, specificity, and accuracy, were calculated using the confusion matrix to arrive at the best prediction model.

TABLE 3 CONFUSION MATRIX

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

Accuracy refers to the overall accuracy of the model and is calculated as follows:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (1)$$

Sensitivity refers to the percentage of true positive values and is calculated as follows:

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (2)$$

Specificity refers to the percentage of true negative values and is calculated as follows:

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (3)$$

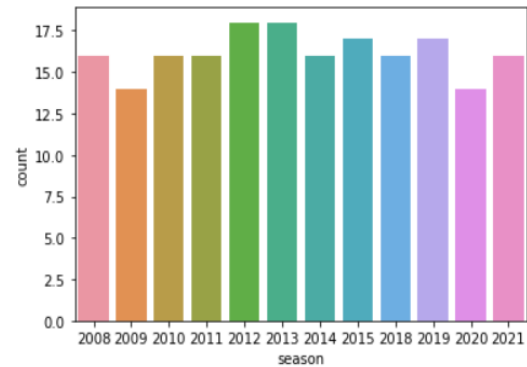
IV. ANALYSIS AND DISCUSSION

The exploratory data analysis helped in visualising the past performance of the CSK team. Figure 2 shows the number of games played by CSK in each season. Every franchise gets to play 14 mandatory matches, and from the

graph, it is clear that CSK has made it to the play-offs in almost 10 seasons.

Fig. 2: Total no. of matches played

Fig. 3 displays the most favourite stadium for the team in the past seasons, and it can be inferred that CSK has an



advantage of winning the game in MA Chidambaram (Chepauk) stadium, which is the home ground for the team. Wankhade and Dubai stadium are the next favourite grounds for the team. Fig. 4 displays the toss decisions taken by CSK and the outcome of the game. The probability of winning the game is higher when CSK has won the toss, and there is a slightly higher probability of winning when the decision to bat is taken after winning the toss.

Fig. 3: Favorite venue

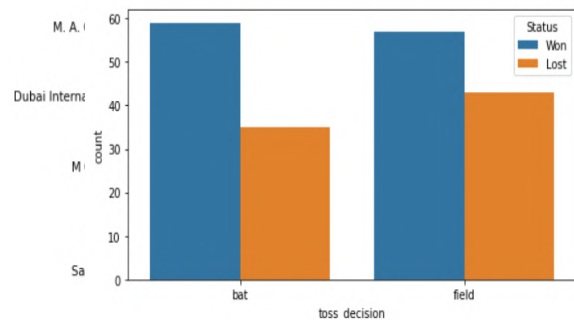


Fig. 4: Toss decision

CSK stands second next to Mumbai Indians in terms of title winners. The franchise has won the IPL title four times in 2010, 2011, 2018, and 2021. The year 2020 was the least performed season in the IPL's history, with only 6 wins and 8 losses. Under the captaincy of MS Dhoni, the team has a winning record of 117 matches in total.

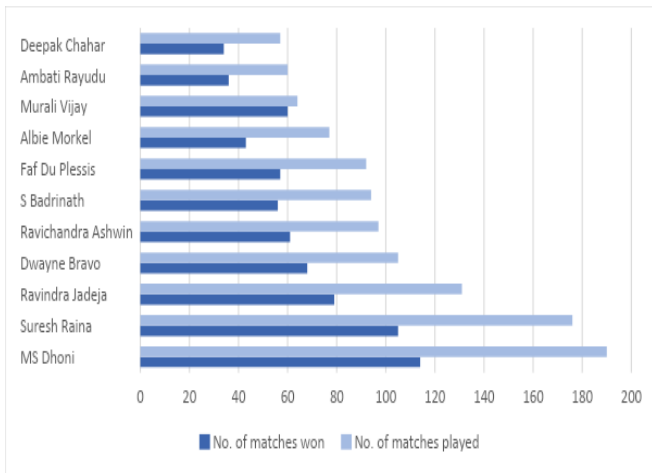
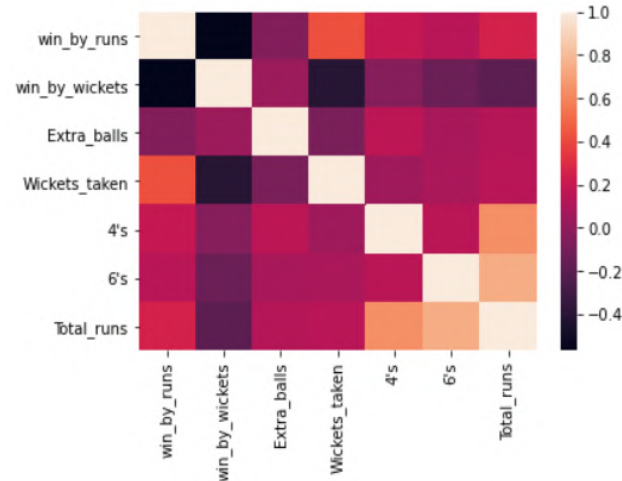


Fig. 5: CSK Players insights

From the above Fig. 5, we can understand that out of 194 matches played by CSK, MS Dhoni and Suresh Raina have been part of the franchise since 2008 and have played 190 and 176 matches, respectively. They are followed by Ravindra Jadeja, Dwayne Bravo, Faf Du Plessis, and Murali Vijay, who played for CSK till the 2021



IPL.

Fig. 6 Correlation heatmap of the continuous variables in the dataset

From the above heat map(Fig.6), the correlation between the continuous variables is inferred. The total runs scored by the team is highly correlated to the number of fours and sixes scored. The higher the number of wickets taken, the lower the opponent team's score would be; hence, the higher the chance of winning by runs. On the other hand, the win_by_runs and wickets_taken variables are negatively correlated to the extra_balls feature. This indicates that the possibility of winning decreases when a high number of extra balls are thrown at the opponent team. In the next stage, prediction models were built with the help of 11 key features, and the accuracy of the various machine learning models is shown in Fig. 7. From the Fig., it is clear that out of the 5 models used, support vector machines and logistic

regression models showed the highest accuracy of 95% and 92%, respectively.

Fig. 7: Prediction accuracy of the models

The model performance metrics are presented in Table 4.

TABLE 4: MODEL PERFORMANCE METRICS

Model	Accuracy	Sensitivity	Specificity
Logistic Regression	92%	92%	92%
k-Nearest Neighbors algorithm	75%	76%	76%
Support Vector Machine	95%	94%	95%
Random Forest	84%	84%	86%
Decision Tree	80%	80%	80%

The support vector machine and logistic regression models gave the best results, with an accuracy of 95% and 92%, respectively. Another metric to arrive at the accuracy of the model is the area under the curve and receiver operating characteristic. They indicate the degree of the model's ability to classify the positive and negative classes. The higher the AUC value, the more accurate the model is, and the AUC value for the support vector machine learning model was at 0.956 (Fig. 8), indicating a high level of accuracy. The variable importance plot (Fig. 9) shows that the number of wickets taken by the team, the number of fours and sixes scored, conceding scores with minimum extra balls, the ground venue, and the toss decision are the key variables that influence CSK's win in a game.

Fig. 8: ROC curve

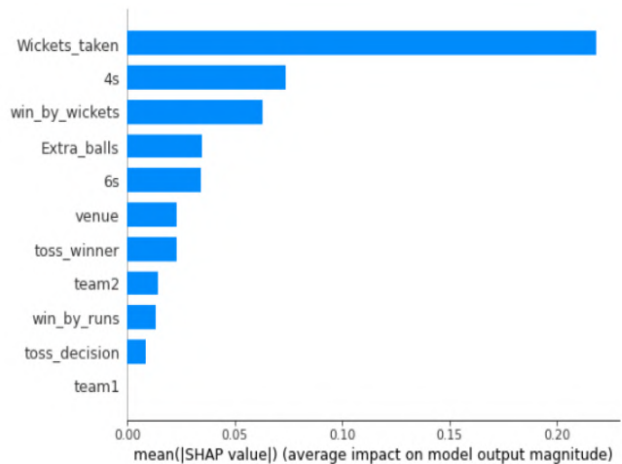
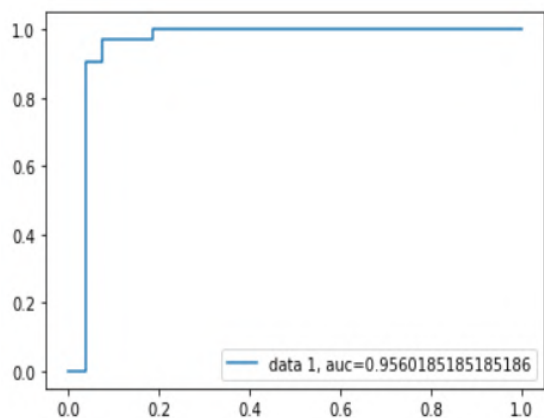
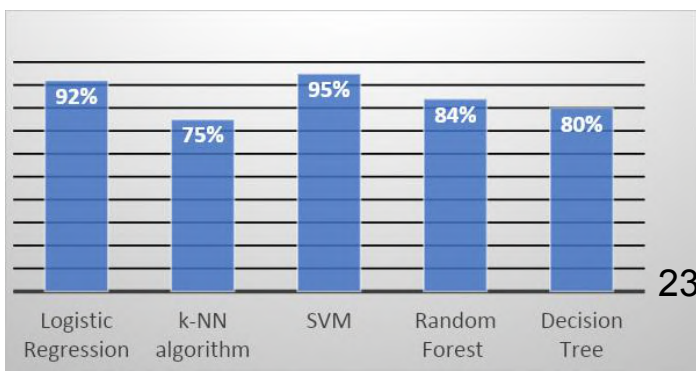


Fig. 9: Variable importance plot



V. IMPLICATIONS AND CONCLUSION

Prediction of outcomes in sports used to be a challenging exercise due to the intricate nuances of the game. But the application of analytics in the field of sports has made the process simpler and has generated very important insights that could transform the way the game is played and viewed. This paper focused on the application of machine learning techniques to create a prediction model to predict the team CSK's win in the upcoming IPL season of 2022. The IPL dataset was collated, preprocessed, and prepared for model building. Algorithms like Logistic regression, Random Forest, Support Vector Machines, K nearest neighbors, and Decision Tree were used and the highest accuracy of 95% was observed from the SVM model. The analysis also brought out the key factors that had a considerable influence on the team's chances of winning the game. The team has a high chance of winning the game when they play on their home ground. They have an average of 160 runs scored at the Chepauk stadium, with the highest total of 246 runs and a winning streak of 71% at the home venue. The early wickets taken during the powerplay overs have increased the probability of winning the game and also the number of fours and sixes put up a strong total on the board. Considering the toss, the team has won 66 games whenever they won the toss, regardless of their decision to bat or bowl after winning the toss. For about 50 matches, the team has lost the toss but won the game. Hence, winning the toss and the decision taken have a moderate impact. MS Dhoni stands as a valuable player and a pillar of support for the CSK team. He has been leading the franchise since 2008 as the captain, along with the title of the most successful wicket-keeper and the best finisher in IPL history. The study gives insights on the factors which need to be considered by CSK for winning the matches in the forthcoming IPL 2022 season. The number of wickets taken by the team is an important factor. Therefore, concentrating on capturing early wickets during the start of the game could enhance the chances of winning the game. The number of fours and sixes scored is also equally important. This model could be deployed by the team to formulate in-game strategies and also provide training for the players before the commencement of the match.

REFERENCES

- [1] Agarwal, Shubham, Lavish Yadav, and Shikha Mehta. "Cricket team prediction with hadoop: statistical modeling approach." *Procedia Computer Science* 122 (2017): 525-532.
- [2] Awan, Mazhar Javed, Syed Arbaz Haider Gilani, Hamza Ramzan, Haitham Nobanee, Awais Yasin, Azlan Mohd Zain, and Rabia Javed. "Cricket match analytics using the big data approach." *Electronics* 10, no. 19 (2021): 2350.
- [3] Bunker, Rory P., and Fadi Thabtah. "A machine learning framework for sport result prediction." *Applied computing and informatics* 15, no. 1 (2019): 27-33.
- [4] García-Aliaga, Abraham, Moisés Marquina, Javier Coterón, Asier Rodríguez-González, and Sergio Luengo-Sánchez. "In-game behaviour analysis of football players using machine learning techniques based on player statistics." *International Journal of Sports Science & Coaching* 16, no. 1 (2021): 148-157.
- [5] Goel, Rajesh, Jerryl Davis, Amit Bhatia, Pulkit Malhotra, Harsh Bhardwaj, Vikas Hooda, and Ankit Goel. "Dynamic cricket match outcome prediction." *Journal of Sports Analytics Preprint*: 1-12.
- [6] Kapadia, Kumash, Hussein Abdel-Jaber, Fadi Thabtah, and Wael Hadi. "Sport analytics for cricket game results using machine learning: An experimental study." *Applied Computing and Informatics* (2020).
- [7] Khetan, Aryan, Baibhav Kumar, Divy Tolani, and Harshal Patel. "Prediction of IPL Match Outcome Using Machine Learning Techniques." *arXiv preprint arXiv:2110.01395* (2021).
- [8] Kulkarni, Apurva, Aditya Vidyadhar Kamath, Aadith Menon, Prajwal Dhatwalia, and D. Rishabh. "Prediction of Player Price in IPL Auction Using Machine Learning Regression Algorithms." In *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, pp. 1-6. IEEE, 2020.
- [9] Rodrigues, Nigel, Nelson Sequeira, Stephen Rodrigues, and Varsha Shrivastava. "Cricket Squad Analysis Using Multiple Random Forest Regression." In *2019 1st International Conference on Advances in Information Technology (ICAIT)*, pp. 104-108. IEEE, 2019.
- [10] Rommers, Nikki, Roland Rössler, Evert Verhagen, Florian Vandecasteele, Steven Verstockt, Roel Vaeyens, Matthieu Lenoir, Eva D'Hondt, and Erik Witvrouw. "A machine learning approach to assess injury risk in elite youth football players." *Medicine and science in sports and exercise* 52, no. 8 (2020): 1745-1751.
- [11] Shailaja, Varagiri, Rayala Lohitha, Sreethi Musunuru, K. Deepthi Reddy, and J. Padma Priya. "Predictive Analytics of Performance of India in the Olympics using Machine Learning Algorithms." *International Journal* 8, no. 5 (2020).
- [12] Simon, Annina, and Mahima Singh. "An overview of M learning and its Ap." *International Journal of Electrical Sciences Electrical Sciences & Engineering (IJESE)* 22 (2015).
- [13] Sinha, Anurag. "Application of Machine Learning in Cricket and Predictive Analytics of IPL 2020." (2020).
- [14] Sun, Xiangyu, Jack Davis, Oliver Schulte, and Guiliang Liu. "Cracking the black box: Distilling deep sports analytics." In *Proceedings of the 26th acm sigkdd international conference on knowledge discovery & data mining*, pp. 3154-3162. 2020.
- [15] Tekade, Pallavi, Kunal Markad, Aniket Amage, and Bhagwat Natekar. "Cricket Match Outcome Prediction Using Machine Learning." *International Journal* 5, no. 7 (2020).
- [16] Tian, Changjia, Varuna De Silva, Michael Caine, and Steve Swanson. "Use of machine learning to automate the identification of basketball strategies using whole team player tracking data." *Applied Sciences* 10, no. 1 (2020): 24.

STUDENT ATTENDANCE AUTOMATION SYSTEM USING FACIAL RECOGNITION

R. Vaibhav¹, D.Sudhagar²

¹ BTech(IT), ² Associate Professor
Jerusalem College of Engineering, Chennai.

Abstract

Attendance Monitoring System is important in many places such as educational institutions, offices for checking the performance of the students and employees and it's not an easy task to check every student is present or not. In all organizations, attendance is taken manually by calling their register numbers or names and noted in attendance registers issued by the department heads as proof and in some organizations the pupils want to sign in these sheets which are stored for future references. This technique is repetitive, complex work and results in errors as few students regularly sign for their absent students or tell proxy attendance of the absent students. This method additionally makes it more complex to track all the students' attendance and difficult to monitor the individual student attendance in a big classroom atmosphere. In this article, we use the technique of utilization face detection and recognition framework to continuously recognize students going to class or not and mark their attendance by comparing their faces with a database to match and mark attendance. This facial biometric framework takes a picture of a person using a camera and contrasts that image and compares the image with the image which is stored at the time of enrolment and if it matches marks the attendance and monitors the student performance continuously. We may use the concept of artificial intelligence to monitor student attendance like capturing the motion pictures of the student when present in class to analyze the student data on how much time the student presents in class.

Keywords: *Artificial Intelligence, Student Attendance System, Face reorganization, Students attendance monitoring system and applications*

I. INTRODUCTION

The student's attendance system using artificial intelligence concept mainly works using the concept of facial recognition system was discussed by Akshara Jadhav et al. Face is considered as a primary key feature to identify and talk with other peoples in the world because the face is considered as a unique identity for every person. The facial features will be unique to the other industries. The Unique features for every indusial make facial recognition in implementing the real world. Humans distinguish a particular person's face based on several factors like colour, nose, eyes, ears, etc but for computers, it's difficult to analyze the data so we may use the concept of Computer vision. The intention of using

computer vision technology is to recognize the human features in a computer.

In recent years we observed remarkable changes in face recognition techniques because of available biometric methods, this is the most unnoticeable technique. The installation of face recognition systems on a large scale is easy but the actual implementation of the face recognition system is ambitious because it has to take into account all potential cases variation caused by a modification in face expressions by light-weight face expressions, different styles, image resolution, sensing element device, viewing distance, etc. [3] Several algorithmic rules are implemented on face recognition and every algorithm has strengths and capabilities by its own. We tend to do face recognition nearly daily. Most of the time we glance at a face and acknowledge by in a flash with the data present already in the database.

This aptitude if potential followed by machines will influence be valuable and should give a vital role in real-world applications like various access management, national and international security and defence, etc. At present mainly two approaches rely on Face reorganization methods. The first and very familiar method is the native face recognition system that depends on facial expressions of a covering for example eyes, nose, colour, etc. to identify the face with someone matching or not. The second approach is the world face recognition system which uses the whole face to identify a person. The two described approaches are enforced by methods by various types of algorithms. The recent implementation using artificial intelligence applications in face recognition attracts many scientists to research this topic. The elaboration of a face features originates from the changes continuously within the facial expression that changes over time. Apart from all these changes, we are ready to acknowledge an individual easily. The idea of developing a self-understanding and self-learning intelligent machine that may require giving sufficient data to the machine was proposed by Pradeepa .M et al.[5] A facial Recognition System can be defined in simple words as the technology that identifies a person and verifies it with the database by comparing the facial features described by Chaitanya Reddy [7].

A. Face Recognition

Scientists started working on computers to recognize human faces in the mid-1900s because of its enormous applications on face recognition has received continuous

attention from researchers. Face recognition may be outlined because of the technique of characteristics by someone based on biometrics by the approach of matching a capturing image or video with the data present in the database. The data flow process in face Recognition systems starts by having the ability to find faces and recognize frontal faces from data input devices like mobile phones, cameras, etc. [10] Practically it has been proven that students attended classes only when there is full control on classroom and attendance monitoring.

II. LITERATURE SURVEY

Patel UA, Swaminarayan Priya R. et. al[1], Development of a student attendance management system using Face recognition. This process will be used effectively in our project since we use Facial Recognition to mark the attendance. The face once recognized will detect the student and then the attendance will be marked in the CSV File.

Jacksi K. et. al[2], Design and Implementation of Attendance System, from here we can see that our main concept of marking attendance is done by using this research paper where the author stresses out to store the attendance in a CSV file rather than in a database such as MySQL or SQLite.

Gangagowri G, Muthuselvi J, Sujitha S. et. al[3], Attendance Management System using OpenCV, here when the face is detected and recognized, the process behind this is Computer Vision, which is a tool exclusively used for projects which deal with image classification and image recognition. OpenCV is the python package for Computer Vision.

Anitha V Pai, Krishna A, Kshama PM, Correa M. et. al[4], Offline service for Student Attendance Management System, the main purpose of using OpenCV is that, since the web camera needs to be opened at the time of recording attendance, this OpenCV package is where it establishes a connection with the local web camera present in the computer and then it switches on for marking the attendance.

M. Turk and A. Pentland et. al[5], Eigenfaces for recognition, Eigenfaces is a facial recognition algorithm which uses a 64 landmark point to recognize the face. This is one of the oldest facial recognition algorithms that are used even today due to its accurate results. Hence Eigenfaces can recognize the faces easily and the computational time is less when compared to any other facial recognition algorithm.

M. Alwakeel and Z. Shaaban et. al[6], Face recognition based on Haar wavelet transform and principal component analysis via Levenberg-Marquardt back propagation neural network can be used where the process of feature extraction can be done using Principal Component Analysis, where the face is divided into features which will be useful to recognize the student.

III. METHODOLOGY

A. Functional Specifications

Functional specifications are the requirements in which requires to operate a system. These requirements are necessary to assemble a system that will be required to attain the objectives embarked on previously was implemented by Jireh Robert Jam [3] . Some of the important functional and non-functional requirements are outlined below by analysing the shopkeeper's story.

- First capturing the facial image by the high-quality camera.
- HD Camera especially professional camera
- Facial features should be detected in the photo.
- Crop the overall ranges of faces detected.
- Resize all the images until the recognition system takes a photo to recognize them.
- Calculating the overall attendance percentage based on facial features matched.
- Storing all the detected face images in a folder.
- Loading the images into the database.
- We want to train facial features to recognize the computer.
- Perform recognition for faces stored on the database.
- Calculate computer facial recognition speed for effective security.
- Performing face recognition sequentially for each image cropped.
- Displaying input and output cropped images side by side on the same slot to recognize and compare the features by machine.
- After recognizing the face displays the name of the output image above the image in the given area to identify easily.

B. Non-Functional Specifications

Non-functional Specifications are the needs based on the specific criteria to evaluate the operation of the system. These requirements are collected and analysed based on the client needs and exceptions, security and working etc implementation issues by Jireh Robert Jam [3] specifically:

- The first and most important thing is that users need to find easy to take pictures.
- The system can be easily installed.
- The operator will give clear instructions on how to pose the face to train the computer.
- The face recognition system is highly secure.
- The response time if the system is very less i.e. 10 seconds.
- The face recognition system must be fast, reliable and 100% efficient.

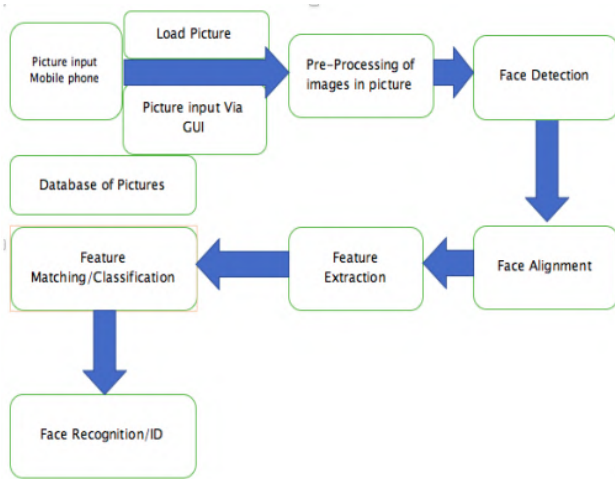


Fig-1: Flow Diagram describing the face detection

From the above use-case mentioned Face Detector detects the face inputted in the given image or video and loads the image to the system. Face localization can notice wherever the face is situated within the inputted image or video marked using the bounding boxes. The landmarks of the face like eyes, colour, nose, mouth etc are done for feature extraction from the system using face localization. The key features are extracted using face extraction to undergo tracking effectively. Facial features are matched and classified with the databases stored. Face detection offers two outputs either a positive output or negative output for the supported image from the set of stored database collection of images.

C. Working on the face recognition system

The working of the face recognition system is mainly classified into two types of algorithms [3]. They are holistic matching algorithms and feature-based algorithms. The entire face is considered as input data to identify a particular person from the database in the holistic matching method and the face is divided based on the facial features like eyes, colour, skin tone and eyebrows etc in the feature-based method. Apart from the above two, in recent days three dimensions face recognition technology is used to capture the 3D faces using sensors to identify people in the real world very accurately. Real-time face detection using 3D sensor-based applications determines a

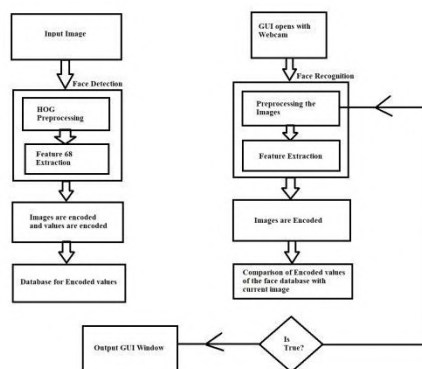


Fig-2: Architecture of Student Attendance Management System.

a person with various facial images in different angles, on different light conditions, and various poses with different expressions make more efficient in all applications.

IV. ARTIFICIAL INTELLIGENCE TO MARK ATTENDANCE

Artificial Intelligence enabled Face detection based applications to become world-famous. This was discussed by Santana Fell [6]. In Washington, an American private elementary school in Seattle begins implementing an automatic face detection technique called SAFR (Secure accurate Facial Recognition). In this method, the faces of parents are added to the database.

The security gates of the school opened only to secure accurate Facial Recognition that recognizes the right person. The working of this method is based on scanning This method works by the method of scanning the person's face and checking with the database and the person need not show their face to the scanner it will automatically detect the face and check with the database if both match the doors will open Advanced Artificial intelligence-enabled techniques are started implemented in classrooms. As books are replaced by tablets and smart board's blackboards, Smart mobiles will replace attendance registers. A girl's government high school in Chennai, Tamil Nadu, started implementing an attendance system based on artificial intelligence-enabled and this is India's first school implemented and got popular in Japan and the US. The working of that method is based on the mobile application which contains all the relevant basic details of the student with the student photo.

The faculty needs to just click the photo which automatically marks the attendance of those who present and stores the attendance in the database. A Company in china has programmed the world's 1st Artificial Intelligence teaching assistant designed to control through a whiteboard along with a commanding person. Artificial Intelligence assistant, reacting to the user through voice recognition. The main feature of automatic face recognition is to find that all students attending or not. It is also used to mark grades and use to prepare customized exercises for college activities. Citytech Software firm in Calcutta organized an in-house innovation competition based on trending technology. One group in that contest started working on Artificial Intelligence enabled cameras to require pictures of individuals coming into the workplace notice the personage, feelings, gender and alerts the concerned person about the security thread discussed by Shreyak Sawhney et al. [9] The CCTV cameras which detects the employee's face and marks the attendance of that person and calculates their salary based on how many hours they worked in the office.

V. REAL-WORLD APPLICATIONS OF FACIAL RECOGNITION

- In recent Years, Face Recognition and identification is being extensively used in security surveillance Real-time systems to Identify individuals on the spot proposed by Shreyas Iyer et al [8].
- In crime detection and forensic analysis, it plays a major role. Using drivers' licenses to identify criminals using facial recognition system is used in the US Federal Bureau of Investigation. Artificial Intelligence enabled cameras has been checked to identify those smuggling persons in the UK.
- The face identification method plays a key role in making secure payments using online payment. Online payments as it are a more trusted feature in which only the account holder can access the account.
- The personal information in mobiles can be protected in smartphones to check no one can access the personal data even the smartphone is stolen. It is being used in mobile phones for unlocking. This method is very high secured.

A. Advertising makes more responsive

Using the face identification method we can make advertisements a lot more participating and makes more personalized for various types of users. To customize the audience interests Some branded companies have already implemented automatic face identification methods in the digital world to customize their campaigns by scanning a face, based on his age and gender ads are played. Apart from that this system identifies expressions of people to understand their emotions like sad, happy, disgusted, etc. based on that displays ads for a particular product in which the user likes by understanding facial emotional features.

B. Airport security increases

Airports are the busiest place in which high chances of criminal and terrorist activities because of this reason several airlines started implementing face identification systems to check baggage and flight boarding makes the process quicker. Moreover, the Artificial Intelligence face authentication application implemented with surveillance cameras which helps in identifying a terrorist who might be involved in some disastrous activity by understanding the unusual behaviour of the person and the facial expressions of the person to recognize the criminals to make the airport place safe.

C. Diagnose rare genetic diseases

Artificial Intelligence enabled automatic face identification application will facilitate the medical business to diagnose sickness that leads to an amendment in appearances like spreading eyes or drooping ears. A face recognition scanning will become a part of standard medical checkups that

will identify genetic disorders such as Disgorge syndrome, Engelmann syndrome, Cornelia de Lange syndrome etc which bring gradual changes in facial expressions. Now clinical diagnoses for various genetic diseases and their treatment will become faster than before with face recognition in this way facial technology is also implemented in the medical field.

D. Provide driver safety and personalization

Automobile companies like Tesla, Subaru etc are increasing their services in numerous ways by utilizing face recognition systems to recognize drivers. The main use of face recognition systems is to begin the car using face recognition rather than using a key to start the car. Face recognition can scan the facial features of the driver, monitor the focus of the driver and alert the concerned person if they are losing concentration. Face identification identifies the preference of the driver like favourite stations, calendar and the position of the seat with a detailed report which increases the customer and driver safety.

E. Helpful in VIP identification

A face recognition system will establish prestigious guests whereas getting into a building or attending any event automatically identifies the person which boosts their loyalty greatly. In hospitals, the face recognition method helps in identifying a returning guest at the entrance and displays the preference of food; room etc once the face is scanned. In the event, the organizer can easily identify the VIP guest among all the fans early skips the queue and provide all VIP benefits to them.

F. Stops retail crime

In the retail business, the face recognition and identification system could be a future decider because it identifies the person instantly once the person enters it searches for a thief, criminal or person with a fraudulent history. The security officer of the particular retail shop is informed immediately when the criminal enters. When a criminal enters the store the image of the criminal is matching with the huge database of criminals for any criminal records pending. With the help of face recognition technology, retail crimes are gradually decreasing and there is no case of retail crimes shop robberies etc.

VI. ADVANTAGES

A. Improvement of Security Level

Every organization needs to secure their premises for unknown entry into that place. They also wish to monitor the employees and industrial entry into that place. Those who are entering the organization premises without proper access they

are captured in the security surveillance system and noticed to the respective person and alerts instantly concerning the person who doesn't have permission.[11]

B. Straightforward Integration method

The automatic face detection tools work effectively with the current authentication code that organizations have developed. The technique is straightforward to code the system to access organizations' automatic data processing which makes the method very clear.

C. High Accuracy Rates

The main advantage is its Accuracy. The system checks and gives the output without any misunderstanding and bad face detection system. The authorized person will be detected at the right time due to the high accuracy levels. The manual recognition, which is done by securities outside of the organization's premises may use the face recognition technology to automate the process of identification and assure its perfection without changes. We don't want an additional employee to monitor the working of cameras 24/7. The main objective of Automation means to reduce the human effects and reduce the cost of employees too. Then an organization can recognize the fact that usage of automated face identification is highly secure with accurate data.

D. Forget the Time Fraud

The massive advantage of using the automatic face recognition method is to provide a time tracking attendance system to enable avoid time fraudulence between employees. It is not possible to any colleagues to favour their friends because everyone needs to pass the entrance gate where the face recognition camera catches their photo and matches with the database and avoids time fraud among workers. This is very beneficial for the employees who work based on hourly starts check in the time starts counting from that moment until a similar check-out did which is beneficial to the organization they need not monitor their workers and the method is quick because the staff doesn't have to prove industrial identities by scanning their smart cards on the scanner. It is very difficult for the business heads to monitor all the employees are attending or not. The main problem is time fraud among the employees can be avoided using an artificial intelligence-enabled face detection system.

VII. LIMITATIONS

A. Processing & Storing

Storages play a major role in the practical world in avoiding wasting huge amounts of information and needs to store for future use. Storing HD-Videos in a very low resolution also needs a huge amount of. We are wasting a

large number of resources in processing high-quality image frames every time which is not required. We need to process all the data and want to store all the data creates a huge space. But using face recognition we can do this job in a fraction of seconds using artificial intelligence concepts. To reduce the speed of processing photo frames professional agencies will use clusters of computers. But each additional system means that appreciable information transfer via the internet creates a problem.

B. Size & Quality

To operate the face recognition system correctly and perfectly we required an advanced software system using high-quality digital cameras. The identification system takes a photo of a person or takes a screenshot from the video and starts comparing it with the actual image hence here the storage matters and affects the storage and the image to reduce the size of the image it will affect the quality of the system in recognizing the face. For example, consider CCTV and the person who is far away the camera recognizes the person and takes a photo from a long distance if the resolution of the image is decreed then we can't identify the person if we didn't decrease the quality of the image the storage size of the image is very high. To avoid bad detection of images and speed up the recognition process we need to permit the identification in face-size range. The main problem with the system is an initial investment in such a huge software system is very expensive and the processing speed will decrease due to the high quality of the image.

C. Surveillance Angle

The police identification method additionally creates problems in many ways which were answerable for the selected face capturing by the camera to register a face using software recognition system, so many angles are getting used like 45 degrees, frontal face etc. To get the transparent model for the image, we have to use the frontal part. One can fool the face recognition system by the hair on the face, spectacles etc but using the high intention image with direct angle goes with the enrolled and compared with the actual image to get accurate results. Even a person can fool the system by appearing suddenly or removing a beard or mask on the face etc. We can simply avoid the above-mentioned problems by updating the databases regular to avoid such failures; the databases must be regularly updated with all the recent changes.

VIII. RELATED WORKS

A. Fingerprint-based attendance System

This attendance system works on the moveable fingerprint device developed and can pass the device to all students to scan their thump impression on the device during

class time to mark attendance without faculty involvement. This method assures the never failure method for taking the attendance. The main issue with this method is passing the fingerprint device to all the students in class time that may distrust the students' concentration which is of no use for the students and difficult to pass the device without disturbing the class to all students to mark the attendance by putting their thumb on the device was discussed by Akshara Jadhav et al [1].

B. Radio Frequency Identification-based Attendance System

The works connected to this identification attendance system are still present in the practical society described by the author Akshara Jadhav et al. they proposed a Radio frequency identification system. The working is this type system every student needs to carry an ID card called radio frequency tag and wants to put the ID card on the card reader to read and to store the attendance of the student in database. The main problem with this technique is unauthorized access. The person using another person's id card can enter into the organization which the process is less secure and even they mark attendance of their friends by approved ID card scanning on the scanner.

C. Iris-Recognition Based Attendance System

It is a biometric system that uses Iris-Recognition management .this system uses the concept of an iris recognition system in which the iris are scanned and extracted the features and matched with the database. The main problem with this system is placing the transmission lines of the scanner in a good condition of light to scan the entire irises passing through the system. It is based on the real-time face detection system which is highly secure, reliable and fast to access but needs a lot of development in different lighting conditions as discussed by Akshara Jadhav et al[1].

XI. RESULTS AND DISCUSSIONS

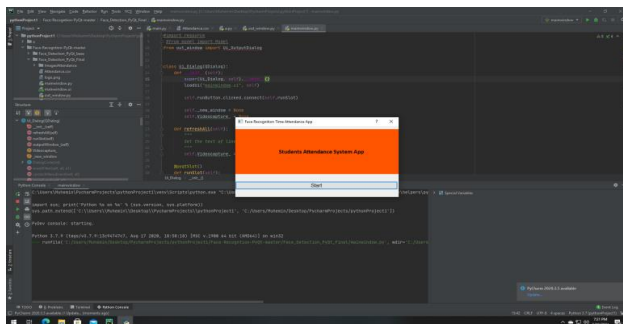


Fig-3: Starting the Application

Firstly, the program is being run and the GUI Dialog box appears which will display the start button.

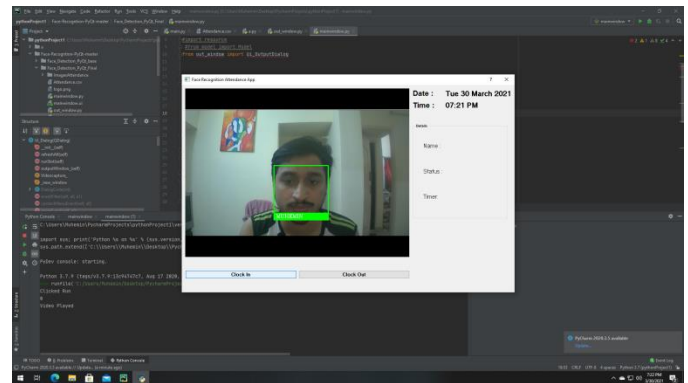


Fig-4; Face is being detected

In this window, the face is detected and recognized. After recognition, we will click the Check-In button for the validation of the face.

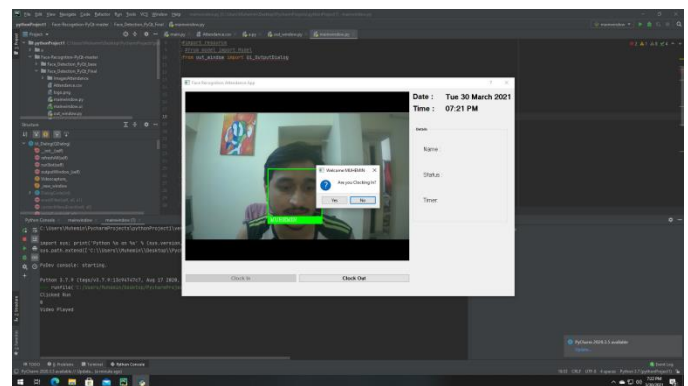


Fig-5: Verification for Clock-In

Here it verifies the user asking him where he has to check-in or check out. If we press the check-in button, our information will be displayed on the right hand side window which shows the Check-In status which includes name, date, status and timer.

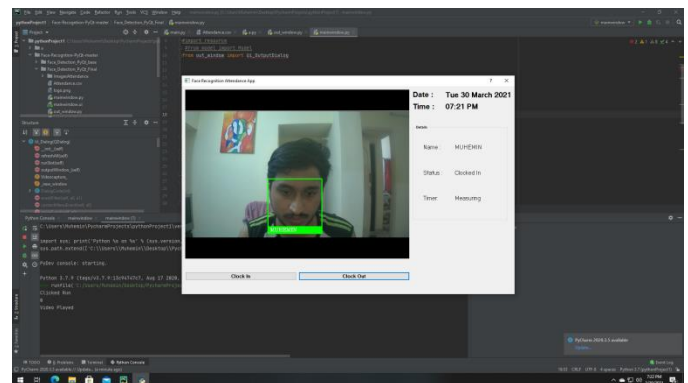


Fig-6: Measuring Process

Now, when we click the check-out button, the time which has been measured, would display on the output window and the status would show as check-out.

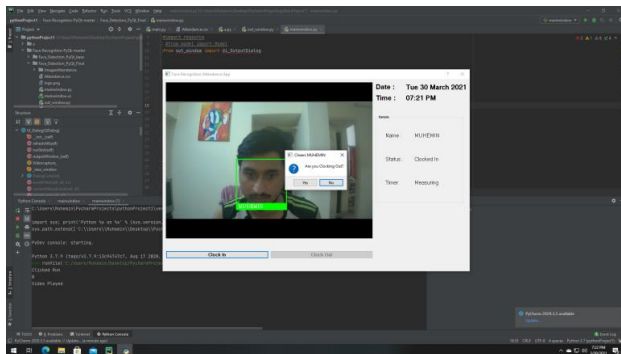


Fig-7: Check-In/Check-Out

Thus this process will be repeated for multiple people.

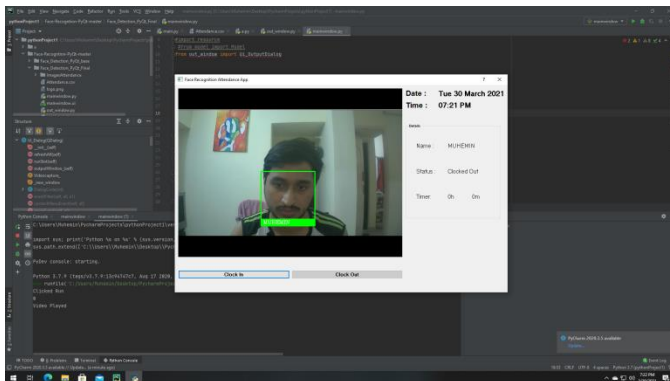


Fig-8: Check Out is pressed and the attendance is registered

Once after clicking the check-out button, the attendance will be registered in a CSV File.

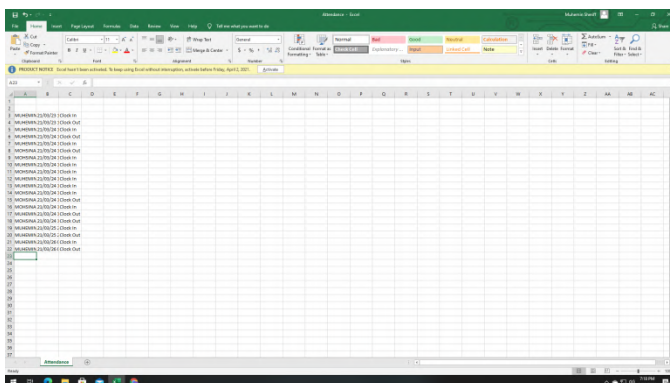


Fig-9: Output Screen (CSV File)

The attendance register is stored as data in a CSV File which contains the details such as name, date, time, check-in/check-out. This can be viewed in the excel sheet.

CONCLUSION

We are automating the attendance system to decrease the errors that occur due to manual taking attendance. If the cameras monitoring into classrooms to evaluate their interest and to mark attendance, students tend to pay attention if Artificial Intelligence enabled method can monitor and mark their attendance and faculties will at least come to school or college every day because in early times they are coming and putting sign and they are letting the school or college now it's not possible if the faculty left the college the system automatically marks as absent so everyone will come to school or organization regularly. Using the artificial intelligence concept the attendance monitoring system is very secure, accurate and easy to monitor students' and faculty's attendance.[8]

FUTURE ENHANCEMENTS

The project has a very vast scope in future. The project can be implemented on the internet in future. The project can be updated in the near future as and when the requirement for the same arises, as it is very flexible in terms of expansion. With the proposed software of database space manager ready and fully functional, the client is now able to manage and hence run the entire work in a much better, accurate and error-free manner. The following are the future scope for the project.

- Discontinuation of particular students eliminates potential attendance.
- Bar-code reader based attendance system.

REFERENCES

- [1] Patel UA, Swaminarayan Priya R. (2017) Development of a student attendance management system using Face recognition. (pp. 151-155) Charleston, SC; May 7-9.
- [2] Jacksi K. et. al[2], Design and Implementation of Attendance System AAFPRS(2017), Vol. 11, No. 3. Page 3.
- [3] Gangagowri G, Muthuselvi J, Sujitha S., Attendance Management System using OpenCV (2018). International Journal of Man-Machine Studies, 15:137-178.
- [4] Anitha V Pai, Krishna A, Kshama PM, Correa M. (2018), Offline service for student attendance management system, A.I. Memo No. 1536, C.B.C.L. Paper No. 121. MIT [5] M.

Turk and A. Pentland (2019), Eigenfaces for recognition., Eidgenössischen Technischen Hochschule, Zurich.

[6] M. Alwakeel and Z. Shaaban, (2019) Face recognition based on Haar wavelet transform and principal component analysis. M.S. Thesis. MIT.

[7] S. E. Handy, S. Lukas, and H. Margaretha (2019), et.al[7], “Further tests for face recognition using discrete cosine transform and Hidden Markov Model, ISRCJ, 15(10):1042-

1052 [8] L. Samuel., A. R. Mitra, R. I. Desanti., D. Krisnadi, (2019). Student Attendance System in Classroom Using Face Recognition Technique, CJS 5:183-187, February. [9]

S.R.Bharamagoudar, Geeta R.B., S.G.Totad (2020), Web-Based Student Information Management System. Perception. 27(10):1233-1243

[10] A. Ahmedi and S. Nandyal (2020), An Automatic Attendance System Using Image Processing. International Conference on Cyberworlds (CW) (Doi:10.1109/Cw.2017.34)

[11] Refik Samet, Muhammed Tanriverdi “Face Recognition-Based Mobile Automatic Classroom Attendance Management System” Published In Ieee 2017 International Conference On Cyberworlds (Cw) (Doi: 10.1109/Cw.2017.34)

ANALYZING DATA MINING TECHNIQUES IN EDUCATIONAL SYSTEMS

¹ Ms. Nisha Raveendran, ² Dr. N. Vijayalakshmi,

¹ Assistant Professor, Christ College (Autonomous), Kerala-680125

² Assistant Professor (Sr.G), SRM IST Ramapuram, Chennai - 89

Abstract

In the educational arena, data mining has proven to be helpful in extracting information and analyzing them to arrive at conclusions, assist in improving curriculums, predict student performances to aid in taking action for positive outcomes, and determine factors affecting enrolment, drop-out ratios, or low placement rates. The video watching patterns of students in converged educational environment were examined using data mining techniques. Since the emergence of Educational Data Mining (EDM) in the past decade, it has become increasingly important to explore the unique types of data that are generated in educational settings. The purpose of this research is to determine which characteristics are influential in determining which field of study students chooses at higher education institutions. Students' behavior, attitudes, and performance will be forecast by predictive tools and procedures in order to determine their choice of higher education. For the purpose of improving teaching and learning through educational data mining and learning analytics, this paper discusses its applications and the techniques that need to be adopted.

Keywords: *Data Mining, Education, Machine Learning, Prediction, Student Performance*

1.Introduction

Universities are interested in predicting the paths of students and alumni, so they can identify which students will participate in particular programs and which students will need a lot of debate time. Educational institutions are today faced with the challenge of using educational data to improve decision-making quality due to the explosion in educational data. The goal of data mining is to extract meaningful knowledge from these large data sets by using analytical tools. Various systems are used to meet these challenges, including ERP, DWH, etc.

Researchers have developed data mining techniques that allow them to reduce manpower, duplicate human intelligence, and perform more efficiently than a human could. Through data mining and statistical methods,

a huge amount of data and experiences are recorded, and the techniques must be able to learn autonomously. In addition to this, data mining processes can be used to integrate large amounts of unrelated data, uncover useful correlations, and obtain valuable information from the data. There are five types of data mining techniques: statistics, classification, clustering, regression, and association (Chou and Lee, 2020; Lee et al., 2018; Romero and Ventura, 2010). According to Lee et al. (2018), they were able to predict programmers' levels of tasks and expertise using user profile data. Based on the results, user profile data can provide insights into a programmer's ability to perform easy or difficult tasks. Thus, data mining techniques are used to analyze the generated systems based on data from behavior. As an example, watching multimedia videos to gather effective information and provide information to multimedia designers and developers to improve future Web multimedia systems.

1.1 Enterprise Resource Planning (ERP)

Enterprise resource planning (ERP) integrates internal and external management information of an entire organization like finance/accounting information, manufacturing, sales, service information and most important its customer relationship management. ERP systems automate all these activities to facilitate the flow of information between all business functions inside the boundaries of the organization and manage the connections to outside world [1]. ERP systems operate in real time i.e. without relying on periodic updates [2, 3] along with common database, which supports all applications. ERP plays a crucial role in decision making for businessmen to get quick decisions with fewer errors.

Today, education is one of India's most important industries. It is bigger than both the software industry and the automobile industry combined. Educational institutes have become complex organizations as a result of exponential growth in the educational sector. They are no longer limited to deliver education only, but to manage a large range of activities like marketing of institutes for student admission and corporate student's for placements,

managing internal operations like smooth conducting of classes or recruitment and motivation of human resources like faculty and staff, financial and cash flow planning, co-ordination with regulatory and statutory authorities. As a result of stiff competition and demanding customers (students and corporate), institutes are also subject to the vagaries of market forces. To manage their internal and external operations, educational institutions need modern management practices and the latest technology. As a result, the software industry began developing automated solutions to support educational institutions.

2. Data Mining Techniques

In this section, we will review the main data mining techniques [9] that are used to analyze data.

2.1 Clustering Analysis

Clustering is the process of grouping objects so that they are similar (in some sense or another) to each other than to items in other clusters. Many fields use this approach, such as machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. The clustering or grouping of objects is based on the principle of maximizing intra-class similarity and minimizing inter-class similarity. It is possible to consider each cluster as a class of objects from which rules can be derived. Through the use of clustering in education, institutes can classify students according to similar behavior. Cluster students according to their abilities, so that students in the same cluster (e.g. Average) are similar to each other but different from students in other clusters (e.g. Intelligent, Weak).

- A choice of an appropriate algorithm for a particular problem involves many decisions. Some of them are
- Connectivity models: These are based on distance connectivity.
- Centroid models: This algorithm represents each cluster by a single mean vector.
- Distribution models: Clusters are modeled using statistic distributions, such as multivariate normal distributions used by the Expectation-maximization algorithm.
- Density models: It defines clusters as connected dense regions in the data space.
- Subspace models: Clusters are modeled based upon both cluster members and relevant attributes.

2.2 Decision Tree

In a decision support system, a decision tree is a tree-like diagram that represents decisions and their probable outcomes, including resource costs and utility. It is a visual representation of an algorithm. In operations research, decision trees are used to identify the best strategy to achieve a goal, specifically in decision analysis. Decision trees can also be used as a means to calculate conditional probabilities. Figure 3 shows how decision trees can be used to analyze an institute's admission criteria. Even small data sets can be analyzed with decision trees since they are simple to understand and interpret. Data that include categorical variables with a different number of levels may not be suitable for this approach.

2.3 Factor Analysis

The factor analysis procedure is a statistical approach for describing the correlation between observed variables in terms of a lower number of unobserved, uncorrelated variables.

Factor analysis looks for such joint variations in response to latent variables that are unobserved. A linear combination of the potential factors with "error" terms is used to model the observed variables. A dataset can be reduced by using the information obtained about the interdependencies between observed variables. Originating with psychometrics, factor analysis has been applied to behavioral sciences, social sciences, marketing, product management, operations research, and other fields of applied sciences dealing with large quantities of data. It can be of two types, Exploratory factor analysis (EFA) and Confirmatory factor analysis (CFA).

Exploratory factor analysis (EFA): is used to uncover the underlying structure of a relatively large set of variables. A priori, the researcher assumes that any indicator may be related to any factor. Factor analyses are often based on this assumption.

Confirmatory factor analysis (CFA): determines if the number of factors and loadings of measured (indicator) variables are in accordance with expectations based on theory. An analysis of factors is done to determine whether the indicator variables load the way they should based on prior theory.

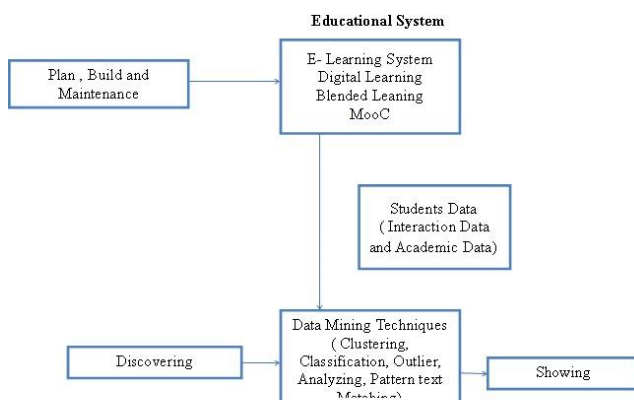
2.4 Regression Analysis

The technique of regression analysis includes modeling and analyzing several variables when a dependent variable and one or more independent variables are considered. In particular, regression analysis enables us to investigate how the dependent variable's standard value

changes when any one of the independent variables is varied, whereas the other independent variables remain constant. Regression analysis is commonly used to estimate the conditional expectation of the dependent variable with respect to independent variables. The estimation target in such cases is a function of the independent variables referred to as a regression function. In regression analysis, it is also important to understand how the dependent variable varies around the regression function, which can be modelled by a probability distribution. Forecasting and prediction are carried out using regression analysis, which explores the relationship between independent and dependent variables. In most cases, linear regression and ordinary least squares regression are used. The most commonly used regression methods are linear regression and ordinary least squares regression.

3. Incorporating data mining into Education

Education institutes can use clustering to group students based on similar behaviors. Organize the students into groups, so that students in one group (e.g. Average) can be compared to each other while they are dissimilar to students in another group (e.g. Intelligent, Weak).



Picture 1: Incorporating data mining techniques in Education System

By applying data mining techniques, traditional classrooms and web-based educational systems can extract knowledge, helping educators and students to make informed decisions.

4. Data mining techniques for educational tasks

Data Mining has been used to resolve many applications or tasks in educational environments. According to Baker, Educational Data Mining could be used in four ways: Improving student models Improving domain models. Assessing how learning software supports pedagogical practice. Studying how learners learn.

And five approaches/methods:

Prediction, Clustering, Relationship mining, Distillation of data for human judgement, Discovery with models.

4.1 Predicting Student Performance

By identifying the dropouts and students who need special attention sooner, the teacher can provide appropriate advising/counseling. The purpose of this paper is to examine students' performance in courses through the use of data mining methodologies. Data mining provides a variety of tasks that can be used to analyze student performance. We extract knowledge about students' performances in end-of-semester exams through this task. In the classification task, students are evaluated based on their performance. As many approaches can be used for data classification, here we use the decision tree method. In order to predict performance at the end of the semester, student management system data such as attendance, tests, seminars, and assignments were collected.

5. Conclusion

In this study, we present various data mining techniques for supporting education systems by generating strategic information. Because data mining offers a number of advantages to higher education institutions, it is recommended applying these techniques in the areas of resource optimization and student performance prediction. Using a student database, the classification task is used to predict the students' division based on previous student data. The decision tree method is used here because there are many approaches to data classification. In order to predict the student's performance at the end of the semester, attendance, class test, seminar and assignment marks were collected from the student's previous database.

References

- [1] Arcinas, Myla M., et al. "Role of Data Mining in Education for Improving Students Performance for Social Change." *Turkish Journal of Physiotherapy and Rehabilitation* 32.3 (2021): 204-226.
- [2] Su, Yu-Sheng, and Sheng-Yi Wu. "Applying data mining techniques to explore user behaviors and watching video patterns in converged IT environments." *Journal of Ambient Intelligence and Humanized Computing* (2021): 1-8.
- [3] Govindarajan, M. "Educational Data Mining Techniques and Applications." *Advancing the Power of*

Learning Analytics and Big Data in Education. IGI Global, 2021. 234-251.

[4] C. Romero, S. Ventura "Educational data Mining: A Survey from 1995 to 2005", Expert Systems with Applications (33), pp. 135-146, 2007

[5] Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar, M. Inayat Khan, "Data Mining Model for Higher Education System", European Journal of Scientific Research, Vol.43, No.1, pp.24-29, 2010

[6] K. H. Rashan, Anushka Peiris, "Data Mining Applications in the Education Sector", MSIT, Carnegie Mellon University, retrieved on 28/01/2011

[7] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann, 2000.

[8] Kumar, V. (2011). An Empirical Study of the Applications of Data Mining Techniques in Higher Education. IJACSA - International Journal of Advanced Computer Science and Applications, 2(3), 80-84. Retrieved from <http://ijacsa.thesai.org>.

[9] Manoj Bala, Dr.D.Bojha , "Study of Applications of Data Mining Techniques in Education". IJRST – International Journal of Research in Science And Technology.

GENERATIVE ADVERSARIAL NETWORKS BASED APPROACH ON REAL WORLD :THEORY AND APPLICATIONS

1st Ashly Ann Jo, 2nd Ebin Deni Raj

ashlyannjo.phd2112@iitkottayam.ac.in,ebindeniraj@iitkottayam.ac.in

Abstract—Generative adversarial network (GANs) is one of the most important research topics in AI, and their excellent data production capacity has attracted a lot of interest in recent years. For computer vision and healthcare field researchers, this paper provides an overview of GANs and its application. The structure and principle operation of GANs, as well as the basic GAN models proposed to date and the theory of GANs, were reviewed. In this paper also discussed about the development of a GAN selection methodology for the GAN applications. Also discuss about the GANs progress in image synthesis tasks, such as image-to-image translation, image super-resolution and text-to-image synthesis, sequential data generation and GANs in healthcare. The limitations of GANs and the solution were present in this paper.

Index Terms—Generative adversarial Networks, Computer Vision, Image Synthesis

I. INTRODUCTION

Deep generative models are neural networks with many hidden layers trained to solve complex problems which required large number of samples for training. These models have recently drawn significant attention in capturing the sequence data such as audio or video and synthesize new samples. The generative models include variational autoencoders (VAE) and Autoregressive models have used Markov chain Monte Carlo based algorithms. But sampling from the Markov Chain in High dimensional data are computationally slow and inaccurate. To address this problem Goodfellow, et, al [1] proposed Generative Adversarial Networks (GANs). GANs are a type of generative models that can generate a random variable with respect to a specific probability distribution. GANs can create images that look like photographs of human faces, even though the faces don't belong to any real person. The rest of the paper is organized as follows: Section II gives a brief introduction of GANs. Section III introduces GAN variants. Section IV focuses on GAN Selection. Section V discusses several applications of GAN. Section VI reviews the current challenges and limitations of GAN-based methods. Conclusions are given in Section VIII.

II. GENERATIVE ADVERSARIAL NETWORKS(GANS)

A. GANs Architecture

GANs are trained using two neural network models. One model is called the generator or generative network model that learns to generate a fake sample. The other model is called the discriminator or discriminative network and learns to differentiate generated examples from real examples. The

generator output is connected directly to the discriminator input. Through backpropagation, the discriminator's classification provides a signal that the generator uses to update its weights. The discriminator in a GAN is simply a classifier. It tries to distinguish real data from the data created by the generator. It could use any network architecture appropriate to the type of data it's classifying. GAN's are just like a game where discriminator and generator compete with each other. The generator tries to fool the discriminator while discriminator tries not to get fooled by the generator. The training continues until the generator wins the adversarial game. The whole process can be regarded as a two-player min-max game where the main aim of GAN training is to achieve the Nash equilibrium.

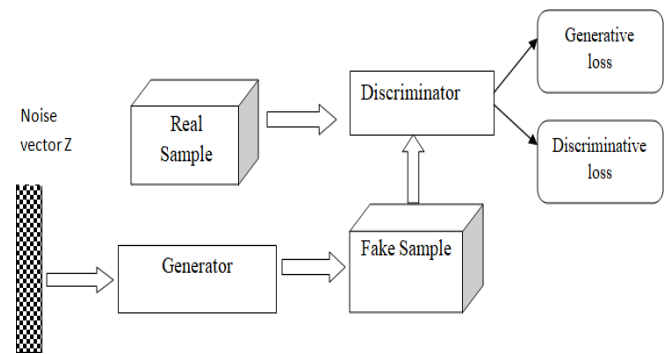


Fig. 1. GAN Framework

B. Objective Function

A GAN can have two loss functions: one for generator training and one for discriminator training. The loss function of GAN is based on the minimax game which includes two neural networks competing with each other in the framework. The discriminator needs to check whether the input sample is real or fake and optimize the weights of the network model by the backpropagation algorithm. During training the generator and the discriminator do not change the model parameters. GAN will not stop training until the two networks reach the Nash equilibrium. The loss function of the GAN is formulated as follows:

$$\min_G \max_D V(D, G) = E_{x \sim P_{data}(x)} [\log D(x)] + E_{z \sim P_z(z)} [\log (1 - D(G(z)))]$$

where $P_{data}(x)$ denotes the true data distribution, $P_z(z)$ denote the noise distribution. The selection of loss function is the most important problem for the GAN model. The loss function has a direct influence on the gradient disappearance and the model collapse.

C. Evaluation metrics

GAN generator evaluation refers to the calculation of specific numerical scores used to summarize the quality of generated images. The basic GAN model used the Average Log-likelihood method, also referred to as kernel estimation or Parzen density estimation, to summarize the quality of the generated images. The Inception Score and the Fréchet Inception Distance are two extensively used metrics for analyzing created images.

- Inception Score (IS), which captures both the quality and diversity of the generated images. A higher score indicates better-quality generated images.
- Fréchet Inception Distance (FID) which compares the real vs. fake images and doesn't just evaluate the generated images in isolation. A lower FID score indicates more realistic images that match the statistical properties of real images.

III. GAN VARIANTS

Since the original Generative Adversarial Networks paper there has been many GAN variants. They tend to build upon each other, either to solve a particular training issue or to create new GANs architectures for the better result.

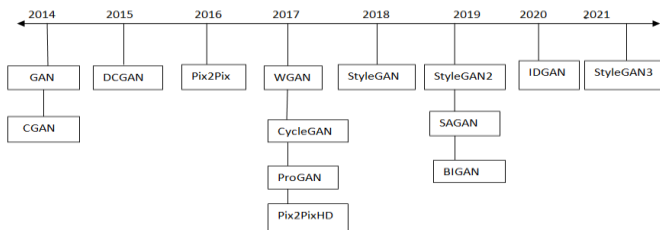


Fig. 2. Timeline of architecture-variant GANs presented in this article

A. Deep Convolutional Generative Adversarial Networks (DCGAN)

Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks [2] was the first GAN proposal using Convolutional Neural Network (CNN). Most of the GAN variations existing nowadays are built upon the DCGAN. DCGAN uses convolutional layers in the generator and convolutional-transpose layers in the discriminator. It replaces the generator and discriminator in classic GAN with two CNN without changing the basic structure of the GAN, a step size convolution is used instead of the upsampling

layer, and a convolutional layer are used to replace the full connection layer to increase the stability of the training. The discriminator consists of strided convolution layers, batch normalization layers, and LeakyRelu as activation function. It takes a $3 \times 64 \times 64$ input image. The generator consists of convolutional-transpose layers, batch normalization layers, and ReLU activations. The output will be a $3 \times 64 \times 64$ RGB image.

B. Wasserstein GAN (WGAN)

WGAN [3] and WGAN-GP [4] were created to solve mode collapse. It occurs when the generator produces the same images or a small subset of the training image repeatedly. Instead of using a discriminator to classify or predict the probability of generated images as being real or fake, the WGAN replaces the discriminator model with a critic that scores the realness or fakeness of a given image. WGAN-GP improves upon WGAN by using gradient penalty instead of weight clipping for training stability.

C. Conditional Generative Adversarial Nets (CGANs)

CGANs [5] first introduced the concept of generating images based on a condition, which could be an image class label. Pix2Pix [7] and CycleGAN [8] are both conditional GANs, using images as conditions for image-to-image translation. The conditional GAN adds a condition to the generation and discrimination process of GANs. cGANs are not strictly unsupervised learning algorithms because they require labeled data as input to the additional layer.

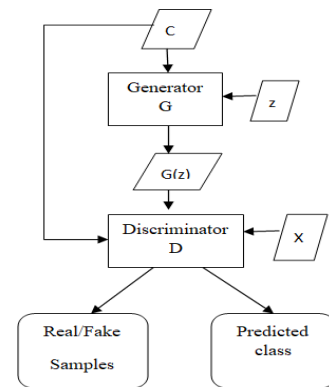


Fig. 3. Architecture of Conditional GAN

D. Pix2PixHD

High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs [6] disentangles the effects of multiple input conditions. It consists of coarse-to-fine generator and multi-scale discriminator. In addition, it can generate realistic 2k high-resolution images. It also generates multi-modals.

E. Self-Attention Generative Adversarial Networks(SAGAN)

SAGAN [9] improves image synthesis quality. It generates details using cues from all feature locations by applying the self-attention module. SAGAN is a convolutional GAN that uses a self-attention layer in the generator model, and spectral normalization on both the generator and discriminator. It trained by using the two time-scale update rule (TTUR) and the adversarial loss.

F. BigGAN

Large Scale GAN Training for High Fidelity Natural Image Synthesis[10] can create high-resolution and high-fidelity image. BigGAN design is based on SAGAN and to improve the training by increasing the number of channels for each layer and the batch size. The BigGAN uses model architecture with attention modules from SAGAN and is trained via hinge loss. The class information is provided to the generator model via class-conditional batch normalization. The BigGAN slightly modify and updates the discriminator model twice before updating the generator model in each training iteration. This method achieves the goal of generating high resolution and diverse samples from the complex dataset ImageNet successfully. It is the largest scale of Generative Adversarial Networks that have been trained so far and can generate images of high quality. BigGAN make use of orthogonal regularization to the generator to handle the specific instability .

G. ProGAN, StyleGAN StyleGAN2

Progressive Growing GAN(ProGAN) is an extension to the GAN that allows generation of large high-quality images, such as realistic faces by efficient training of generator model[11].In ProGAN Generator network starts with less Convolution layers

to output low-resolution images and then increments layers to high resolution images .Similarly Discriminator network follows same approach, starts with smaller network taking the low-resolution images and outputs the probability. It then expands its network to intake the high-resolution images from generator and classify them as real or fake. This incremental expansion of both G and D networks allows the models to effectively learn high level details first and later focus on understanding the fine features in high-resolution 1024×1024 pixel images. It also improves model stability and lowering the probability of mode collapse. A Style-Based Generator Architecture for Generative Adversarial Networks known as StyleGAN was developed to have control over the style of generated images[12].StyleGAN is an extension to the basic ProGAN architecture, with the ability to control over the disentangled style properties of the generated images .It controls the visual features by modifying the input of each level in the network separately, from coarse features to fine details. StyleGAN not only produces high-quality and realistic images but also provides better control and understanding of the generated images. But sometimes unnatural parts are generated in images such Droplet like artifacts and phase artifacts. StyleGAN2 [13] improves the original StyleGAN by

Configuration	
StyleGAN	StyleGAN2
Base ProGAN[11]	Base StyleGAN[12]
Tuning	Weight demodulation
Add mapping and styles	Lazy regularization
Remove traditional Input	Path length regularization
Add noise inputs	No growing in each level
Mixing Regularization	Large network

TABLE 1

STYLEGAN AND STYLEGAN2 CONFIGURATION

making several improvements in areas such as normalization, progressively growing and regularization techniques.

H. SmileGAN

Smile-GAN is a semi-supervised clustering method[14] which is designed to identify disease-related clustering . Semi-supervised clustering of Smile-GAN is achieved through joint training of the mapping and clustering function.

I. StyleGAN3

StyleGAN3 [15]is the improvement of StyleGAN2[13] and it overcome all the limitations of the latter.StyleGAN3 is improved with Alias-free generator architecture and training configurations. Also Introduced tools for interactive visualization, spectral analysis and video generation. This advancement will be very useful in the improvements of models that generate video and animation. The new generator is more computationally complex than StyleGAN2 but still have high FID. The major improvements of StyleGAN3 over StyleGAN2 are reduced memory usage, faster training and error fixing.

IV. GAN SELECTION

In this paper proposed a systematic literature review approach for the selection of GAN to solve a particular problem.First need to choose a certain database of the selection. Then develop search using keywords to identify the general application in each paper. The possible keywords are "GAN models", OR "GAN variants", OR "GAN applications",OR "Image Synthesis using GAN" AND GAN in Healthcare". The search result is used to classify each paper into each class. This will be much helpful in the future process. Then choose the class of the specific application for the further process. If the search is for a particular application then choose the class of the specif application, among that papers filter the paper the based on year. Then remove the duplicate entries in the records. After removal check whether the remaining records meet the objective.

The Original papers, reviews and survey are considered for the selection process. These papers were characterized as three criteria such as to check whether the selected present the certain application-based publication, discuss GAN progress in real-world and check whether the paper accurately address the research question and objectives. The records that does not meet the objective, paper that is not proper language and studies with no precise knowledge about how these models are evaluated and which industry is connected with the specific application in GAN.The detailed flowchart of the GANs selection process is given in fig4.

Criteria	Learning	Architecture	Improvements	Loss Functions	Performance Metrics
DCGAN	Unsupervised	Convolutional Networks	CNN structures, batch normalization	Binary loss	Accuracy and error rate, FID
WGAN	Unsupervised	Basic GAN	Increased stability of optimization	Wasserstein Loss	Kernel distance(KID)
CGAN	Supervised	Multilayer Perceptron	Sample generation based on a condition	Standard GAN loss	FID
Pix2PixHD	Supervised	Dual Generator and multi-scale Discriminator	synthesizing portraits from face label maps	Feature Matching Loss	FID
SAGAN	Unsupervised	Self Utilization layers	Self attention	Hinge Loss	FID
BigGAN	Supervised	Self attention module	Deeper net and large batch size	Hinge Loss	FID
ProGAN	Unsupervised	Multi-layer Perceptron	Progressive growing during training	Wasserstein loss	FID, IS
StyleGAN 3	Unsupervised	Mapping Network and Synthesis network	Memory Reduction, Faster training	Non-saturating logistic loss	FID
SmileGAN	Semi-Supervised	Continuous Mapping and Clustering	Clustering Learning	$L_{adv}, L_{reg}, L_{cluster}$	IS, FID

TABLE III
COMPARATIVE STUDY OF DIFFERENT GAN VARIANTS

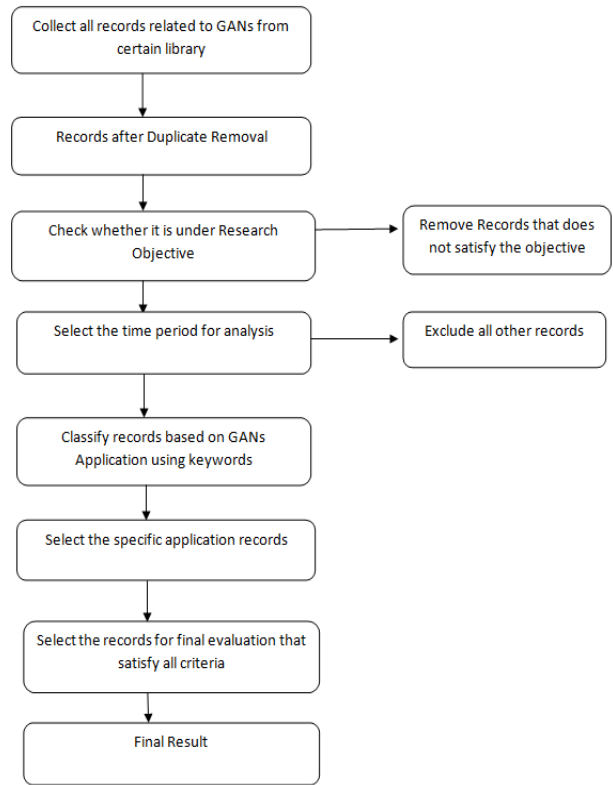


Fig. 4. Flowchart of GAN Selection

V. GANs APPLICATIONS

A. Image Generation

Ian Goodfellow, et al.[1] proposed Generative Adversarial Networks in which GANs are used to generate new examples for the MNIST handwritten digit dataset, the CIFAR-10 small object photograph dataset, and the Toronto Face Database. The improvement of GANs leads to the introduction of Unsupervised Representation learning with Deep Convolutional Generative Adversarial Networks(DCGAN).

Tero Karras et al demonstrate the generation of realistic photographs of human faces using Progressive Growing of GAN[11]. The human face generations was trained on celebrity dataset. This method is also used to demonstrate the generation of objects and scenes. In Progressive Growing GAN, the generator and discriminator model expanded by adding layers during the training process. It takes input as a low-resolution image and generates a high resolution image of desired scaling factor. Andrew Brock, et al. demonstrate the generation of synthetic photographs with the technique BigGAN[10] that are practically indistinguishable from real photographs. BigGAN is a type of generative adversarial network that was designed for scaling generation to high-resolution and high-fidelity images. BigGAN scale up the model size and batch size to generate larger and quality images. It is specifically designed for class-conditional image generation. CIFAR

or ImageNet datasets are mainly used to train class-conditional GANs. The base architecture of this model is the Self-Attention GAN (SAGAN) described by Han Zhang, et al.[9] The SAGAN achieved best results from ImageNet dataset.

Karras et.al[12] proposed new generator architecture for generative adversarial networks, borrowing from existing style transfer methodologies. Style GAN proposes a lot of changes in the generator part which allows it to generate the photo-realistic high-quality images as well as modify some part of the generator part. This architecture change the attributes such as style of human faces as there are many features that can be altered in the generation of various images such pose, hair color, eyes color etc. A similar approach is used by Huang et al.[16] with GANs operating on intermediate representations rather than lower resolution images. LAPGAN also extended the conditional version of the GAN model where both G and D networks receive additional label information as input. This technique has proved useful and is now a common practice to improve image quality. This idea of GAN conditioning was later extended to incorporate natural language.

B. GANs in Healthcare

Various GAN architectures have been proposed for medical applications, which provided promising results in generating realistic looking images. In one study compare the traditional data augmentation with synthetic data augmentation utilizing DCGAN for a liver lesion classification task [17]. It demonstrated that the use of synthetic samples significantly improves classification performance even on a small dataset consisting of computed tomography images of 182 liver lesions [16]. While unconditional GAN architectures such as DCGAN address the instability problem of GANs, they do not work well at relatively low resolutions. Qin et.al. exploited progressive growing of GANs (ProGAN) to synthesize skin lesion images at high resolution[18], which produced highly realistic synthetic images that expert dermatologists had difficulty distinguishing them from real images.

GANs can be used in medical tumor detection[19]. The neural network can be used to identify tumors by comparing images with a dataset of images of healthy organs. The neural network can detect anomalies in the patient's scans and images by identifying differences when comparing them to the dataset images. Using generative adversarial networks results in faster and accurate detection of cancerous tumors. It helps save costs for patients as well as doctors. However, most importantly, generative adversarial networks can potentially help save human lives.

Another area in the healthcare domain where generative adversarial networks can assist is drug discovery[20]. The networks can be used for generating molecular structures for medicines that can be utilized in targeting and curing diseases. Train the generator with the existing database to find new compounds that can potentially be used to treat new diseases. This helps to search the entire database for compounds that can help fight new diseases. The algorithm automatically identifies such compounds and helps reduce the time required for research and

development of such drugs[21]. As the data scarcity remains a major obstacle for medical imaging, GANs are likely to become a standard practice to solve this problem

C. Image to Image Translation

Image to Image translation problem is mostly occurred problem in Computer vision and Image processing. Conditional adversarial networks[5] are the most suitable solution for translating an input image into an output image.

The pix2pix

[7] model learns the mapping from input image to output image and constructs a loss function to train this mapping. CycleGAN extended this work by introducing a cycle consistency loss that attempts to preserve the original image after a cycle of translation and reverse translation[8]. In this formulation, matching pairs of images are no longer needed for training. This makes data preparation much simpler, and opens the technique to a larger family of applications. For example, Translation from photograph to artistic painting style, translation of photograph from summer to winter and Translation of satellite photograph to Google Maps view. Huang et al. [22] proposed a Multimodal Unsupervised Image-to-image Translation (MUNIT) framework. In this framework the image representation is decomposed into a content code that is domain-invariant, and a style code that captures domain-specific properties. To translate an image to another domain, combine its content code with a random style code. Image-to-Image translation is similar to style-transfer where input receives the style and content image. It is not only transferring the image styles but also style of the style image Yu et al. proposed a novel method, SingleGAN[23], to perform multi-domain image-to-image translations with a single generator. SingleGAN uses domain code to explicitly control the different generative tasks and integrate multiple optimization goals to ensure the translation.

D. Text to Image

Han Zhang, et al. proposed StackGAN[24] to generate realistic looking photographs from textual descriptions of simple objects like birds and flowers. Fedus et al. [25] introduced MASKGAN to improve sample quality which explicitly trains the generator to produce high quality samples. MASKGAN uses actor-critic conditional GAN that fills the missing text on the context. Automatic synthesis of realistic images from text would be interesting and useful. Denton et al. used a Laplacian pyramid of adversarial generator and discriminators[16] to synthesize images at multiple resolutions. Radford et al. used a standard convolutional decoder[2], but developed a highly effective and stable architecture incorporating batch normalization to achieve striking image synthesis results. Reed et al. used GAN architecture to synthesize images from text descriptions[26], which one might describe as reverse captioning.

E. Anime Character Generation

Yanghua Jin, et al. [27] demonstrates the training and use of a GAN for generating faces of anime characters. They applied

GAN for creating automatic anime characters by combining a clean dataset and several GAN training. Inspired by the anime examples, a number of people have tried to generate Pokemon characters, such as the pokeGAN[28] and the Generate Pokemon with DCGAN, with limited success. hen et al. proposed a solution to transforming photos of real-world scenes into cartoon style images. The proposed solution, CartoonGAN[29], a framework for cartoon stylizing takes unpaired photos and cartoon images for training.

The video game industry can benefit hugely from generative adversarial networks. Developers and designers will have their work cut short[30]. GANs can be used to automatically generate 3D models required in video games, animated movies, or cartoons. The network can create new 3D models based on the existing dataset of 2D images provided. The neural network can analyze the 2D photos to recreate the 3D models of the same in a short period of time. For example, 3D objects such as tables, chairs, cars, and guns can be generated by providing 2D images of these objects to the neural network. This will significantly help animators save time and cost. Another exciting application of the generative adversarial network is creating emojis from human photographs[31]. The neural network analyzes facial features to create a cartoonish version of individuals. Major technology companies such as Apple support the technology to generate custom emojis similar to an individual's facial features.

F. Human Pose Estimation

Human pose estimation is the process of estimating the pose of the body from a single image. Ian et. al. proposed CR-GAN to address the problem of human pose estimation[32]. In addition to the single reconstruction path, introduced a two-pathway framework. The two-pathway framework makes it possible to combine both labeled and unlabeled data for self-supervised learning.

The issue of pose variation in person images has also been addressed by Ge et al. A new framework Known as a Feature Distilling Generative Adversarial Network (FD-GAN)[33] is introduced to learn identity-related and pose-unrelated representations. After learning pose-unrelated person features with pose guidance, no auxiliary pose information and additional computational cost are required during testing. The proposed FD-GAN obtained better performance on three-person reidentification datasets.

G. Image inpainting

Image inpainting refers to the technique of restoring and reconstructing images based on background information. High-quality image inpainting required the interpretation of the generated content to be reasonable but also required the texture of generated image clear. Dolhansky et al. [37] proposed an image inpainting approach called Exemplar GANs (ExGANs). They used exemplar information as a reference image of the region to inpaint a person with closed eyes in a natural picture which can produce high-quality inpainting results.

H. DeepFakes

GANs can be used for a number of exciting things but what has caught the public's imagination is the use of GANs to create deepFakes[35], i.e. to create videos of talking people where the face has been swapped with someone else. GANs can be used to create photos of imaginary fashion models. Facial manipulation is usually conducted with DeepFakes and can be categorized as face synthesis, face swap and facial attributes and expression. Facial attributes and expression manipulation consist of modifying attributes of the face such as the color of the hair or the skin, the age, the gender, and the expression of the face by making it happy, sad, or angry. One of the best performing methods is StarGAN [33] that uses a single model trained across multiple attributes' domains instead of training multiple generators for every domain.

I. Audio generation

Although GANs are best known for their use in image generation, they have also been used successfully for generating sequential data such as audio, natural language, and time-series. Audio generation can be applicable in specific domains such as speech synthesis and music generation. On the other hand, GAN-based models for speech synthesis have been shown to work much more efficiently. GANs hold great potential for data augmentation in speech recognition models. Moreover, the methods used in speech synthesis can be generalized to any form of audio. Music generation is another application area in which GANs are used for producing new music[38]. Different representations of sound can be more desirable in some applications of audio generation. In WaveGAN [40] and SpecGAN[41], which uses raw waveform of audio and spectrogram respectively. When comparing these two representations, they obtained promising results in both approaches.

J. Video Generation

The generative adversarial network (GAN) framework has emerged as a powerful tool for various image and video synthesis tasks, allowing the synthesis of visual content in an unconditional manner. DVD-GAN (dual video discriminator)[42] for video generation on large-scale datasets can produce videos at resolutions up to 256×256 and lengths up to 48 frames. DVD-GAN is able to generate longer and higher resolution videos based on the BigGAN architecture while introducing scalable, video specific generator and discriminator architecture. Carl Vondrick, et al.[43] describes the use of GANs for video prediction, specifically predicting up to a second of video frames with success, mainly for static elements of the scene. DCVGAN[44] produces better video samples than ones by conventional method in terms of both variety and quality when evaluating on facial expression and hand gestured datasets. The below table IV shows the summary of different GAN applications.

Application	Model	Usecases
Image Generation	DCGAN PGAN BigGAN SAGAN StyleGAN LAPGAN	High Resolution Image, Sample Dataset
GANs in Healthcare	DCGAN PGAN QUGAN	Medical Image Synthesis, liver lesion Classification, skin lesion images, Tumor detection, Drug Discovery
Image to Image	CGAN, Pix2Pix CycleGAN MUNIT StyleGAN SingleGAN	Photo to Painting, Satellite images to Google map view
Text to Image	StackGAN MASKGAN LAPGAN	Image Synthesis
Anime Character Generation	cartoonGAN PokeGAN	Cartoon style Images, Video Games
Human Pose Estimation	FD-GAN	Re-identification
Image Inpainting	ExGANs	Image Reconstruction
DeepFakes	StarGAN	Face Modulation Entertainment
Audio Generation	WaveGAN SpecGAN	Audio Synthesis, Music Generation
Video Generation	DVDGAN MoCoGAN DCVGAN	Video Generation, Prediction

TABLE IV
SUMMARY OF DIFFERENT GAN
APPLICATIONS

VI. GAN LIMITATIONS

1) *Mode Collapse*: Each iteration of generator over-optimizes for a particular discriminator, and the discriminator never manages to learn its way out of the trap. As a result the generators rotate through a small set of output types. This form of GAN failure is called mode collapse. There are certain methods to rectify the failure. Arjovsky et al. [3] proposed to compare distributions based on a Wasserstein distance rather than a KL-based divergence in DCGAN. Melz et. al [36] proposed Unrolled GANs that use a generator loss function that incorporates not only the current discriminator's classifications, but also the outputs of future discriminator versions. So the generator can't over-optimize for a single discriminator.

2) *Attaining Nash Equilibrium*: The discriminator and generator fight in an adversarial manner to get the better of each other. At any given step, either the generator is improved (gradient descent) keeping the discriminator constant or the discriminator learns while the generator stays the same. This concurrent learning of the two new neural networks does not guarantee a convergence to Nash Equilibrium. Saliman et al. [38] propose feature matching that prevents the generator from over-training for any given discriminator using statistics of real data distribution where the discriminator specifies the statistics worth matching. However, GAN architecture has some limitations that affect our society. The images created by GAN look misleadingly

like a photograph of a real person based on the analysis of portraits. Different concern by the people has been raised for using the human image synthesis by GAN potentially by frauds thereby producing the fake and photographs and videos without permission. On social media, fake profiles can be prevented using GANs for generating the unique pictures of persons that do not exist.

VII. FUTURE SCOPE

However, research that has applied GANs to the more challenging scenario of video is limited. Furthermore, in terms of performance and capabilities, GAN-related research in other domains such as time-series generation and natural language processing lags behind in computer vision. there are clearly more opportunities for future research and application in these fields.

VIII. CONCLUSION

In this paper, reviewed GAN fundamentals and existing GAN variants from 2014 for different applications based on architecture, loss function and Performance metrics. Also developed a GAN selection methodology for the selection of model corresponding to the application. The real world applications of GANs were reviewed in detail. The limitations and challenge of GANs are also discussed in the paper along with the future scope.

REFERENCES

- [1] Goodfellow, Ian, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative adversarial nets." *Advances in neural information processing systems* 27 (2014).
- [2] Radford, Alec, Luke Metz, and Soumith Chintala. "Unsupervised representation learning with deep convolutional generative adversarial networks." *arXiv preprint arXiv:1511.06434* (2015).
- [3] Arjovsky, Martin, Soumith Chintala, and Léon Bottou. "Wasserstein generative adversarial networks." In *International conference on machine learning*, pp. 214-223. PMLR, 2017.
- [4] Gulrajani, Ishaan, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. "Improved training of wasserstein gans." *arXiv preprint arXiv:1704.00028* (2017).
- [5] Mirza, Mehdi, and Simon Osindero. "Conditional generative adversarial nets." *arXiv preprint arXiv:1411.1784* (2014).
- [6] Wang, Ting-Chun, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. "High-resolution image synthesis and semantic manipulation with conditional gans." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8798-8807. 2018.
- [7] Isola, Phillip, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. "Image-to-image translation with conditional adversarial networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125-1134. 2017.
- [8] Zhu, Jun-Yan, Taesung Park, Phillip Isola, and Alexei A. Efros. "Unpaired image-to-image translation using cycle-consistent adversarial networks." In *Proceedings of the IEEE international conference on computer vision*, pp. 2223-2232. 2017.
- [9] Zhang, Han, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. "Self-attention generative adversarial networks." In *International conference on machine learning*, pp. 7354-7363. PMLR, 2019.
- [10] Brock, Andrew, Jeff Donahue, and Karen Simonyan. "Large scale GAN training for high fidelity natural image synthesis." *arXiv preprint arXiv:1809.11096* (2018).
- [11] Karras, Tero, Timo Aila, Samuli Laine, and Jaakko Lehtinen. "Progressive growing of gans for improved quality, stability, and variation." *arXiv preprint arXiv:1710.10196* (2017).

- [12] Karras, Tero, Samuli Laine, and Timo Aila. "A style-based generator architecture for generative adversarial networks." arXiv preprint arXiv:1812.04948 (2018).
- [13] Karras, Tero, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. "Analyzing and improving the image quality of stylegan." In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8110-8119. 2020.
- [14] Yang, Zhijian, Junhao Wen, and Christos Davatzikos. "Smile-GANs: Semi-supervised clustering via GANs for dissecting brain disease heterogeneity from medical images." arXiv preprint arXiv:2006.15255 (2020).
- [15] Karras, Tero, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. "Alias-free generative adversarial networks." arXiv preprint arXiv:2106.12423 (2021).
- [16] Denton, Emily, Soumith Chintala, Arthur Szlam, and Rob Fergus. "Deep generative image models using a laplacian pyramid of adversarial networks." arXiv preprint arXiv:1506.05751 (2015).
- [17] Frid-Adar, Maayan, Eyal Klang, Michal Amitai, Jacob Goldberger, and Hayit Greenspan. "Synthetic data augmentation using GAN for improved liver lesion classification." In 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018), pp. 289-293. IEEE, 2018.
- [18] Qin, Zhiwei, Zhao Liu, Ping Zhu, and Yongbo Xue. "A GAN-based image synthesis method for skin lesion classification." *Computer Methods and Programs in Biomedicine* 195 (2020): 105568.
- [19] Han, Changhee, Leonardo Rundo, Ryosuke Araki, Yujiro Furukawa, Giancarlo Mauri, Hideki Nakayama, and Hideaki Hayashi. "Infinite brain MR images: PGGAN-based data augmentation for tumor detection." In *Neural approaches to dynamics of signal exchanges*, pp. 291-303. Springer, Singapore, 2020.
- [20] Chen, Hongming, Ola Engkvist, Yinhai Wang, Marcus Olivecrona, and Thomas Blaschke. "The rise of deep learning in drug discovery." *Drug discovery today* 23, no. 6 (2018): 1241-1250.
- [21] Li, Junde, Rasit Topaloglu, and Swaroop Ghosh. "Quantum generative models for small molecule drug discovery." arXiv preprint arXiv:2101.03438 (2021).
- [22] Huang, Xun, Ming-Yu Liu, Serge Belongie, and Jan Kautz. "Multimodal unsupervised image-to-image translation." In Proceedings of the European conference on computer vision (ECCV), pp. 172-189. 2018.
- [23] Yu, Xiaoming, Xing Cai, Zhenqiang Ying, Thomas Li, and Ge Li. "Singlegan: Image-to-image translation by a single-generator network using multiple generative adversarial learning." In *Asian Conference on Computer Vision*, pp. 341-356. Springer, Cham, 2018.
- [24] Zhang, Han, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks." In Proceedings of the IEEE international conference on computer vision, pp. 5907-5915. 2017.
- [25] Fedus, William, Ian Goodfellow, and Andrew M. Dai. "Maskgan: better text generation via filling in the " arXiv preprint arXiv:1801.07736 (2018).
- [26] Reed, Scott E., Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. "Learning what and where to draw." *Advances in neural information processing systems* 29 (2016): 217-225.
- [27] Jin, Yanghua, Jiakai Zhang, Minjun Li, Yingtao Tian, Huachun Zhu, and Zhihao Fang. "Towards the automatic anime characters creation with generative adversarial networks." arXiv preprint arXiv:1708.05509 (2017).
- [28] Huu, Phat Nguyen, Thuong Nguyen Thi Mai, Quang Tran Minh, and Hieu Nguyen Trong. "Proposing the Development of Dataset of Cartoon Character using DCGAN Model." In *International Conference on Future Data and Security Engineering*, pp. 325-339. Springer, Singapore, 2020.
- [29] Chen, Yang, Yu-Kun Lai, and Yong-Jin Liu. "Cartoongan: Generative adversarial networks for photo cartoonization." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 9465-9474. 2018.
- [30] Acornley, Christopher. "Using generative adversarial networks to create graphical user interfaces for video games." In Proceedings of the 2021 International Conference on Multimodal Interaction, pp. 802-806. 2021.
- [31] Mittal, Paritosh, Kunal Aggarwal, Pragya Paramita Sahu, Vishal Vatsalya, Soumyajit Mitra, Vikrant Singh, Viswanath Veera, and Shankar M. Venkatesan. "Photo-realistic emoticon generation using multi-modal input." In Proceedings of the 25th International Conference on Intelligent User Interfaces, pp. 254-258. 2020.
- [31] Semi-supervised, How Does GAN-based, and Generative Partial Multi-View Clustering. "CR-GAN: Learning Complete Representations for Multi-view Generation."
- [32] Ge, Yixiao, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, and Hongsheng Li. "Fd-gan: Pose-guided feature distilling gan for robust person re-identification." arXiv preprint arXiv:1810.02936 (2018).
- [33] Choi, Yunjey, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. "Stargan: Unified generative adversarial networks for multi-domain image-to-image translation." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 8789-8797. 2018.
- [34] Masood, Momina, Marriam Nawaz, Khalid Mahmood Malik, Ali Javed, and Aun Irtaza. "DeepFakes Generation and Detection: State-of-the-art, open challenges, countermeasures, and way forward." arXiv preprint arXiv:2103.00484 (2021).
- [35] Metz, Luke, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. "Unrolled generative adversarial networks." arXiv preprint arXiv:1611.02163 (2016).
- [36] Dolhansky, Brian, and Cristian Canton Ferrer. "Eye in-painting with exemplar generative adversarial networks." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7902-7911. 2018.
- [37] Salimans, Tim, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. "Improved techniques for training gans." *Advances in neural information processing systems* 29 (2016): 2234-2242.
- [38] Engel, Jesse, Kumar Krishna Agrawal, Shuo Chen, Ishaan Gulrajani, Chris Donahue, and Adam Roberts. "Gansynth: Adversarial neural audio synthesis." arXiv preprint arXiv:1902.08710 (2019).
- [39] Bińkowski, Mikołaj, Jeff Donahue, Sander Dieleman, Aidan Clark, Erich Elsen, Norman Casagrande, Luis C. Cobo, and Karen Simonyan. "High fidelity speech synthesis with adversarial networks." arXiv preprint arXiv:1909.11646 (2019).
- [40] Wu, Yi-Chiao, Tomoki Hayashi, Takuma Okamoto, Hisashi Kawai, and Tomoki Toda. "Quasi-periodic Parallel WaveGAN: A non-autoregressive raw waveform generative model with pitch-dependent dilated convolution neural network." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021): 792-806.
- [41] Thiem, Nathan, Marko Orescanin, and James Bret Michael. "Reducing Artifacts in GAN Audio Synthesis." In 2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA), pp. 1268-1275. IEEE, 2020.
- [42] Clark, Aidan, Jeff Donahue, and Karen Simonyan. "Efficient video generation on complex datasets." (2019).
- [43] Vondrick, Carl, Hamed Pirsiavash, and Antonio Torralba. "Generating videos with scene dynamics." *Advances in neural information processing systems* 29 (2016): 613-621.
- [44] Nakahira, Yuki, and Kazuhiko Kawamoto. "DCVGAN: Depth conditional video generation." In 2019 IEEE International Conference on Image Processing (ICIP), pp. 749-753. IEEE, 2019.

POINT OF VIEW ON ENTERPRISE-WIDE TEXT SUMMARIZATION APPROACHES

Sabari Rajan
Accenture, Advanced Technology
Centers in India
Chennai, India
rajan.k.sabari@accenture.com

Balasubramanian Vijayasankar
Accenture, Advanced Technology
Centers in India
Chennai, India
b.vijayasankar@accenture.com

Selvakuberan karuapasamy
Accenture, Advanced Technology
Centers in India
Chennai, India
s.b.karuppasamy@accenture.com

Subhashini Lakshminarayanan
Accenture, Advanced Technology
Centers in India
Chennai, India
s.j.lakshminarayanan@accenture.com

Abstract—At present, there has been a boom in the amount of text data generated from various sources at the enterprise level. The huge volume of text data is more valuable, and the knowledge extracted from the text data needs to be effectively synthesized to be a value. The manual text summarization process requires a lot of manual time and effort, and it will become cumbersome if the text data is huge. Summarization has been a key area of research in NLP for the past three decades. There are many state-of-art models available in the market for summarization techniques. In this paper, we have explored and made a comparative analysis of the best state-of-the-art text summarization models, and frameworks that are widely used across enterprises. Our observations indicate that T5 and BART are well suited for short summaries, Whereas BERT and BART are well suited for generating long summaries.

Keywords—Text summarization, Abstractive, Extractive, Features

I. INTRODUCTION

The summarization technique is a process of extracting the most important and relevant information from the source. In addition to text, summarization can be done in images as well as videos. The term image summarization refers to the extraction of the most relevant representative image frames from images, video summarization extracts the most important frames from video content. Text summarization is a process of generating a concise and significant summary of text from books, news articles, blog posts, research papers, emails, tweets, etc.

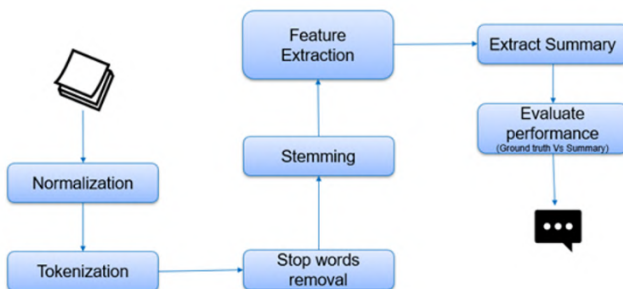


Fig. 1. Summarization process flow

Below are the steps for Summarization:

1. Importing the text data
2. Normalizing the text data
3. Tokenization
4. Removing stop words,
5. Stemming/lemmatization,
6. Extracting the relevant features
7. Final Summary output

Text summarization is referred to as be a common use case in the field of NLP and many research outcomes enhanced the efficiency of processing techniques followed for the summarization process.

Based on the nature of the text summarization process it can be divided into below categories,

1. Extractive Summarization
2. Abstractive Summarization

A. Extractive summarization.

It identifies sentences/phrases directly from the source document betting on their significance by decipherment necessary sections of the text present in the document and generating them verbatim, manufacturing a set of the sentences from the source textbook.

B. Abstractive summarization.

It generates a caption or summary conforming to many rulings that capture a composition's salient ideas or a passage. It involves a human-like approach, understanding the linguistics and context, and presenting it in a concise form.

The summary is formed by newly constructed meaningful sentences, which don't necessarily appear in the original input document. It helps us to avoid using pre-written text at the same time it needs large-scale data during training.

II. LITERATURE REVIEW

Abigail See et al [1], have proposed a solution that uses the seq-to-seq attention model to perform text summarization. The attention model is used to create a pointer-generator network that is created based on source text and then coverage is used to keep track of summarized text to avoid repetition. The model has shown an improvement of 2 ROGUE points over the current models on CNN / Daily Mail data.

Yisong Chen et al [2], have proposed a method that improvises on existing Text Summarization models involving TextRank and BART. Their proposed approach is that involves taking results from TextRank and BART splicing the data to form new text by increasing the weight of key sentences in the article later giving this text as input to a BART Model again. It is found that the results of

Rogue-1, Rogue-2, and Rogue-L recall metrics are increased by 1.5%, 0.5%, and 1.3% respectively.

Rada Mihalcea et al [3], have proposed an approach called TextRank. This involved creating a text graph from natural language texts based on Unsupervised learning algorithms without requiring any domain or language-specific annotated corpora. It can be used for key text extraction and key sentence extraction. It has also shown that this methodology has provided competitive results when compared with state-of-the-art methods.

Mike Lewis et al [4], have given a new approach for text generation using a method called BART (Bidirectional and Auto-Regressive Transformer). It is built by corrupting a text with noise and reconstructing it using a Transformer-based neural machine translation architecture. While this method has shown comparative results with other NLP tasks, it has shown the state-of-the-art results for Text generation tasks.

François St-Amant has mentioned that [5], BART Large can be used to create efficient models with very less training. They have pointed out that, the model contains 1024 hidden layers and 406 million parameters obtained by using CNN news data. They have also mentioned that it will cause memory and computing overhead due to the parameter and architecture size.

Yang Liu et al [6], have proposed a BERT-based approach for Text summarization problems. They have achieved this by stacking many Transformer layers for sentences and for Abstraction summarization they have used fine-tuning layers to avoid the mismatch between encoder and decoder layers. This method was found to have achieved staggering results in both human and automated evaluations.

Tinghuai Ma et al [7], have proposed an approach involving a Topic-aware BERT (Bidirectional Encoder Representations from Transformers) system. Here they have encoded the topic representation through neural topic model (NTM), then the data is sent through the transformer layer to learn the dependencies, and finally, LSTM layers are used to capture sequence information. The result of this model is found to be achieving state-of-the-art results.

Zolotareva et al [8], have proposed an approach involving seq-to-seq recurrent neural networks and transfer learning with unified text-to-text approaches. The text-to-text sequence-based transfer learning model has provided a better result than other methods involving linear models with handcrafted features.

Mehdi Allahyari et al [9], have surveyed the Text summarization techniques, Machine Learning algorithms that were used and their results, and the Evaluation methods available for this problem. It has pointed out the differences between Extractive and Abstraction summarization methods, Topic representation approaches, Knowledge base creation for automatic summarization, the impact of context in summarization, and Evaluation metrics for this problem.

III. EXPLORATORY ANALYSIS

At present, we have multiple texts summarizing models available in the market. Selecting the best summarization model that will be fit for a particular business problem needs certain features and prerequisites to be sorted out. for

example, the following are some of the features/characteristics for the evaluation of the summarization model

1. Summarization support – abstractive/extractive
2. Custom training
3. Training infrastructure
4. Framework support
5. Generate alternate summarization
6. Custom corpus training
7. Multiple document summarization

Before that, to understand the working of Sequence-to-Sequence models find the below part

A. Sequence to sequence model architecture

Sequence to sequence models comprises an encoder and decoder setup. When the input is passed on to the encoder setup. It generates the vectorized form of the input as an output which is then fed to the decoder set up as an input. The output generated by the encoder can be fed directly to the decoder as a block or it can be connected directly to the hidden units of the layers present in the decoder network at every step concerning time. Text summarization is one of the use cases of the sequence-to-sequence model architecture. The encoder-decoder network (see fig 2) can be a simple recurrent neural network or LSTM or GRU networks. The input text and its reference summary are fed as input to the encoder to train the model to generate the vectorized output.

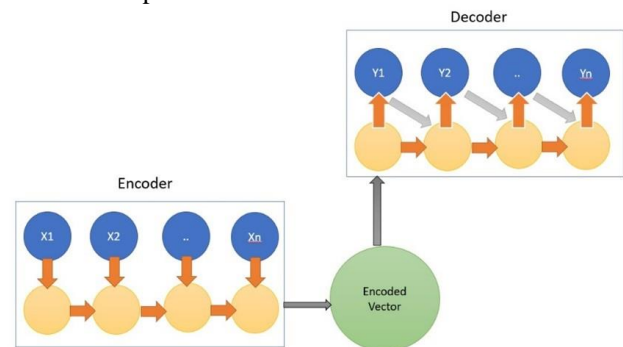


Fig. 2. Encoder-decoder Architecture

The output from the encoder network is given as an input to the decoder unit to decode the encoded version. After successive training progress, the model will be available to take new input & make an inference and provide machine-generated summaries.

B. Dataset:

To evaluate and compare the efficiency of the industry-driven models, the volume of data chosen needs to be adequate to feed the process of training, validate and test the models. We have utilized CNN/daily mail summarization dataset for this experiment. The CNN/daily mail summarization dataset contains news narrations from CNN and daily mail webpages. It includes news stories along with human-generated summaries. The corpus we framed contained 273,614 training sets and 15,597 validation sets & 14,294 sets for test. To get a better comparison, the same set of data sources was used for all models and results have been evaluated for the metrics.

We have considered the top five text summarizing models and their features for our exploratory analysis to solve our enterprise business problem. The top five summarization models include

1. Text Rank
2. BART
3. BARTLarge
4. BERT
5. T5

a) TextRank: TextRank is a graph-based unsupervised text summarization technique that supports Extractive Summarization (learning without knowing output class). It uses the PageRank-like technique to create a graph linking the tokens obtained from the input documents. Where [1-3] at first the text documents are converted into text data, which is then split into sentences, Later sentences are converted into numerical vectors, the similarity between these vectors are calculated and stored as similarity matrices and finally based on the ranking top sentences are returned as summary.

b) BART: Bidirectional and Auto-Regressive Transformer (or BART) is a sequence-to-sequence model which is pre-trained by adding noise to the data (i. e, Denoising Autoencoder) which makes it a better algorithm for Abstractive text Summarization tasks, as it helps in avoiding noise in data and capture only the relevant information. Since BART has [4] an autoregressive decoder it can be fine-tuned for summarization tasks directly by manipulating the input which is like the denoising pre-training task for which the model is built.

c) BART Large: BART Large model is developed by Facebook pretrained using CNN news data, it contains 1024 hidden layers and 406 million parameters which is the reason it is called BART Large. This model is suitable for Abstractive Summarization. Here, [5] the weights from the pre-trained model are used as a starting point, on top of this, the current problem data is fed into the system and trained to tune the weights for this specific problem.

d) BERT: The Bidirectional Encoder Representations from Transformers (BERT) differs from other machine learning algorithms as it can perform bi-directional training of a Transformer for language modeling which provides state-of-the-art results in both Extractive and Abstract text Summarization, this is possible because bi-directional training provides a better sense of data. tuning it for Extractive Summarization [6][7] CLS symbols are inserted to learn the sentence representation several inter-sentence Transformer layers are stacked on top of BERT outputs to obtain important features in the Document. For Abstractive, pretrained BERTSUM is used as an encoder and a 6 layered Transformer is used as a decoder.

e) T5: Text-to-Text-Transfer-Transformer (or T5) is a transformer-based model that supports Abstractive text summarization. This model is an extensive transfer-learning-based model trained on Colossal Clean Crawled Corpus (or C4) data which can perform most of the downstream text-to-text NLP tasks with state-of-the-art results. Here [8] three Long-Short Term Memory (LSTM) layers are used for the encoder and one LSTM Layer is used

additionally a custom attention layer is used to handle lengthy sequences. Later the pretrained T5 model is trained on our dataset to make it suitable for the specific task.

IV. EVALUATION METRICS

A. ROGUE

ROGUE is mainly used for evaluating the summary output [9]. It measures recall, which means the total number of words or n-grams in the human-generated reference summary that got replicated in the machine-generated summary. Naturally, these results are complementing, as is often the case in precision vs recall. The value of Bleu will be high when more number of words or n-grams from the system results are replicated in the human references, and the value of Rouge will be high when more no of words or n-grams from the human references are replicated in the system generated results.

B. Precision and recall in ROGUE

In terms of ROUGE, recall measures how many percentages of reference summaries are getting captured or recovered in a system-generated summary.

$$\frac{\text{No of overlapping words}}{\text{Total words in the refernce summary}}$$

In most cases, the system-generated summary will be very long & it may capture all the words that are available in the reference summary. Sometimes the words present in the system summary might be useless and will make the summarization process cumbersome. At that point, precision comes into play. It measures how much % of the system summary was appropriate.

Whereas precision is measured as:

$$\text{Precision} = \frac{\text{No of overlapping words}}{\text{Total words in the system summary}}$$

C. ROGUE1

It is used to calculate the amount of overlapping of unigrams between system and reference summary, ROUGE2 will measure the overlap of bigrams between system and reference summary.

D. ROGUEL

This measures the LCS- the longest common subsequence between system summary and reference summary.

V. RESULTS AND DISCUSSION

We have conducted multiple experiments on the text summarization models. Table 1. briefly explains the model features and fitment to the business problem.

Features	BERT	BART	T5	TextRank	BARTLong
Summarization type	both	Abstractive	Abstractive	Extractive	Abstractive
Framework Support	GPU/TPU	GPU/TPU	GPU/TPU	NA	GPU/TPU
Custom Training	Yes	Yes	Yes	No	Yes
Framework Support	TensorFlow	PyTorch	both	PyTorch	PyTorch
Generate alternate summarizations	NA	yes	No	No	yes
Multi-document	NA	No	Yes	No	No

Table 1. Models & features

Based on our experiments we have calculated ROUGE scores for the models. The results are depicted in Table 2.

Models / Approach	ROUGE1			ROUGE2			ROUGEL		
	precision	recall	F1	precision	recall	F1	precision	recall	F1
BERT	0.7	0.38	0.51	0.65	0.32	0.43	0.68	0.34	0.45
T5	0.69	0.22	0.34	0.47	0.15	0.23	0.64	0.21	0.31
BART	0.97	0.51	0.67	0.98	0.5	0.66	0.73	0.37	0.49
BART Long	0.8	0.51	0.62	0.69	0.44	0.53	0.55	0.35	0.43

Table 2. Models and features

VI. CONCLUSION

In this paper, we have evaluated the top five state of art models for summarization that is being widely used in enterprises. From our exploratory results, we conclude that T5 and BART models are well suited for generating short summaries out of the huge text. Comparatively, BART has high precision and recall value than T5. Similarly, BERT

and BART Long are well suited for generating long summaries.

REFERENCES

- [1] Abigail See, Peter J. Liu, and Christopher D. Manning: Get to The Point: Summarization with Pointer-Generator Networks, 2017, <https://arxiv.org/pdf/1704.04368.pdf>
- [2] Y. Cheng and Q. Song, "News Text Summarization Method based on BART-TextRank Model," 2021 IEEE 5th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC), 2021, pp. 2005-2010, DOI: 10.1109/IAEAC50856.2021.9390683.
- [3] A. Kazemi, Veronica P´erez-Rosas, Rada Mihalcea: Biased TextRank: Unsupervised Graph-Based Content Extraction: 2020, <https://arxiv.org/pdf/2011.01026.pdf>
- [4] M. Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, Luke Zettlemoyer: BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension, 2020, <https://arxiv.org/abs/1910.13461>
- [5] François St-Amant, Fine-Tuning the BART Large Model for Text Summarization: 2021. <https://towardsdatascience.com/fine-tuning-the-bart-large-model-for-text-summarization-3c69e4c04582>
- [6] Yang Liu and Mirella Lapata: Text Summarization with Pretrained Encoders-, 2019, <https://arxiv.org/pdf/1908.08345v2.pdf>
- [7] T. Ma, Q. Pan, H. Rong, Y. Qian, Y. Tian, and N. Al-Nabhan, "T-BERTSum: Topic-Aware Text Summarization Based on BERT," in IEEE Transactions on Computational Social Systems, 2021, DOI: 10.1109/TCSS.2021.3088506.
- [8] Zolotareva, Ekaterina & Misikir Tashu, Tsegaye & Horvath, Tomas. Abstractive Text Summarization using Transfer Learning. 2020. https://www.researchgate.net/publication/343863087_Abstractive_Text_Summarization_using_Transfer_Learning
- [9] Mehdi Allahyari, Seyedamin Pouriye, Mehdi Assefi, Saeid Safaei, Elizabeth D. Trippe Young, Juan B. Gutierrez, Krys Kochut: Text Summarization Techniques: A Brief Survey 2017, <https://arxiv.org/pdf/1707.02268.pdf>.

ENHANCED CRYPTOGRAPHIC SOLUTIONS TO PROTECT MESSAGES DURING COMMUNICATION IN VEHICULAR CLOUD COMPUTING

A. SHEELA RINI

Research Scholar, Avinashilingam Institute for Home Science & Higher Education for Women, Coimbatore.

sheelarini.a@gmail.com

Dr C. MEENA

Computer Center Incharge, Avinashilingam Institute for Home Science & Higher Education for Women,
Coimbatore.

cccmeena@gmail.com

ABSTRACT

VCC or Vehicular Cloud Computing is a network that combines VANET (Vehicular Adhoc NETWORK) and Cloud Computing principles along with wireless concepts to provide services related to safe transportation and traffic management. One important concern during VCC communication is the security of the messages transmitted. In this research work, hybrid cryptographic algorithms that combines layered and combination methods are proposed. The methodology works in two stages. In the first stage, 2 layered and 2 combination algorithms are designed, from which, the best performing method is selected. The selected methods are then hybridized to protect the VCC messages. Experimental results showed that the proposed hybrid cryptographic method is efficient when compared to the existing algorithms.

KEYWORDS

Hybrid Cryptography, Layered Cryptography, Combination Cryptography, Vehicular Cloud Computing, Secure Communication, Message Safety.

1. INTRODUCTION

The technological breakthroughs in software, hardware and communication has evolved in different types of networks, specially constructed to suit different environments. One such network is VANET (Vehicular Adhoc NETWORK). A VANET is a network that uses a set of moving vehicles to form a wireless network that can apply Information Communication Technology (ICT) to provide efficient services related to transportation and traffic management (Kugali and Kadadevar, 2020). The VANET consists of various components to help vehicles to communicate with each other. These components include, Road Side Units (RSUs), GPS (Global Positioning System) devices, cameras, radio transceiver, sensors, moving vehicles and the cloud servers. The huge amount of data sensed from these components need huge storage units along with fast computing devices and methods. As this requirement is difficult to handle by VANET components, VCC (Vehicular Cloud Computing) was introduced. This new technology is a part of Intelligent Transportation system and is designed as a hybrid system that combines the advantages of cloud computing with VANET (Antonio *et al.*, 2020). Currently, VCC has received more attention as it can provide efficient solutions in areas related to vehicle and road safety, improve traffic management, provide efficient entertainment services and provide better utilization of traffic

signals. VCC helps to improve communication between vehicles and work to provide a safe and efficient travelling environment.

The main objective of VCC, as mentioned earlier, is to create a safe and efficient travelling environment. However, VCC has several security holes that make the network vulnerable against attacks. Examples of such attacks include jamming (that prevent communication between vehicles), forging or falsifying fake hazard warning messages, message hampering (dropping or altering messages) and privacy violations. Previously, in order to ensure secure vehicular communication, a machine learning-based method was proposed to detect hacked vehicles. However, due to the high mobility characteristics of the vehicles, VCC also faces serious security issues, like authentication, message confidentiality, safety of messages communicated and secure location information.

This paper, focuses on techniques that ensures safety of messages communicated using cryptography. Cryptography is defined as secure communication techniques that allow only the source and the intended destination vehicles to access and view the message content. These algorithms transfer the messages into a hard to decipher form, which can be converted to its original state only by the intended destination vehicles. Cryptographic algorithms have envisaged huge advancements in the past few decades. Initially, the advancements were in the form of mono-alphabetic ciphers, polyalphabetic substitution ciphers, transposition ciphers and block ciphers (Aung *et al.*, 2019). Later on, more advancements were implemented using sophisticated algorithms like AES (Advanced Encryption Standard), DES (Data Encryption Standard), RSA (Rivest–Shamir–Adleman) and SHA (Secure Hash Algorithm). Each of these algorithms have their own merits and demerits.

Recent researches are focused on developing hybrid cryptographic algorithms that can combine their advantages in order to improve its efficiency in protecting messages being transmitted over VCC and thus construct a safe communication environment (Kumar *et al.*, 2021). This work, motivated by the success of hybrid algorithms, also proposes an enhanced 2-level hybrid algorithm that combines the advantages of multiple cryptographic algorithms to provide both vehicle level and message level security. The rest of the paper is organized as follows. Section 2 provides the methodology behind the design of hybrid cryptography algorithm. The algorithms used are BlowFish, RSA, 3DES, AES and MD5 (Message Digest-5). Section 3 analyses the performance of the proposed hybrid cryptographic algorithms and compare their results with the existing algorithms. Section 4 concludes the work with future research directions.

2. METHODOLOGY

According to Ekwonwune and Enyinnaya (2020), a hybrid algorithm refers to the usage of two or more cryptographic algorithms with the aim of creating a robust VCC model that can protect messages transmitted. In order to construct a secure communication VCC model, this work proposes a 2-level Hybrid Cryptographic (2-HC) system, where the first level focuses on providing vehicle level security, while second level focuses on message level security.

In general, safety messages are broadcasted every 100 to 300 milliseconds (Liu *et al.*, 2020). At the receiving end, the sender's identity is verified for authenticity. However, as VCC is a high speed network, where even a small delay can cause catastrophic situations, the authentication has to be done in a fast manner. In this paper, to solve this issue, the first level of 2-HC system is focused on correctly identifying valid vehicles by making sure that only registered user's access data. That is, the method allows only vehicles which are part of

clusters involved in communication, to access the messages. The rest of the vehicles (that is, public) cannot perform any operation on them. This is implemented using a method that is similar to login, password system commonly used in networks. Here, the vehicle's license plate along with driver's license is used as password to get access to cloud messages.

The second level of 2-HC system focuses on message level security where multiple cryptographic algorithms are used to protect the messages send using cloud systems. In practice, two types of methodology are used to combine cryptographic algorithms (Chakraborty *et al.*, 2020). They are,

1. Layered Cryptographic Algorithms : These algorithms provide the ability to use different encryption algorithms on different portions of a message. The advantage of this methodology is accessing a single part would not reveal the whole message.
2. Combination Cryptographic Algorithms : These algorithms apply multiple encryption algorithms on the whole message, thus making it difficult to hack the message as the hacker has to handle two or more encryption algorithms.

This work designs 2 layered and 2 combination algorithms, from which the efficient ones are selected and fused to form a hybrid algorithm. Thus, the proposed hybrid algorithm combines layered and combined algorithm.

The design of layered algorithm uses two popularly used cryptographic method, namely, Blowfish (B) and DES (D). This algorithms performs cryptography in two steps (Figure 1). The first step splits the message into two blocks. While the second method applies, different cryptographic algorithm to each block. During the design of layered algorithms, the order of applying the cryptographic algorithms is critical. In this work, two layered algorithms are proposed, where the first method applies Blowfish to encrypt block 1 and DES to encrypt block 2. The second method, on the other hand, applies DES to encrypt block 1 and blowfish to encrypt block 2. The methods respectively are termed as LC_B+D algorithm and LC_D+B algorithm.

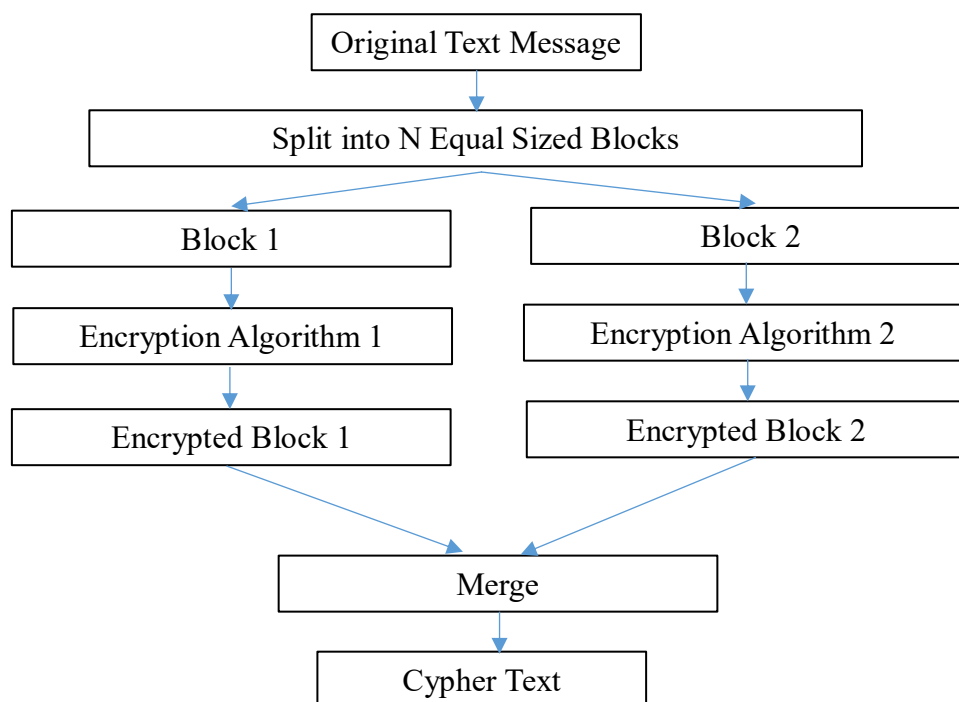


Figure 1 : Layered Cryptographic Algorithm

The design of combination cryptographic uses RSA and AES algorithms. As with layered algorithm, two methods, which differ in the order of applying the RSA and AES algorithms, are proposed. The methods respectively are termed as CC_R+A (RSA is applied first, followed by the application of AES) and CC_A+R (AES is applied first, followed by RSA). The steps involved are presented in Figure 2.

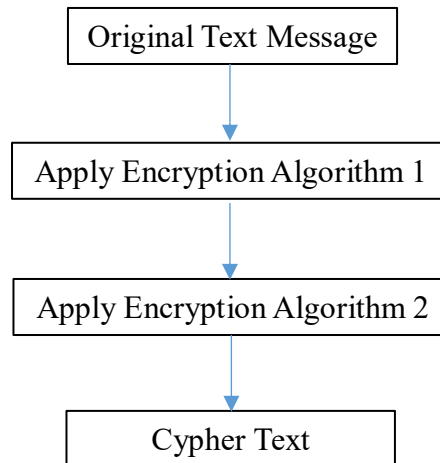


Figure 2 : Combination Cryptographic Algorithm

Performance evaluation of the above four designs showed that both layered and combination methods improve message security and motivated by these results, hybrid algorithms that joined layered and combination methods are proposed. This algorithm is termed as Hybrid cryptographic algorithms. Again, two types of hybridization are designed, which differed in the order of using layered and combination methods. The first applies layered algorithm followed by combination algorithm and is termed as HLC Cryptographic Algorithm. The second applies combination first followed by layered and is termed as HCL Cryptographic Algorithm. The steps involved are respectively shown in Figures 3a and 3b. In both the algorithms, the MD5 algorithm is included to improve integrity.

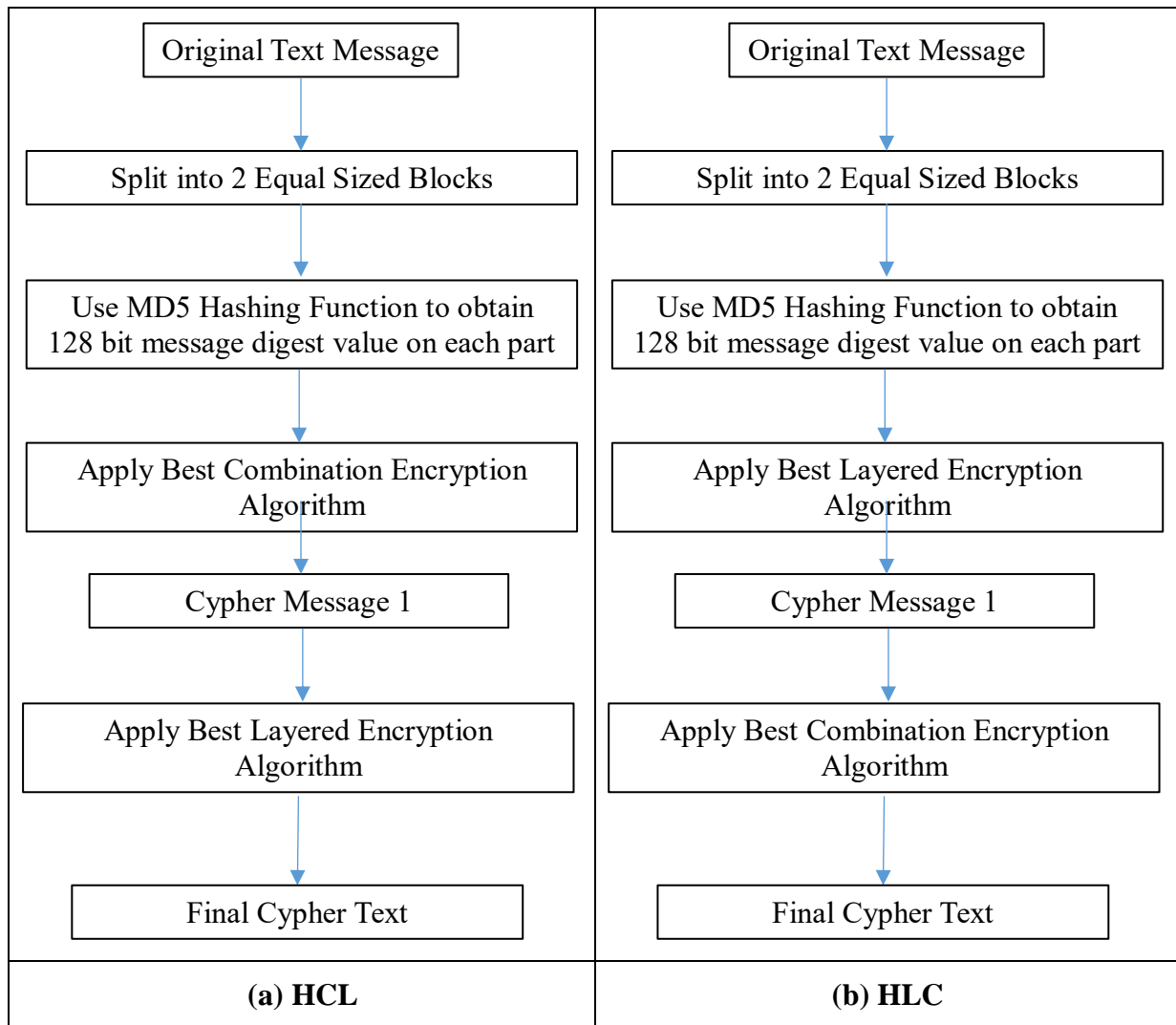


Figure 3 : Steps in Hybrid Cryptographic Algorithm

Both the fusion algorithms, HCL and HLC, combines the advantages of layered and combination cryptography, thus providing a secure message transmission environment.

3. EXPERIMENTAL RESULTS

Several experiments were conducted to evaluate the performance of the proposed algorithms. Computation overhead of encryption and decryption algorithms, measured in seconds, were used as performance measure. The coding scheme used is presented in Table 1. Figures 1a,b to 3a,b show the encryption and decryption time taken by the layered, combination and proposed hybrid algorithms respectively..

TABLE 1 : CODING SCHEME

Code	Description
B	BlowFish Algorithm
D	3DES Algorithm
A	AES Algorithm
R	RSA Algorithm

LC_B+D	Layered Cryptographic Algorithm Using Blowfish and DES
LC_D+B	Layered Cryptographic Algorithm Using DES and Blowfish
CC_A+R	Combination Cryptographic Algorithm Using AES and RSA
CC_R+A	Combination Cryptographic Algorithm Using RSA and AES
HLC	Hybrid Layered and Combination Cryptographic Algorithm
HCL	Hybrid Combination and Layered Cryptographic Algorithm

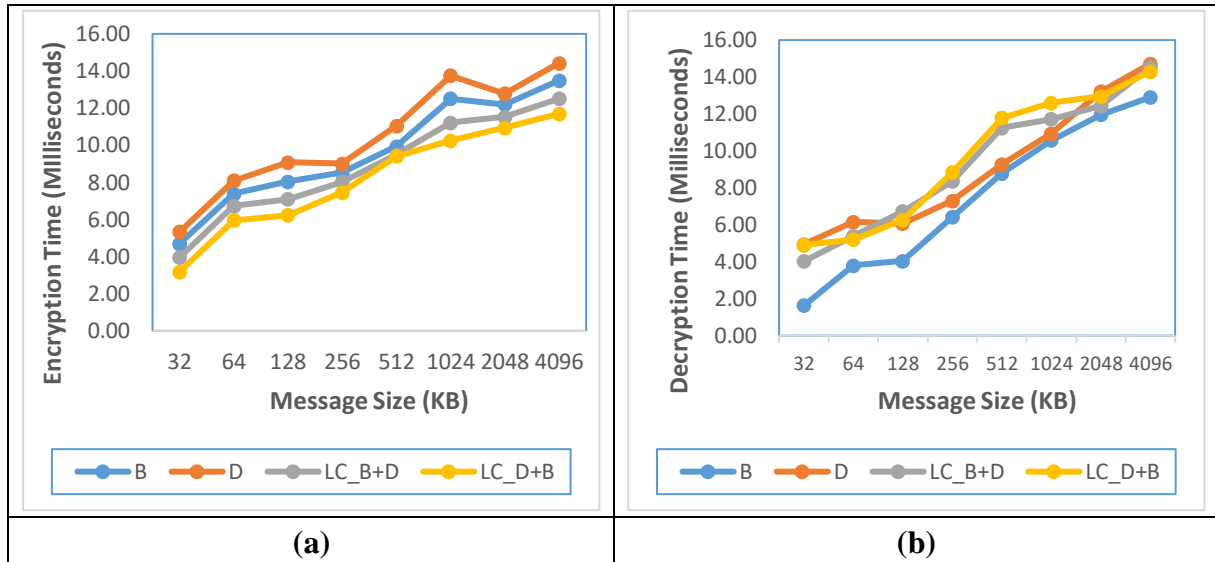


Figure 1 : Analysis of Layered Cryptographic Algorithms

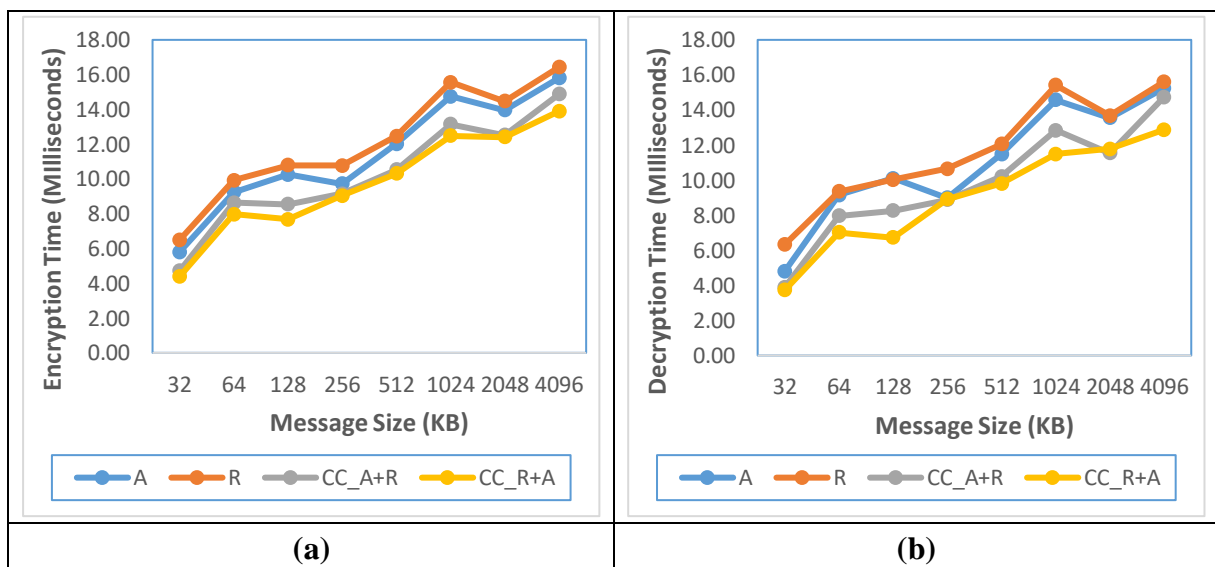


Figure 2 : Analysis of Combination Cryptographic Algorithms

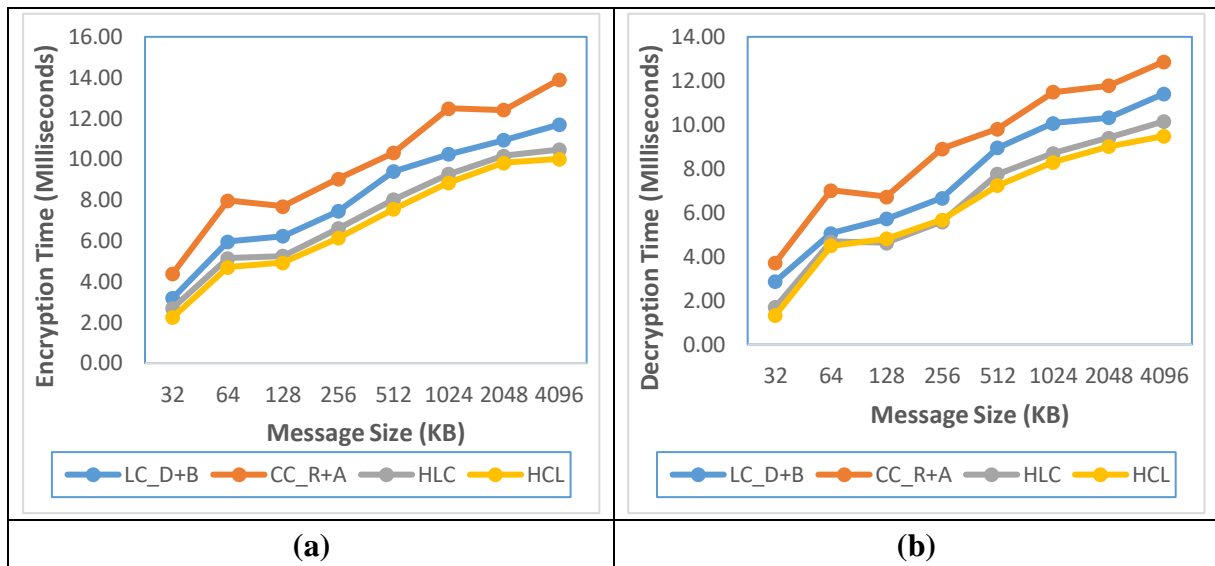


Figure 3 : Analysis of Hybrid Cryptographic Algorithms

From the results, it is understood that the layered algorithms are faster to produce encrypted and decrypted message when compared with combination algorithms. Among the layered algorithms, the algorithm that applied D first and then B had less time complexity. Among the combination algorithms, the algorithm that applied R first followed by A produced results in a fast manner. However, maximum efficiency was produced by the proposed hybrid algorithms with respect to both encryption and decryption time. Among the proposed hybrid algorithms, the algorithm that used layered algorithm on the result of combination algorithm was the fastest.

4. CONCLUSION

During VCC communication, it is very important to protect the messages being transmitted by the vehicles. In this paper, cryptographic algorithms are proposed to protect the messages. Hybrid cryptographic algorithms that combine layered and communication based methodologies are proposed. For this purpose, a 2-stage algorithm is proposed. In the first stage two layered and two combined algorithms are designed, from which, the best performing algorithm is selected. These algorithms are then combined to form the proposed hybrid method. For this purpose, four algorithms, namely, Blowfish, 3DES, AES and RSA algorithms, are considered. Experiments showed that the method combined 3DES with Blowfish along with the method that combined RSA with AES algorithm, using layered-combination fashion of hybridization produced maximum efficiency. In future, methods that can further protect the messages, like signcryptography, will be analyzed and explored.

REFERENCES

- Antonio, G., Sameer, S.M., Jun, L. and Wensong, W. (2020) Security and Privacy in Vehicular Ad Hoc Network and Vehicle Cloud Computing: A Survey, Article ID 5129620, Vol. 2020, Pp. 1-25.
- Aung, T.M., Naing, H.H. and Hla, N.N. (2019) Complex Transformation of Monoalphabetic Cipher to Polyalphabetic Cipher : (Vigenère-Affine Cipher), International Journal of Machine Learning and Computing, Vol. 9, No. 3, Pp. 296-303.

- Chakraborty, R., Bairagi, A. and Bandyopadhyay, S.K. (2020) Design and implementation of two-layer encryption system in cryptography, Vol. 5, Issue 2, Pp. 424-427.
- Ekwonwune, E. and Enyinnaya, V. (2020) Design and Implementation of End to End Encrypted Short Message Service (SMS) Using Hybrid Cipher Algorithm. Journal of Software Engineering and Applications, 13, 25-40.
- Kugali, S.N. and Kadadevar, S. (2020) Vehicular ADHOC Network (VANET):-A Brief Knowledge, International Journal of Engineering Research & Technology, Vol. 09, Issue 06, Pp. 1026-1029.
- Kumar, S., Karnani, G., Gaur, M.S. and Mishra, A. (2021) Cloud Security using Hybrid Cryptography Algorithms, 2nd International Conference on Intelligent Engineering and Managem, Pp. 597-602.
- Liu, Y., Wang, L. and Chen, H.H. (2015) Message Authentication Using Proxy Vehicles in Vehicular Ad Hoc Networks, IEEE Transactions on Vehicular Technology, Vol. 64, No. 8, Pp. 3697-3710.

REVIEW ON INTRUSION DETECTION SYSTEM IN WIRELESS SENSOR NETWORK

Jyoti Srivastava^{1*}, Jay Prakash², Anu Raj³

^{1,2,3} Department of ITCA, Madan Mohan Malviya University of Technology, Gorakhpur

sriv.jyoti1996@gmail.com, jpr_1998@yahoo.co.in, anu.raj10@yahoo.com

Abstract

The sensor nodes are installed in Wireless Sensor Networks (WSNs) to collect information from the external surroundings. WSNs are very prone to security at many levels because of their dispersed nature, multihop data transmission, and open wireless channels. Intrusion Detection Systems (IDSs) are useful for detecting and preventing security breaches. This paper discusses existing Machine Learning (ML) approaches for IDS as well as a brief introduction and the types of IDS in WSN security.

Keywords- *Wireless Sensor Network, Intrusion Detection System, Machine Learning, Support Vector Machine, Firewall.*

1.INTRODUCTION

The ID is a security system for the computer & network. An ID system collects and examines information from multiple parts of a computer or network to identify possible threats to security, including (outside-the-board attacks) intrusions, or harassment (in-office attacks). ID uses a scanning technology to test operating applications or network security. ID provides an evaluation of a vulnerability. It uses an evaluation of vulnerability. An IDS is considered a burglar detector[1]. For eg, the locking mechanism in the house defends the house from burglary. In comparison, if someone cracks a house lock or attempts to enter, a burglar alert is violated but alarms the owner to this by an alert. Firewalls are

also a very effective firewall to stop the firewall from coming into internet traffic. For example, the modem of the company's private network may allow external users to link to Intranet, and firewalls cannot detect such access[2]. Intrusion means that without permission, someone is disturbed. An intrusion is an act that causes accidental damage without privileges utilizing system resources. Intrusion detection is every method to identify an intrusion. The IDS controls network traffic and possible security determinants. It monitors the traffic in the network[3]. If a threat is detected, the system or network administrator may warn. Intrusions and alerts may be observed by IDS. IDS is a category of technology or approaches only often used to classify suspicious activities at the network or host level[4].

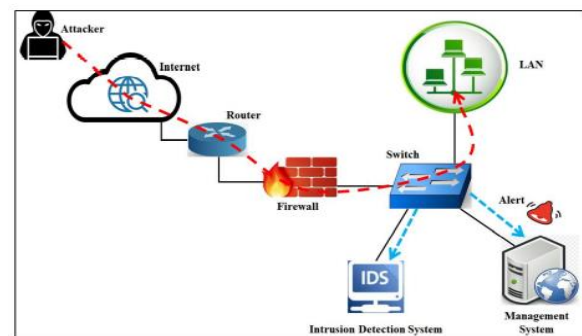


Figure 1- : Intrusion Detection System

The term fraud applies to the abuse of the system of a profitable company without contributing to legal consequences. Fraud could become a business-critical problem in a competitive environment if it is pervasive and prevention methods are not unsafe. As part of

overall fraud prevention, fraud detection automates or helps to reduce the scanning but testing process manual sections. It is a renowned application in the data mining industry/government[5]. The validity and motive behind a request or transaction cannot be completely assured. The best economic choice is to use mathematical algorithms to eliminate potential evidence of fraud of available data[6]. Fig. 1 shows how entry and warning are maintained by IDS.

1.2 Types of IDSs

There are two types of IDS based on their installed placement, host-based IDS and network-based IDS.

A. Host-based IDS (HIDS)

A HIDS is a type of IDS that runs on a computer, node, or device. Even though many varieties of HIDS have been produced that can be used to detect networks, its primary use is internal surveillance. A HIDS can identify malicious software that unexpectedly utilizes a system's resources or discovers that a program has manipulated the registry in a hazardous way by analyzing the entire communication stream and alerting administrators.

HIDS is a software agent that monitors and analyses the activities of particular hosts, such as files, processes, and system logs. HIDSs have a variety of tools at their disposal. The existence of unauthorized or suspect activities can be checked by comparing system snapshots. Multiple failed login attempts, as well as unusually high CPU utilization for an extended period, are indicators of possible assaults. By evaluating system calls and modifications to system binaries, some HIDSs can also do kernel-based detection. They could also be used to spy on people by tracking their movements[6].

B. Network-based IDS (NIDS)

Sensors are often used by NIDS at different locations throughout the network. The sensor analyses the traffic on its own or with the help of a central controller. Because NIDSs are more flexible and cross-platform than HIDSs, they are used to safeguard a company's IT equipment more frequently. These methods, on the other hand, can be used in conjunction to provide a better level of protection[7].

The purpose of NIDS is to monitor and analyze network traffic. Its main function is to safeguard the computer from network-based risks by detecting unauthorized harmful access to a LAN and examining traffic that passes via numerous hosts across the wire. Detection techniques examine inbound and outgoing packets for any suspicious network, prompting NIDS to send an alert to the administration. This system includes three network topologies: direct connection to a switch spanning port, network tap, and inline connection. Conventional IT security precautions are delivered through IDS technology, as well as customized ones tailored to the unique aspects of ICS. We felt it was critical to have a clear comparison of the two systems, therefore we summarised some of the benefits and drawbacks of both types in table 1.1.

Table 1- Summarised some of the benefits and drawbacks of HIDS & NIDS.

IDS Type	Advantages	Disadvantages
HIDS	Examines the complete operating system and can examine the ongoing communication stream. Insider assaults that do not require network traffic can be detected,	When installing settings for each host, more administrative work is

	<p>and end-to-end encrypted communications can be checked.</p> <p>It does not necessitate the installation of any additional hardware.</p> <p>Checks system calls, system directories, application logs, and user actions for intrusions.</p>	<p>required.</p> <p>When such forms of DOS attacks[8] occur, this feature may be deactivated, resulting in HIDS capabilities being lost.</p> <p>Because OS audit logs take up a lot of space, it uses up a lot of resources on the host.</p> <p>Only keeps track of local attacks on the machine where it's installed.</p> <p>Delays in reporting assaults are possible.</p>
NIDS	<p>By monitoring network traffic, it identifies attacks.</p> <p>Multiple hosts on the system are monitored for any unusual behavior.</p> <p>Direct attacks are not possible, and attackers are unable to discover them.</p> <p>Checks a wide range of</p>	<p>When the network volume is enormous, it can be difficult to distinguish assaults from high-speed encrypted traffic.</p> <p>Because</p>

	<p>network protocols (TCP/UDP/ICMP/SNMP) as well as router NetFlow records.</p>	<p>switches have restricted surveillance port functionality, certain networks are unable to deliver all data processing.</p> <p>A certain piece of hardware is required.</p> <p>Only network assaults are identified.</p>
--	---	---

Data-gathering sensors are used to keep an eye on the equipment that collects data and to keep track of specific processes or protocols. They classify the data they acquire from their location in the first place.

Detector Engine This module compares the collected data to the set of rules that have been defined. When IDS detect a change in the normal status, it sounds the alarm.

Storage Module It holds the IDS rule sets, in which the detectors compare the gathered information[9].

Response When an alarm goes off, it performs a predetermined action. Regardless of the type of warning, the IDS may be able to perform a specified action, such as discarding malicious packets. It can also be a passive response, such as logging activity and allowing the human factor to determine what to do next[10].

The following is an example of a typical IDS architecture:

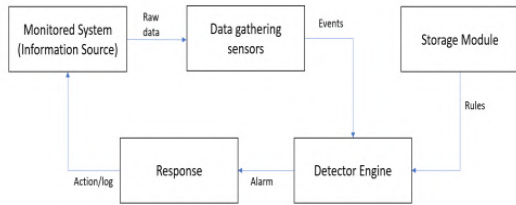


Figure 2- Architecture of IDS

2. Classification of Algorithms

Classification is an admirable activity in ML, specifically in future designs & exploration of information. In this section, important classification algorithms were described.

A. Naive Bayes (NB)- NB is a classification algorithm. It is for binary (twin-class) or multi-class problems and seems to be easier to appreciate when binary or definite input values are represented. The estimation of the probability for each hypothesis is made easier to make the estimate tractable. It is called Naive Bayes. These are meant to be conditionally independent provided the target value rather than to try to compute values for each attribute. This is an exceptionally solid presumption that is most improbable in genuine information that is not working with attributes. Nevertheless, the approach does not incorporate this statement surprisingly well in detail. Its primary difficulty is that it can't learn interactions amid features[11].

B. Logistic Regression (LR)- LR is a good method where the independent variable is boolean. Islogistic regression in statistical modeling. It is used to describe data or to explain the relationship among a binary variable that depends on it or a similar or even more specific interval variable or ratio point. It provides several ways to regularize your layout unlike Naive Bayes, you do not have to be worried about the connection between your apps. If you require a probabilistic framework or if

you want to obtain further training knowledge in the future, with which you need to be able to easily become a guide, use this to efficiently adjust grouping limitations[12].

C. Gaussian Distribution- Gaussian distribution (Normal distribution) is very common by nature. Almost all variables are distributed approximately normally. An incredible number of procedures in nature and sociologies normally pursue Gaussian distribution. Even after they don't, Gaussian provides the best model estimation for these procedures. Even though they are just roughly ordinary, they are commonly very close. The idea behind it is a central limit theorem. The central limit theorem states that once we include an extensive number of autonomous irregular factors, independent of the first circulation of these factors, their standardized aggregate tends towards a Gaussian distribution. Numerous sorts of factual tests are gotten from a normal distribution and function admirably if the distribution is roughly ordinary. A few tests function admirably even with a wide deviation from normality[13].

D. Support Vector Machine- The goal of SVM classification is to discriminate between two groups by providing relevant data with a feature and producing a classifier that performs well on hidden data. The maximum range classification is the most basic sort of SVM. The main classification problem is frequently solved by binary classification of linear separable training data[14].

E. Artificial neural network (ANN)- A biological NN computational model is known as ANN. ANN is another name for NN. The concept of ANN is derived mostly from biology, where the NN plays a fundamental and important role in the human body. In the NN, practice is done on a human body. A NN can be thought of as a collection of connected input/output units, each with its own weigh[15]t.

F. Decision Tree (DT)- DT algorithms are the most widely used algorithms in classification. It also aids in the classifying process. DT provides an easy modeling technique. A decision tree is a simple tool that allows people to quickly inspect a tree structure in order to understand how decisions are made[16].

G. K- Nearest Neighbor (KNN)- The NN approach is used to find the unknown data point by focusing on the nearest neighbor whose value has already been determined. Search for the closest point. The NN mechanism can be divided into two ways. Structure and function are less used NN classification approaches. The scheme classifies K-NN as a less method. The KNN method makes use of the NN for the value of k, which specifies how many NN to a sample data point must be checked in the class description. There are two types of NN strategies: KNN dependent structure and KNN less structure[17].

H. Bayesian network- Belief networks is another name for BN. A BN is a probabilistic visual-spatial distribution. This BN is made up of two parts. The first element is basically a directed acyclic graph (DAG), which refers to the graph nodes as random variables and represents probabilistic addition concerns on the edges observed between nodes or random variables[18].

Table 1- Table showing existing ML approaches for IDS.

Author & Year	Title	Method	Evaluation Results	Reference
Jeng-Shyan Pan et al. [2021]	A Lightweight Intelligent ID	k-nearest neighbor algorithm	Detection Rate=99.206 False	[19]

	Model for Wireless Sensor Networks	m (kNN) and Compact cosine algorithm (CSCA) polymorphic mutation (PM)	Alarm Rate=0.5848 Accuracy Rate=99.327	
Jing Jin [2021]	ID Algorithm and Simulation of WSN under Internet Environment	BiLSTM+C5.0	Accuracy =99.57 Detection Rate=94.61 False Alarm Rate=19.85	[20]
Achmad Akbar Megan tara and Tohari Ahmadi [2021]	A hybrid machine learning method for increasing the performance of network IDS	hybrid machine learning method	Accuracy =91.86 False Alarm Rate=1-2% Specificity Score=99%	[21]
Samir Ifzarne [2021]	Anomaly Detection using ML Techniques in WSN	information gain ratio and the online Passive-aggressive	Accuracy = 96% Precision =0.96 Recall=0.96 F1-	[22]

		ve classifie r (ID- GOPA)	score=0. 96	
Nada M. Alruha ily [2021]	A Multi- layer Machine Learning -based IDS for WSN	Multi- layer ID framew ork	TPR=0.9 9 TNR=0.9 77 FPR=0.0 23 FNR=0.0 02	[23]
Ashwi ni B. Abhal e and S. S. Maniv annan [2020]	Supervis ed ML Classific ation Algorith mic Approac h for Finding Anomaly Type of ID in WSN	SVM	Accuracy = 0.84 Precision =0.87 Recall=0 .86 F1- score=0. 87	[24]
Xiaop eng Tan et al. [2019]	WSN- ID Based on SMOTE and the Random Forest Algorith m	syntheti c minority oversam pling techniqu e (SMOT E) with random forest algorith m	Accuracy Rate= 92.39	[25]

By using ML to the challenge of resource management in WSN, creating a classifier that is well-trained with network patterns, and identifying and preparing a suitable dataset. Furthermore, including smart tactics such as compressing the input dataset, reducing the scale of characteristics set, and simplifying the analytical and decision-making process could help IDS meet the WSN requirement limitation while maintaining security and dependability. The main challenge in developing an IDS for the WSN is to identify attacks with high accuracy while meeting the needed restrictions and obstacles to extend the network's lifetime. This goal could be achieved in a variety of ways. To begin, pay considerably greater attention to assault detection strategies that are defined by efficiency and ability. Secondly, to reduce communication overhead, recreating the detecting system in a distributed manner.

3. CONCLUSION

REFERENCES

- [1] E. Darra and S. K. Katsikas, "A survey of intrusion detection systems in wireless sensor networks," *Intrusion Detect. Prev. Mob. Ecosyst.*, pp. 393–458, 2017, doi: 10.1201/b21885.
- [2] A. Yadav, "Network design: Firewall, IDS/IPS," *INFOSEC*, 2020. <https://resources.infosecinstitute.com/topic/network-design-firewall-idsips/>.
- [3] S. N. Kane, A. Mishra, and A. K. Dutta, "Preface: International Conference on Recent Trends in Physics (ICRTP 2016)," *J. Phys. Conf. Ser.*, vol. 755, no. 1, pp. 0–6, 2016, doi: 10.1088/1742-6596/755/1/011001.
- [4] A. H. Farooqi and F. A. Khan, "A survey of intrusion detection systems for wireless sensor networks," *Int. J. Ad Hoc Ubiquitous Comput.*, vol. 9, no. 2, pp. 69–83, 2012, doi: 10.1504/IJAHUC.2012.045549.
- [5] Pankaj, "Intrusion Detection System (IDS)," *Geeks for Geeks*, 2022. <https://www.geeksforgeeks.org/intrusion-detection-system-ids/>.
- [6] I. Butun, S. D. Morgera, and R. Sankar, "A survey of intrusion detection systems in wireless sensor networks," *IEEE Commun. Surv. Tutorials*, vol. 16, no. 1, pp. 266–282, 2014, doi: 10.1109/SURV.2013.050113.00191.
- [7] R. Plaka, "INTRUSION DETECTION USING MACHINE LEARNING FOR INDUSTRIAL CONTROL SYSTEMS Examiner: Radu Dobrin Supervisors: Sasikumar Punnekkat," *Mälardalen Univ. Sch. Innov. Des. Eng.*, 2021.
- [8] B. S. Sarang, H. H. Patel, and K. M. Patel, "Intrusion Detection and Prevention Techniques for Dos Attack with Security Patterns," vol. 8, no. 07, pp. 320–322, 2019.
- [9] J. Kizza and F. Migga Kizza, "Intrusion Detection and Prevention Systems," *Secur. Inf. Infrastruct.*, pp. 239–258, 2011, doi: 10.4018/978-1-59904-379-1.ch012.
- [10] N. M. Shanono, N. A. Abu, and W. Mohamed, "Intrusion Detection System Architecture: Issues and Challenges," *Glob. J. Comput. Sci. Technol.*, vol. 62, no. 7, 2020.
- [11] Y. El Mourabit, A. Toumanari, A. Bouirden, and N. El Moussaid, "A comparative evaluation of intrusion detection techniques in wireless sensor network," *J. Theor. Appl. Inf. Technol.*, vol. 76, no. 1, pp. 27–35, 2015.
- [12] Y. Natsume, "Machine Learning 102: Logistic Regression," *towardsdatascience*, 2022. <https://towardsdatascience.com/machine-learning-102-logistic-regression-9e6dc2807772>.
- [13] H. Nhs, "Gaussian Distribution for Machine Learning and Data Science (Normal Distribution)," *Medium*, 2019. <https://hemanthnhs.medium.com/gaussian-distribution-for-machine-learning-and-data-science-normal-distribution-bc90139e226d#:~:text=Gaussian or Normal Distribution is a very common,the weight vector for our Linear Regression Model>.

- [14] S. Morris, "Image classification using SVM," *Rpubs.Com*, 2018, [Online]. Available: https://rpubs.com/Sharon_1684/454441.
- [15] N. C. Steven Walczak, "Artificial Neural Network," *Science Direct*, 2003. <https://www.sciencedirect.com/topics/engineering/artificial-neural-network>.
- [16] P. Gupta, "Decision Trees in Machine Learning," *towardsdatascience*, 2017. <https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052>.
- [17] O. Harrison, "Machine Learning Basics with the K-Nearest Neighbors Algorithm," *towardsdatascience*, 2018. <https://towardsdatascience.com/machine-learning-basics-with-the-k-nearest-neighbors-algorithm-6a6e71d01761>.
- [18] jason brownlee, "A Gentle Introduction to Bayesian Belief Networks," *machine learning mastery*, 2019. <https://machinelearningmastery.com/introduction-to-bayesian-belief-networks/>.
- [19] J. S. Pan, F. Fan, S. C. Chu, H. Q. Zhao, and G. Y. Liu, "A Lightweight Intelligent Intrusion Detection Model for Wireless Sensor Networks," *Secur. Commun. Networks*, vol. 2021, 2021, doi: 10.1155/2021/5540895.
- [20] J. Jin, "Intrusion Detection Algorithm and Simulation of Wireless Sensor Network under Internet Environment," *J. Sensors*, vol. 2021, 2021, doi: 10.1155/2021/9089370.
- [21] A. A. Megantara and T. Ahmad, "A hybrid machine learning method for increasing the performance of network intrusion detection systems," *J. Big Data*, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00531-w.
- [22] S. Ifzarne, H. Tabbaa, I. Hafidi, and N. Lamghari, "Anomaly Detection using Machine Learning Techniques in Wireless Sensor Networks," *J. Phys. Conf. Ser.*, vol. 1743, no. 1, 2021, doi: 10.1088/1742-6596/1743/1/012021.
- [23] N. M. Alruhaily and D. M. Ibrahim, "A Multi-layer Machine Learning-based Intrusion Detection System for Wireless Sensor Networks," no. May, 2021.
- [24] B. A. Ashwini and S. S. Manivannan, "Supervised Machine Learning Classification Algorithmic Approach for Finding Anomaly Type of Intrusion Detection in Wireless Sensor Network," *Opt. Mem. Neural Networks (Information Opt.*, vol. 29, no. 3, pp. 244–256, 2020, doi: 10.3103/S1060992X20030029.
- [25] X. Tan *et al.*, "Wireless sensor networks intrusion detection based on SMOTE and the random forest algorithm," *Sensors (Switzerland)*, vol. 19, no. 1, 2019, doi: 10.3390/s19010203.

ANIMAL DETECTION BY YOLO COCO MODEL USING IMAGES

G.Elaiyaraja¹, T.K.Kalaiarasan² and C.Manikanta³

¹Professor, Department of ECE, VEMU Institute of Technology, P. Kothakota, Chittoor-517112, Andhra Pradesh, India,

²Assistant Professor, Department of ECE, VEMU Institute of Technology, P. Kothakota, Chittoor-517112, Andhra Pradesh, India,

³Assistant Professor, Department of ECE, VEMU Institute of Technology, P. Kothakota, Chittoor-517112, Andhra Pradesh, India,

ABSTRACT

Object detection is an important and challenging field in computer vision that allows us to identify and locate objects in an image or video. A broad range of techniques in computer vision and deep-learning has shown enormous potential to identify objects in images. In previous algorithm such as Faster R-CNN only object detection takes place, it does not specify the class of the objects in image. To overcome this problem 'You Only Look Once' (YOLO) is used where it detects and specifies class of the object simultaneously. The YOLOv3 uses Darknet-53 which is a convolutional neural network and it has 106 layers. In this the prediction in the entire image is done in a single algorithm run. By using the COCO dataset, the object Detection model is trained. COCO (Common objects in context) is a dataset which is used in large-scale Object detection, segmentation and captioning the objects in images. This model is fine-tuned for identifying ten different types of animals. The detection of YOLOv3 is very accurate and efficient.

Keywords: *YOLOv3, CNN, Accuracy*

1. INTRODUCTION

1.1 Introduction to Computer Vision

The computer vision-based methods are being used increasingly as tools to assist wild animal object recognition. The ability to identify individual animals from images enables population surveys through sight-resight statistics and forms the basis for demographic studies. The pipeline of processing for animal recognition includes several stages, starting with the detection of animals in images and ending with identification decisions. By making all stages of this pipeline more reliable and automated, animal identification studies can be increased in spatial and temporal resolution, provide better conservation statistics and importantly allow citizens without specialized training to participate in engaging census data collection events.

There have been increasing reports of wild animals entering villages or towns, especially in settlements surrounding forest areas, endangering human lives. In-turions by animals cause huge

losses are it in terms of crop loss or cattle being attacked. Increasing human population leading to decreasing forest cover is one of the leading causes for rise in human animal conflicts. Current methods to reduce such conflicts include installation of electric fences or have sentries watch for animals through the night. Electric fences cause severe injury to animals. Moreover, they require enormous initial investment and additionally have high maintenance costs. A recent development in the field of computer science enables use of technology to create low-cost solutions to such problems.

Computer vision is one such technology which could potentially solve most of the associated problems. Uses of deep learning methods to classify images that contain entities of in-terest are gaining popularity. Deep Convolutional Neural Networks (DCNNs) are known to be accurate, and outperform all other existing methods in the task of image classification. Krizhevsky et al., who submitted the winning entry for the ImageNet classification challenge, introduced a Deep Neural Network (DNN) based solution for image classification. It is now considered a landmark achievement in computer vision, and has contributed to increased research in the field of object detection. The main objective of DCNNs is to describe the design for a computer vision system, capable of detecting wild animals and tracking their movement. DCNN's could be leveraged to detect the presence of animals in the captured images. In addition to detecting the presence of an animal, in order to effectively track them and monitor their actions, it is also necessary to localize the animals within the image. This is the task of object detection. Object detection systems predict regions of interest within images, and in addition classify entities within these regions. Thus, object detection is the ideal choice for the system proposed in this project. This method introduces a novel method of reducing human animal conflicts, through constant and automatic monitoring of vulnerable areas using a system of cameras. The proposed solution is accurate and cost effective and to an extent, can be customized specifically for a particular region.

1.2. Artificial intelligence

Artificial intelligence (AI) is the ability of a computer program or a machine to think and learn. It is also a field of study which tries to make

computers "smart". As machines become increasingly capable, mental facilities once thought to require intelligence are removed from the definition. AI is an area of computer sciences that emphasizes the creation of intelligent machines that work and reacts like humans. Some of the activities computers with artificial intelligence are designed for include: Face recognition, Learning, Planning, Decision making etc., Artificial intelligence is the use of computer science programming to imitate human thought and action by analyzing data and surroundings, solving or anticipating problems and learning or self-teaching to adapt to a variety of tasks.

Machine learning is a subsection of Artificial Intelligence (AI) that imparts the system, the benefits to automatically learn from the concepts and knowledge without being explicitly programmed. It starts with observations such as the direct experiences to prepare for the features and patterns in data and producing better results and decisions in the future. Deep learning relies on the collection of machine learning algorithms which models high-level abstractions in the data with multiple nonlinear transformations.

A deep learning technology works on the artificial neural network system (ANNs). These ANNs constantly take learning algorithms and by continuously increasing the amounts of data, the efficiency of training processes can be improved. The efficiency is dependent on the larger data volumes. The training process is called deep because the number of levels of neural network increases with the time. The working of the deep learning process is purely dependent on two phases which are called the training phase and inferring phase. The training phase includes labeling of large amounts of data and determining their matching characteristics and the inferring phase deals with making conclusions and label new unexposed data using their previous knowledge. Deep-learning is such an approach that helps the system to understand the complex perception tasks with the maximum accuracy.

Deep learning is also known as deep structured learning and hierarchical learning that consists of multiple layers which includes nonlinear processing units for the purpose of conversion and feature extraction. Every subsequent layer takes the results from the previous layer as the input. The learning process takes place in either supervised or unsupervised way by using distinctive stages of abstraction and manifold levels of representations. Deep learning or the deep neural network uses the fundamental computational unit, i.e., the neuron that takes multiple signals as input. It integrates these signals linearly with the weight and transfers the combined signals over the nonlinear tasks to produce outputs. In the "deep learning" methodology, the term "deep" enumerates the

concept of numerous layers through which the data is transformed. These systems consist of very special credit assignment path (CAP) depth which means the steps of conversions from input to output and represents the impulsive connection between the input layer and the output layer.

It must be noted that there is a difference between deep learning and representational learning. Representational learning includes the set of methods that helps the machine to take the raw data as input and determines the representations for the detection and classification purpose. Deep learning techniques are purely such kind of learning methods that have multiple levels of representation and at more abstract level.

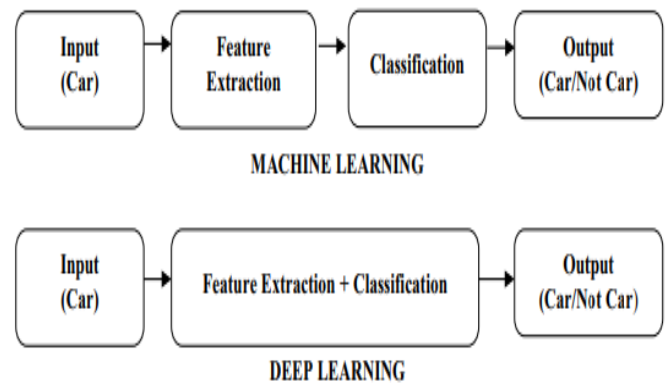


Figure 1. Difference between Machine Learning and the Deep Learning

Fig.1 depicts the differences between the machine learning and deep learning. Deep learning techniques use nonlinear transformations and model abstractions at a high level in large databases. It also describes that a machine transforms its internal attributes, which are required to enumerate the descriptions in each layer, by accepting the abstractions and representations from the previous layer. This novel learning approach is widely used in the fields of adaptive testing, big data, cancer detection, data flow, document analysis and recognition, health care, object detection, speech recognition, image classification, pedestrian detection, natural language processing and voice activity detection. Deep learning paradigm uses a massive ground truth designated data to find the unique features, combinations of features and then constructs an integrated feature extraction and classification model to figure out a variety of applications.

1.3. Object Detection and Tracking

There is a wide range of computer vision tasks benefiting society such as object classification, detection, tracking, counting, Semantic Segmentation, Captioning image, etc. Process of identifying objects in an image and finding its position is known as object detection. Various object detection tasks. With advancements in field of computer vision assisted by AI,

realization of tasks was realizable along t time scale. Semantic segmentation task of clustering pixels based on similarities. Classification + Localization and object detection method of identifying class of object and drawing a bounding box around it to make it distinct. Instance segmentation is semantic segmentation applied to multi objects.

The general intuition to perform the task is to apply CNN over the image. CNN works on image patches to carry out the task many such salient regions can be obtained by Region-Proposal Networks like Region Convolution Neural network (R-CNN), Fast-Region Convolutional Neural Network (Fast R-CNN), Faster-Region Convolutional Neural Network (Faster R-CNN). To perform selective search for object recognition Hierarchical Grouping Algorithm is used. Few bottlenecks by these approaches are mitigated by state-of-the-art algorithms like You Only Look Once (YOLO), Single shot Detector (SSD). The efficient object detection algorithm is one which assures to give bounding box to all objects of vivid size to be recognized, with great computational capabilities, faster processing. YOLO and SSD assure to render promising results, but have a tradeoff between speed and accuracy. Hence, selection of algorithm is application specific.

1.4. Convolutional Neural Networks (CNN)

CNN is widely used neural network architecture for computer vision related tasks. Advantage of CNN is that it automatically performs feature extraction on images i.e., important features are detected by the network itself. CNN is made up of three important components called Convolutional Layer, Pooling layer, fully connected Layer as shown in Fig.2. Considering a gray scale image of size 32×32 would have 1024 nodes in multi-layer approach. This process of flattening pixels loses spatial positions of the image. Spatial relationship between picture elements is retained by learning internal feature representation using small squares of input data.

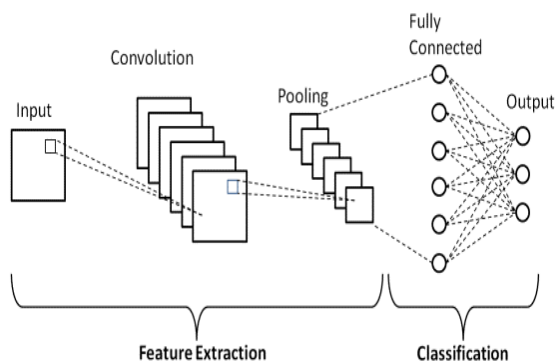


Figure.2. Basic Architecture of CNN

1.4.1. Convolutional layer:

Convolutional Layer encompasses filters and feature maps. Filters are processors of a particular layer. These filters are distinct from one another. They take pixel value as input and gives out feature Map. Feature map is output of one filter layer. Filter is traversed all along the image, moving one pixel at a time. Activation of few neurons takes place resulting in a feature map.

1.4.2. Pooling layer: Pooling layer is employed to reduce dimensionality. Pooling layers are included after one or two convolutional layers to generalize features learnt from previous feature maps. This helps in reducing chances of over fitting from training process.

1.4.3. Fully connected layer: Fully connected layer is used at the end to assign the feature to class probability after extracting and consolidating features from Convolutional Layer and pooling later respectively. These layers use linear activation functions or SoftMax activation function.

1.5. Single Shot Detector (SSD) Algorithm

SSD is a popular object detection algorithm that was developed in Google Inc. It is based on the VGG-16 architecture. Hence SSD is simple and easier to implement. A set of default boxes is made to pass over several feature maps in a convolutional manner. If an object detected is one among the object classifiers during prediction, then a score is generated. The object shape is adjusted to match the localization box. For each box, shape offsets and confidence level are predicted. During training, default boxes are matched to the ground truth boxes. The fully connected layers are discarded by SSD architecture.

The model loss is computed as a weighted sum of confidence loss and localization loss. Measure of the deviation of the predicted box from the ground truth box is localization loss. Confidence is a measure of in which manner confidence the system is that a predicted object is the actual object. Elimination of feature re-sampling and encapsulation of all computation in a single network by SSD makes it simple to train with MobileNets. Compared to YOLO, SSD is faster and a method it performs explicit region proposals and pooling (including Faster R-CNN).

1.6. MobileNets Algorithm

MobileNets uses depth wise separable convolutions that help in building deep neural networks. The MobileNets model is more appropriate for portable and embedded vision-based applications where there is absence of process control. The main objective of MobileNets is to optimize the latency while building small

neural nets at the same time. It concentrates just on size without much focus on speed. MobileNets are constructed from depth wise separable convolutions. In the normal convolution, the input feature map is fragmented into multiple feature maps after the convolution

The number of parameters is reduced significantly by this model through the use of depth wise separable convolutions, when compared to that done by the network with normal convolutions having the same depth in the networks. The reduction of parameters results in the formation of light weight neural network.

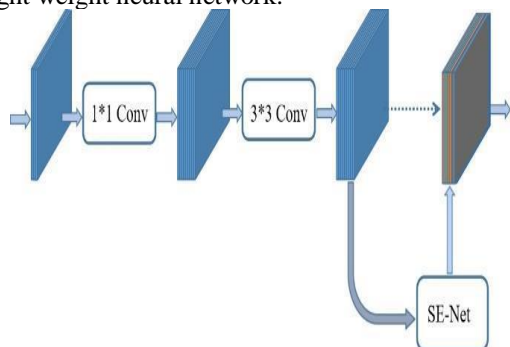


Figure.3 Basic Convolution Structure of MobileNetV3

As shown in Fig.3, MobileNetV3 first uses MnasNet to search for a coarse network structure. Then reinforcement learning algorithm is introduced to select the optimal configuration of model blocks from a set of discrete choices. Finally, MobileNetV3 uses NetAdapt to fine-tune the network architecture, which provides the supplementary information to further boost the robustness of network architecture. Note that, another core idea of MobileNetV3 network is an additional neural network model called "Squeeze-and-Excitation Network" (referred to as SE-Net, which is the champion of 2017 ImageNet Large Scale Visual Recognition Challenge). SE-Net models the relationship between different channels of feature map, and adaptively empowers different weights on the feature map in the channel dimension. According to the importance of task-relevant information, informative channels are weighted a large coefficient, while useless channels are weighted a small coefficient. Through this operation, SE-Net suppresses the useless information and strengthen the informative information of feature maps in the channel dimension.

1.7. You Only Look Once (YOLO)

In this paper, YOLO based model used for effective animal detection. YOLO version 1 and 2 applies soft-max functions convert score into probabilities. This approach is feasible when objects are mutually exclusive only. YOLOv3 employs multi label classification. Independent logistic classifier is used to calculate likelihood of

input belong to a specific label. Loss is calculated using binary-cross entropy of each label. Since we omit the soft-max function complexity is reduced. By using Logistic, regression YOLOv3 predicts the score of presence of object. A ground truth box is defined to all objects, if anchor box overlaps the most with ground truth box, then objectness score is said to be 1. For the anchor boxes whose overlap is greater than the preselected threshold, the anchor box incurs null cost. Every ground truth box is mapped with only one anchor box. If anchor box is not selected and assigned to bounding box then no classification and localization loss is considered, only confidence loss is calculated.

Section 2 presents notable existing research work in the area of animal detection. Section 4 describes the design of the proposed system and highlights the role of various components. Section 6 summarizes the important results of the study, followed by a brief discussion and scope for future enhancements.

2. EXISTING METHOD

2.1. Faster R-CNN Method

Our Existing system describes a system for segmentation of animals from images. The procedure employed uses a multi-level iterative graph cut to generate object region proposals and accurately recognize regions of interest. This is especially useful when the animal blends together with the background and is difficult to identify. These proposals segmented into background and foreground in the second stage. Feature vectors are extracted from each image using AlexNet.

With the surge of deep learning, computer vision has made obvious progress in the recent years. In the context of object detection, deep learning-based methods are in the dominant position and achieve state-of-the-art performance. In 2013, OverFeat firstly applied deep learning methods into object detection community. They propose to use Convolutional Neural Networks (CNNs) to extract image features with a multi-scale sliding window algorithm. After that, Ross Girshick, et al. introduced a R-CNN framework which utilizes Regions of CNN features to extract precise image characteristics. Specifically, in the training process of R-CNN, they have to generate proposals for the dataset in advance. Then they apply the CNN backbone to extract feature maps for every image. Finally, they train a SVM classifiers to identify each object in the image. On the object detection challenge, R-CNN surpasses other methods by near 50% improvement. Then Ross Girshick proposes Fast R-CNN framework, a faster and stronger CNN model for object detection.

Different from R-CNN, Fast R-CNN regards the complete images as inputs and designs

a Selective Search algorithm to generate object proposals, which requires less time for training. Then it introduces Region of Interest (RoI) Pooling to obtain a compact and discriminative feature map for object classification. The feature maps can also be used for bounding box regression. Incorporating the RoI Pooling layer and the following fully connected classification layers, Fast R-CNN can be trained in an end-to-end manner. However, Fast R-CNN heavily relies on Selective Search algorithm (or any other region proposal algorithms), which becomes the bottleneck in the inference phase. Consequently, Faster R-CNN, the newest variety of the R-CNN framework, is proposed to abandon the Selective Search algorithm and it can be trained in an end-to-end manner with the assistant of a Region Proposal Network (RPN). Based on the Faster R-CNN, Xi et al. designed a robust context-aware pedestrian detection method. The pedestrian detection method integrates a de-convolutional module in order to export additional context information to realize effective pedestrian detection. The results show that context information is beneficial to improve the detection accuracy for small-scale pedestrian images. Further, based on Faster R-CNN, Nguyen proposes a novel framework for fast vehicle detection.

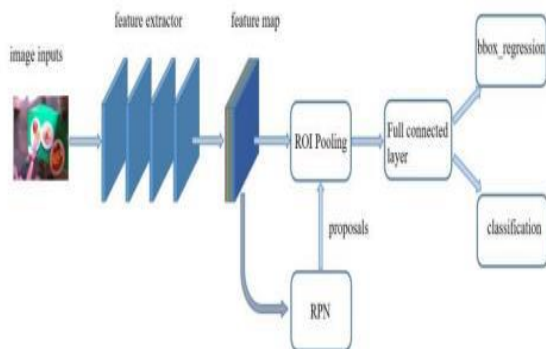


Figure.4. The Architecture of Faster R-CNN

The architecture of Faster R-CNN is illustrated in Fig.4. It regards an image as input and predict the bounding boxes regression and classification results of different objects in the images. Specifically, the input image is firstly passed through a pre-trained CNN backbone (termed feature extractor) to capture the initial feature representation.

It is worth noting that the feature extractor is pre-trained on other datasets to obtain the basic ability of distinguishing different objects, which is a commonly used technique in the context of Transfer Learning. This text will explain the usage of the transfer learning in the later sections. Second, Faster R-CNN applies a Region Proposal Network (RPN) on the initial feature representation to find out the bounding boxes containing different objects. It is an inherent problem for object detection about how to generate a list of bounding boxes. Note that in different scenes, networks

usually have different numbers of bounding boxes. Commonly, the last block of deep neural networks is usually a fixed-sized output vector, which cannot meet the requirement of object detection task. RPN solves the problem by introducing anchor component, i.e., fixed sized bounding boxes for inference. The anchors are uniformly placed throughout the entire original image to guide the search of possible relevant objects. After getting the bounding boxes containing relevant objects, the anchors further introduce the Region of Interest (RoI) Pooling. These modules are encouraged to capture specific features which denote the identity semantic of relevant objects.

Thirdly, the classification loss and regression loss are utilized to jointly optimize the R-CNN model. With the supervisory of category information and position information, R-CNN model can classify the content in the bounding boxes and their positions. Note that background can be regarded as a label, which is discarded in the final results.

2.2. ALGORITHM STEPS FOR EXISTING METHOD

Step 1: Firstly, an input image is taken and passed it to the ConvNet which returns feature maps for the image.

Step 2: Secondly Region Proposal Network (RPN) is applied on these feature maps and we get object proposals.

Step 3: Next ROI pooling layer is applied to bring down all the proposals to the same size.

Step 4: Finally, pass these proposals to a fully connected layer in order to classify any predict the bounding boxes for the image.

2.3. PROBLEM STATEMENT

Wildlife images captured in a field represent a challenging task while classifying animals since they appear with a different pose, cluttered background, different light and climate conditions, different viewpoints, and occlusions. Additionally, animals of different classes look similar. All these challenges necessitate an efficient algorithm for classification.

In this challenge, you will be given 25,000 images of 10 different animal species. Given the image of the animal, your task is to predict the probability for every animal class. The animal class with the highest probability means that the image belongs to that animal class.

3. PROPOSED METHOD:

The proposed method uses the YOLO object detection model to ascertain presence of wild animals in images. The model is fine-tuned for identifying ten different types of animals (Dog, Horse, Butterfly, elephant, Hen, Cat, Cow, Sheep, Spider, and Squirrels.) The system proposed in this

project uses a network of cameras, connected to PIR motion sensors, so that image capture is triggered only when some movement is detected. This enables power conservation. The images captured through these cameras are processed to detect presence of wild animals, and if an animal is found, identify the species. Once identified, the animals are tracked for a suitable time in order to determine their intent – such as to find whether they are moving across the village, or into it. In the latter case, alerts are generated and local authorities are notified through proper channels. Understanding the intent goes a long way to reduce false positives, either due to a false detection or when there is no actual threat posed due to presence of the animal.

3.1 COCO DATASET

The COCO dataset is used for training this model. COCO stands for Common Objects in Context. Images in COCO dataset is taken from everyday scenes thus attaching “context” to the objects captured in the scenes. Let’s say we want to detect a person object in an image. A non-contextual, isolated image will be a close-up photograph of a person. Looking at the photograph, we can only tell that it is an image of a person. However, it will be challenging to describe the environment where the photograph was taken without having other supplementary images that capture not only the person but also the studio or surrounding scene.

As written in the original research paper, there are 91 object categories in COCO. However, only 80 object categories of labelled and segmented images were released in the first publication in 2014. Currently there are two releases of COCO dataset for labelled and segmented images. After the 2014 release, the subsequent release was in 2017. The 80 object category of labelled and segmented images is present in this dataset. The list of 80 classes in the coco dataset are shown :['person', 'bicycle', 'car', 'motorbike', 'aeroplane', 'bus', 'train', 'truck', 'boat', 'traffic light', 'fire hydrant', 'stop sign', 'parking meter', 'bench', 'bird', 'cat', 'dog', 'horse', 'sheep', 'cow', 'elephant', 'bear', 'zebra', 'giraffe', 'backpack', 'umbrella', 'handbag', 'tie', 'suitcase', 'frisbee', 'skis', 'snowboard', 'sports ball', 'kite', 'baseball bat', 'baseball glove', 'skateboard', 'surfboard', 'tennis racket', 'bottle', 'wine glass', 'cup', 'fork', 'knife', 'spoon', 'bowl', 'banana', 'apple', 'sandwich', 'orange', 'broccoli', 'carrot', 'hot dog', 'pizza', 'donut', 'cake', 'chair', 'sofa', 'potted plant', 'bed', 'dining table', 'toilet', 'TV monitor', 'laptop', 'mouse', 'remote', 'keyboard', 'cell phone', 'microwave', 'oven', 'toaster', 'sink', 'refrigerator', 'book', 'clock', 'vase', 'scissors', 'teddy bear', 'hair drier', 'toothbrush']

COCO was an initiative to collect natural images, the images that reflect everyday scene and

provides contextual information. In everyday scene, multiple objects can be found in the same image and each should be labelled as a different object and segmented properly. COCO dataset provides the labelling and segmentation of the objects in the images. This dataset can be used to label and segment the images to create a better performing object detection model.

3.2. SYSTEM ARCHITECTURE

The block diagram of proposed method is shown in Fig.5 The explanation of this block is continued in system modules section.

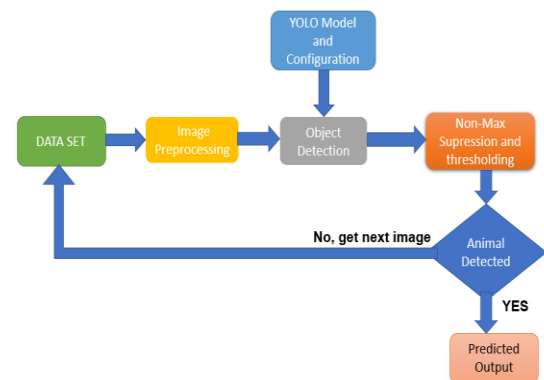


Figure.5 Block Diagram of Proposed Method

3.3. SYSTEM MODULES

- Module 1: Dataset Collection
- Module 2: Pre-processing
- Module 3: Tracking Object
- Module 4: Detection
- Module 5: Accuracy Score Evaluated

Module 1: Dataset Collection

You’re given two types of files (CSV and Images) to download. The train data consists of 25,000 images and the test data consists of 267 images of 10 different species of animals. The image ID and the corresponding animal name are stored in .csv format, while the image files are sorted into separate train and test image folders. Data in the .csv file is in the following format as shown in Table.1:

Table.1 Variable and Description of Given Animal Images

Variable	Description
Image_id	Image name
Animal	Name of the animal

The following are the 10 different species of animals in the dataset:

- Dog
- Horse
- Butterfly
- Elephant
- Hen
- Cat
- Cow
- Sheep
- Spyder
- Squirrels

Module 2: Pre-processing

Once the data is extracted from the twitter source as the datasets, this information has to be passed to the classifier. The classifier cleans the dataset by removing redundant data like stop words, emoticons in order to make sure that non textual content is identified and removed before the analysis.

The `sklearn.pre-processing` package provides several common utility functions and transformer classes to change raw feature vectors into a representation that is more suitable for the downstream estimators.

In general, learning algorithms benefit from standardization of the data set. If some outliers are present in the set, robust scalers or transformers are more appropriate. The behaviors of the different scalers, transformers, and normalizers on a dataset containing marginal outliers are highlighted in Compare the effect of different scalers on data with outliers.

Module 3: Tracking Object

Internet is the main network connecting millions of people in world. Main entertainment factor and the source of greater knowledge is image. Video is collection of frames. The negligible time gap between frames makes the stream of photos looks like flow of scenes. When designing algorithm for video processing. Videos are classified into two classes. Video stream is an ongoing process for video analysis. The processor is not aware of future frames. Video sequence is video of fixed length. All the consecutive frames are obtained prior to processing of current frame. Motion is distinct factor that differentiates video from frame. Motion is a powerful visual Cue. Object properties and action can be realized by noticing only sparse points in the image.

Module 4: Detection

Frames are captured from camera at regular intervals of time. Difference is estimated from the consecutive frames. Optical Flow This technique estimates and calculates the optical flow field with algorithm used for optical flow. A local

mean algorithm is used then to enhance it. To filter noise a self-adaptive algorithm takes place. It contains a wide adaptation to the number and size of the objects and helpful in avoiding time consuming and complicated pre-processing methods. Background Subtraction Background subtraction (BS) method is a rapid method of localizing objects in motion from a video captured by a stationary camera. This forms the primary step of a multi-stage vision system. This type of process separates out background from the foreground object in sequence in images.

Module 5: Accuracy Score Evaluated

The trained model using deep learning must be evaluated for its performance on unseen data called as test dataset. The choice of performance metrics will influence the analysis of algorithms. This helps in identifying reasons for misclassifications so that it can be corrected by taking necessary measures.

1) Confusion Matrix:

It gives prediction information of various objects for binary classification

2) Accuracy and Loss:

Accuracy measure is calculated by using formula. The accuracy measure, as a stand-alone measure is not reliable since it gives equal costs for both type of errors and works well for a well-balanced dataset. The loss is calculated by loss functions of used for training, and average of the loss is calculated when used batch learning that computes loss after each training each batch.

3) Precision and Recall:

Precision is the percentage of classification results that are relevant. Recall is the percentage of total relevant results that are classified correctly by algorithm. Precision and recall values must be maximized to make the model better.

3.3. ARCHITECTURE OF YOLOV3 MODEL

The Fig.6 represents the block diagram of yolov3 model. YOLOv3 uses a variant of Darknet, which originally has 53-layer network trained on ImageNet. For the task of detection, 53 more layers are stacked onto it, giving us a 106 layer fully convolutional underlying architecture for YOLOv3. Here is how the architecture of YOLO now looks like. The newer architecture boasts of residual skip connections, and up-sampling. The most salient feature of v3 is that it makes detections at three different scales. YOLO is a fully convolutional network and its eventual output is generated by applying a 1 x 1 kernel on a feature map. In YOLO v3, the detection is done by applying 1 x 1 detection kernels on feature maps of three different sizes at three different places in the network.

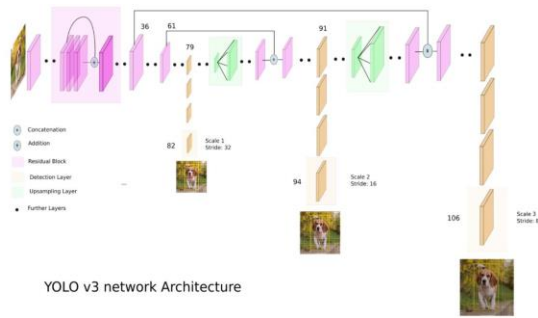


Fig.6. Architecture of YOLOv3 Model

The shape of the detection kernel is $1 \times 1 \times (B \times (5 + C))$. Here B is the number of bounding boxes a cell on the feature map can predict, “5” is for the 4 bounding box attributes and one object confidence, and C is the number of classes. In YOLO v3 trained on COCO, $B = 3$ and $C = 80$, so the kernel size is $1 \times 1 \times 255$. The feature map produced by this kernel has identical height and width of the previous feature map, and has detection attributes along the depth.

Image Grid. The Red Grid is responsible for detecting the dog

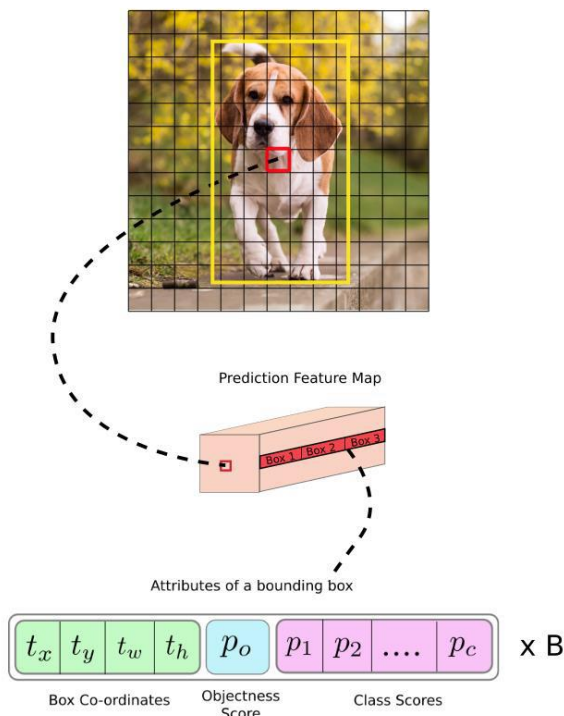


Figure.7 Detection of Objects in Each Scale

Detection of objects in each scale is shown in the Fig.7. The stride of the network, or a layer is defined as the ratio by which it down-samples the input. In the following examples, let us assume we have an input image of size 416×416 .

YOLO v3 makes prediction at three scales, which are precisely given by down-sampling the dimensions of the input image by 32, 16 and 8 respectively.

The first detection is made by the 82nd layer. For the first 81 layers, the image is down sampled by

the network, such that the 81st layer has a stride of 32. If we have an image of 416×416 , the resultant feature map would be of size 13×13 . One detection is made here using the 1×1 detection kernel, giving us a detection feature map of $13 \times 13 \times 255$.

Then, the feature map from layer 79 is subjected to a few convolutional layers before being up sampled by 2x to dimensions of 26×26 . This feature map is then depth concatenated with the feature map from layer 61. Then the combined feature maps is again subjected a few 1×1 convolutional layers to fuse the features from the earlier layer (61). Then, the second detection is made by the 94th layer, yielding a detection feature map of $26 \times 26 \times 255$.

3.4 ALGORITHM OF YOLOv3 MODEL

A similar procedure is followed again, where the feature map from layer 91 is subjected to few convolutional layers before being depth concatenated with a feature map from layer 36. Like before, a few 1×1 convolutional layers follow to fuse the information from the previous layer (36). We make the final of the 3 at 106th layer, yielding feature map of size $52 \times 52 \times 255$. By this method the detection is done in each scale of the given model.

These are the algorithms steps that are to be carried out in this model.

Step 1: The images are down-sampled and strides are used to convert image into $S \times S$ grids.

Step 2: Then a 1×1 kernel is applied on this output image of each layer. The output kernel is called as feature map.

The shape of the kernel for detection is a $1 \times 1 \times (B \times (A+C))$,

Where B = Number of bounding boxes,

C = Number of classes

A =4 bounding box attributes (height, width, x and y coordinates) + Objectness score (P_0)

Step 3: Then pre-defined anchor boxes are used to each cell of the output feature map. 3 anchor boxes are used at each scale and totally 9 anchor boxes are used.

Step 4: These three anchor boxes applied on each cell gives three predefined bounding boxes. Each bounding box contains $(A+C)$ attributes.

Where A = 4 bounding box attributes+ Objectness score(P_0)

C = Number of classes

Step 5: Then it identifies the cell that falls into center of the object by its objectness score. This cell is responsible for detecting the object. This cell gives one ground truth bounding box.

Center cell objectness score = 1

Step 6: We then calculate the bounding box score of the 3 bounding boxes of center cell.

For this we compute element wise product of objectness score and list of confidences.

BB1 score = $P_0 * [P_1 P_2 :: P_c] = PP_1$ (Bounding box score for BB1)

BB2 score = $P_0 * [P_1 P_2 :: P_c] = PP_2$ (Bounding box score for BB2)

BB3 score = $P_0 * [P_1 P_2 :: P_c] = PP_3$ (Bounding box score for BB3)

From this we find the maximum probability by non-max suppression and thresholding and take that class for that object.

Step 7: Next center coordinates, height, width of bounding box is calculated. The Fig.8 shows us the anchor box and predicted bounding box for the given object in the image. By this we can calculate the center coordinates height and width of bounding box.

The Center coordinates of BB (x,y),

$$bx = \sigma(tx) + cx$$

$$by = \sigma(ty) + cy$$

The Width of Bounding box, $bw = pwetw$

The Height of Bounding box, $bh = pheth$

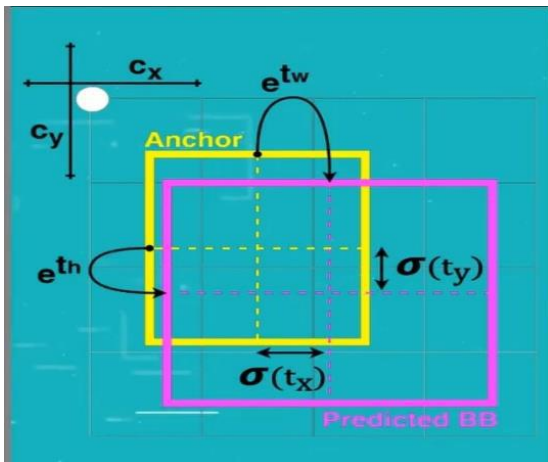


Figure.8. Anchor box and Predicted Bounding Box of Object

Step 8: Then the output of the first detected layer is up-sampled and concatenated with the previous layers so that the input of next layer will have the image same as input.

Step 9: These steps are carried out for the three detection layers of this model.

Step 10: Finally, we get an output with a bounding box around the animal and the class of the animal specified with its probabilities.

3.5. FLOWCHART OF YOLOV3 MODEL

The Fig.9 shows the flowchart of the steps that are carried out in this model.

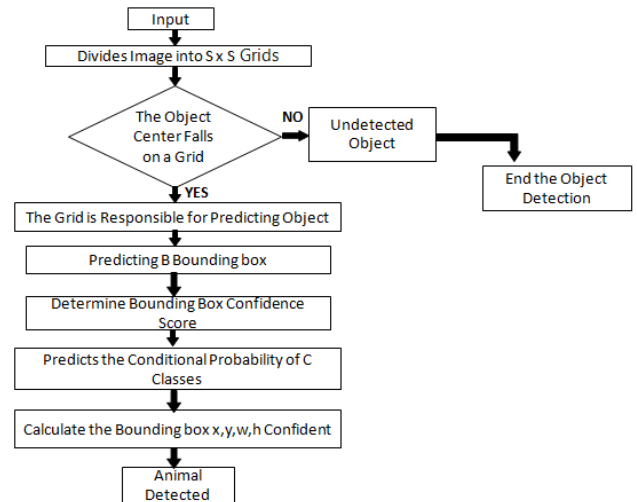


Figure.9: Flow chart of Proposed YOLOv3 Model

4. RESULTS AND DISCUSSION

The dataset here is used for both training and testing of given model. We use two different datasets for training and testing. The dataset images for training and testing with its classes and total number of images are specified below.

4.1 Dataset of Training Images:

Here we use dataset of 10 different categories of animals. The categories of animals and the number of images used for each class are shown in the Table 2. In the Table 2 represents the classes of the images and number of Images used for training.

Table 2: Dataset for Training Images

S. N O	IMAGE CLASS	NO. OF IMAGES
1.	Dog	4863
2.	Horse	2623
3.	Elephant	1446
4.	Butterfly	2112
5.	Hen	3098
6.	Cat	1668
7.	Cow	1866
8.	Sheep	1820
9.	Spider	4821
10.	Squirrel	1862
TOTAL		25,679

4.2 Dataset of Testing Images:

The same animal classes which are used for training is also used for testing. We use dataset of 10 different categories of animals for testing. The categories of animals and the number of images used for each class are shown in the Table 3. Table 3 represents the classes of the images and no. of Images used for testing.

Table 3: Dataset for Testing Images

S.NO	IMAGE CLASS	NO. OF IMAGES
1.	Dog	23
2.	Horse	27
3.	Elephant	25
4.	Butterfly	26
5.	Hen	22
6.	Cat	25
7.	Cow	26
8.	Sheep	35
9.	Spider	25
10.	Squirrel	33
TOTAL		267

4.3 DETECTED ANIMAL IMAGES:

The Figure 10 represents us the detection of animals in the image with its probability of detection.



Fig 10 (a): Detected Sheep Image



Fig 10(b): Detected Dog Image



Fig 10(c): Detected Multiple Pet Image



Fig 10 (d): Detected Cat Image

4.3. PERFORMANCE PARAMETERS

The Table 4 is called as the confusion matrix. The performance parameters are calculated based on this matrix.

Table 4: Confusion Matrix for Testing Images

		PREDICTED CLASS	
		DOG	NOT A DOG
ACTUAL CLASS	DOG	TRUE POSITIVE (TP)	FALSE NEGATIVE (FN)
	NOT A DOG	FALSE POSITIVE (FP)	TRUE NEGATIVE (TN)

- Precision = $(TP)/(TP+FP)$
- Recall = $(TP)/(TP+FN)$
- Accuracy = $(TP+TN)/(TP+TN+FP+FN)$

The Table 5 represents the performance parameters of the detection of images.

Table 5.: Evaluation of Performance Parameters Values

Total No. of Images	TP	TN	FP	FN
23	17	3	2	1

The Table 6 represents the value of precision, recall and accuracy.

Table 6: Performance parameter of YOLOv3

Precision(%)	89.4
Recall(%)	94.4
Accuracy(%)	86.9

4.4. COMPARISON OF FASTER R-CNN AND YOLOV3

The Table 7 shows the comparisons of the precision, accuracy and recall between the existing and proposed system. Here we can observe that YOLOv3 has better performance of precision, recall and accuracy than faster R-CNN.

Table 7: Comparison of Faster R-CNN and YOLOv3

PARAMETERS	Faster R-CNN	YOLOv3
Precision(%)	82.5	89.4
Recall(%)	85.3	94.4
Accuracy(%)	84.6	86.9

4.5 Comparison of Accuracy:

The Fig 11 shows the comparison of accuracy between faster R-CNN and yolov3.

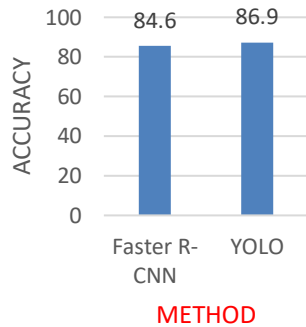


Fig 11. Accuracy Comparison of Faster R-CNN and YOLOv3

4.6 PRECISION vs RECALL GRAPH

The Fig 12 shows us the relation between precision and recall.

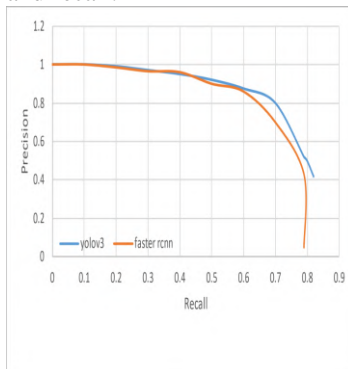


Fig 12. Precision and Recall comparison of Faster R-CNN and YOLOv3

This comparison shows that the yolov3 performs better than the previous detection methods such as Faster R-CNN. The proposed method achieves an accuracy of 85.8% and 90.8% to detect animals respectively.

5. CONCLUSION AND FUTURE WORK

The proposed system attempts to reduce human-animal conflicts by continuous and automatic monitoring of vulnerable areas using computer vision to detect animal intrusions. The intrusion detection pipeline consists of three stages –animal detection, animal tracking and user alerts and notifications. The proposed system is cost-effective and highly efficient, with an average accuracy of 86.9% in detecting and identifying

animal images. Although the prototype described in this paper is trained to recognize five different species of animals, it is easily extendable to detect and track other types of animals with sufficient training data. The choice of species can also be region specific, thereby providing a unique edge over other existing solution. Such a system if implemented on a large scale, has potential to largely reduce casualties due to animal intrusions.

The object detection module is highly accurate. DCNN models for image classification and object detection are widespread in use, and it is evident that given sufficient training data, the models can generalize well in most domains. Similarly, the CSRT tracker is robust and reduces the need for continuous object detection, which is costly and compute intensive. This is especially advantageous, given the use of embedded devices like the Raspberry Pi. The notification system can be customized to dispatch messages using multiple protocols, such as SMS or e-mail. The action taken on animal detection can vary, and could include use of deterrents such as flashing bright lights or playing loud sounds, based on the animal species. The YOLO object detection model is known for its accuracy and ease of use. However, running object detection on embedded devices remains a challenge. A faster and more resource optimal alternative for object detection could be explored. A recent development to create networks specific to mobile devices, such as the MobileNet architecture holds promise, and is a potential candidate to be used for object detection. Another alternative is to use a GPU device, but this would reduce cost-effectiveness of the solution.

YOLO can be programmed for any image domain, expanding the scope of its application. YOLO's technology under goes a research for driverless cars and it can also be used for cancer identification and research. YOLOv3 can also be used in development of responsive robotic systems. In the future this model can be improved to achieve higher accuracy and faster calculation speed. Dataset can also be expanded to more realistic scene images. In addition, the CSRT tracker is a single object tracker – it bears no semantic notation of the object being tracked, and uses visual features to keep the track lets continuous. It is thus, prone to failure if the background closely resembles the appearance of the animal. A more robust tracking mechanism is required, which considers not only visual features but also temporal and spatial features and can effectively track the animal under various conditions.

Use of infrared imagery is yet another area that offers room for improvement. In the proposed system, if the ambient light is not sufficient to capture a reliable image, object detection would fail. Since animal movement generally occurs during the night, use of IR images to detect animals

would make the intrusion detection system more potent, offering a round-the-clock monitoring mechanism.

REFERENCES

1. Caja, Gerardo, J. J. Ghirardi, M. Hernández-Jover, and D. Garín. "Diversity of animal identification techniques: From 'fire age' to 'electronic age'." *ICAR Technical Series* 9 (2004): 21-39.
2. Donovan, John, and Patricia Brown. "Animal identification." *Current protocols in immunology* 73, no. 1 (2018): 1-5.
3. Eradus, Wim J., and Mans B. Jansen. "Animal identification and monitoring." *Computers and Electronics in Agriculture* 24, no. 1-2 (2019): 91-98.
4. Voulodimos, Athanasios S., Charalampos Z. Patrikakis, Alexander B. Sideridis, Vasileios A. Ntafis, and Eftychia M. Xylouri. "A complete farm management system based on animal identification using RFID technology." *Computers and electronics in agriculture* 70, no. 2 (2020): 380-388.
5. Kellenberger, B., Volpi, M., Tuia, D.: Fast animal detection in uav images using convolutional neural networks. In: 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 866–869 (2017)
6. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: F. Pereira, C.J.C. Burges, L. Bottou, K.Q. Weinberger (eds.) *Advances in Neural Information Processing Systems* 25, pp. 1097–1105. Curran Associates, Inc. (2012)
7. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C.: Ssd: Single shot multibox detector. *Lecture Notes in Computer Science* pp. 21–37 (2016)
7. Lowe, D.G.: Distinctive image features from scale-invariant key points. *International Journal of Computer Vision* 60(2), 91–110 (2004)
8. Lukešič, A., Vojtíš, T., ČehovinZajc, L., Matas, J., Kristan, M.: Discriminative correlation filter tracker with channel and spatial reliability. *International Journal of Computer Vision* 126(7), 671–688 (2018)
9. Matuska, S., Hudec, R., Benco, M., Kamencay, P., Zachariasova, M.: A novel system for automatic detection and classification of animal. In: 2014 ELEKTRO, pp. 76–80 (2014)
10. Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M.S., Packer, C., Clune, J.: Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences* 115(25), E5716–E5725 (2018)
11. Parham, J., Stewart, C., Crall, J., Rubenstein, D., Holmberg, J., Berger-Wolf, T.: An animal detection pipeline for identification. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1075–1083 (2018)
12. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788 (2016).
13. De Vasconcellos, Bruno Campos, et al. "Method Applied To Animal Monitoring Through UAV Images." *IEEE Latin America Transactions* 18(07) (2020): 1280-1287.
14. Ren, S., He, K., Girshick, R., & Sun, J. . Faster R-CNN: Towards REAL-TIME Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 39(6), 1137-1149 (2017).
15. W. Hu, Y. Huang, L. Wei, F. Zhang, and H. Li, "Deep convolutional neural networks for hyperspectral image classification," *Journal of Sensors*, vol. 2015, pp. 1-12, Jan. 2015, Art. no. 258619, DOI: 10.1155/2015/258619.
16. C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in *Proc. of the 26th International Conference on Neural Information Processing Systems*, vol. 2, Lake Tahoe, Nevada, USA, 2013, pp. 2553-2561, DOI: 10.5555/2999792.2999897.
17. Zhang, Z., He, Z., Cao, G., & Cao, W. (2016). Animal Detection From Highly Cluttered Natural Scenes Using Spatiotemporal Object Region Proposals and Patch Verification. *IEEE Transactions on Multimedia*, 18(10), 2079–2092. doi:10.1109/tmm.2016.2594138.
18. Suganthi, N., N. Rajathi, and M. Inzamam. "Elephant intrusion detection and repulsive system." *International Journal of Recent Technology and Engineering (IJRTE)* 7.4S (2018).
19. Parham, J., Stewart, C., Crall, J., Rubenstein, D., Holmberg, J., & Berger-Wolf, T. (2018, March). An animal detection pipeline for identification. In 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (pp. 1075-1083). IEEE.

FACE MASK DETECTION USING TENSORFLOW, RESNET, AND MACHINE LEARNING ALGORITHM

Mrs. S. Suriya¹, Agusthiyar R², Shyamala Devi J³, S. ABISHEK⁴, R.HARISH⁵, ANTOS MARIA KARUNAI⁶

^{1,3}Assistant Professor, Department of Computer Applications, SRM IST, Ramapuram Campus.

²Professor & Head, Department of Computer Applications, SRM IST, Ramapuram Campus

^{4,5,6}Final year BCA Students, Department of Computer Applications, SRM IST, Ramapuram Campus

Abstract

The new coronavirus has brought about a new normal life in which keeping a social distance and wearing a face mask plays an important role in controlling the spread of the virus. However, most people do not wear face masks in public places, which increases the spread of the virus. This can lead to the serious problem of increased propagation. Therefore, in order to avoid such a situation, it is necessary to be aware of wearing a face mask. Humans cannot participate in this process because they can be affected by the corona. Therefore, here is the need for artificial intelligence (AI), which is the main topic of our project. Our project uses image processing and AI technology to send alerts to officials to identify who wears and who does not wear face masks in public places. Object detection algorithms are used to distinguish between people with and without face masks. This also tells you how many people have and do not have face masks. It also uses the Internet of Things (IoT) to send alerts. Alerts are sent to the right people via mobile notifications and email.

Foreword.

1.Introduction

Face mask detection detects that no one else is wearing the mask. Face recognition is reverse-engineered and a variety of algorithms are used to recognize faces for security, authentication, and monitoring. In the domains of computer vision and pattern recognition, face recognition is crucial. Face recognition algorithms have benefited from a lot of study in the past. Face recognition was first studied in 2001 and trained successful classifiers for recognition and recognition using traditional machine learning algorithms and craftsmanship feature design. This method has many drawbacks, including high functional complexity and poor recognition accuracy. In recent years, face recognition technology based on a deep convolutional neural network (CNN) has been widely developed to improve recognition performance. Overview

Deep learning is a major advance in the science of artificial intelligence. Recently, we have shown great potential for image analysis to capture small features. With the outbreak of COVID19, different techniques in deep learning are used for the detection of infected patients of the virus. In this sense, more types of viral lung infections are called viral pneumonia, as opposed to bacterial pneumonia. For example, COVID19 infects the lungs by blocking the flow of oxygen that may be

present. This has led scientists to develop a variety of AI-based frameworks and methods for combating infections. So let's break this topic into two parts and take a closer look at the technique. record

The data listed in the COVID 19 Face Mask Detection Dataset section is located in the dataset/directory. Dynamic image three sample images are provided to test the face mask detector. This topic focuses on three Python scripts. Take the input dataset and adapt MobileNetV2 to generate a mask detector. model in trainmaskdetector.py. In addition, a training history plot.png containing the accuracy/loss curve will be created. Detectmaskimage.py is a Python script that detects the face mask of a still image. detectmaskvideo.py: This script uses the website to detect face masks. Each image in the stream. Methodology

This project uses several techniques and modules such as Tensor flow, Moblienet, Deep Learning, Resnet, etc.

Technology Used

Tensor flow

Tensor Flow is a set of techniques for developing and training models in Python or JavaScript and deploying them to the cloud, on-premises, browsers, or devices, regardless of language. The tf data API allows you to build sophisticated input pipelines with simple, reusable components. It can be used in a variety of applications but focuses on training and inference for deep neural networks. Tensor Flow is compatible with various programming languages such as Python, Javascript, C ++, and Java. This adaptability is useful for a wide range of applications across different industries. Automatic image annotation software is based on TensorFlow. Mobilenet's Mobilenet Mask model,

COVID 19, which causes a major health crisis, is being fought in all countries of the world. You can use a regulated face mask to control the spread of the virus. Dey et al. With this configuration, we suggested a deep learning-based approach for recognizing face masks. The Mobilenet Mask model is a multi-phase model. Search for faces in the video stream using a pre-trained model of the Resnet 10 architecture. The loading of the classifier (mobile network), the creation of the FC layer, and the testing phase are all used. Google Colab, which is operating on Colab with an additional 12GB of RAM, is monitoring all of the experimental scenarios. Model performance is evaluated using a variety of performance criteria (accuracy, F1 score, precision, recall).

Deep Learning

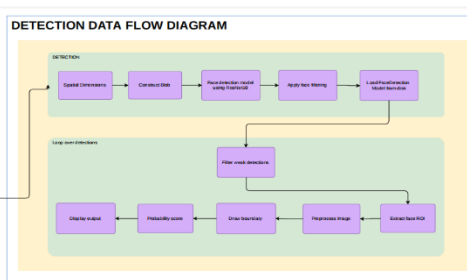
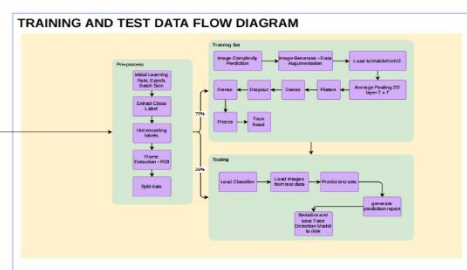
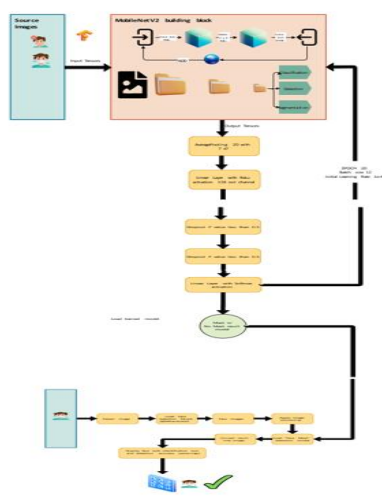
Deep learning is a kind of mechanical machine learning to practice computers to achieve things to achieve what people work instinctively. Home appliances such as mobile phones, tablets, television, and handsfree entries are formed using large amounts of signature and multilayer neuron network topologies. Resnet is an acronym of the residual network. It was a winner of the image net challenge 2015 and served as a backbone for several computer vision applications. It belongs to the reset model family. Overall, there are a total of 48 convolutional layers with the maximum pool layer and the mean pool layer. According to training and verification accuracy and attenuation, the loss is close to zero and the accuracy is about 100%.

ResNet

ResNet is an acronym for Residual Networks. It is a network that serves as the bone of various computer vision applications and was the winner of the 2015 Image Net Challenge. It

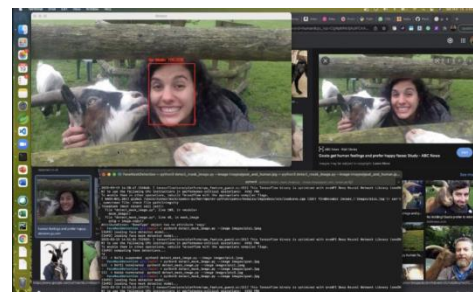
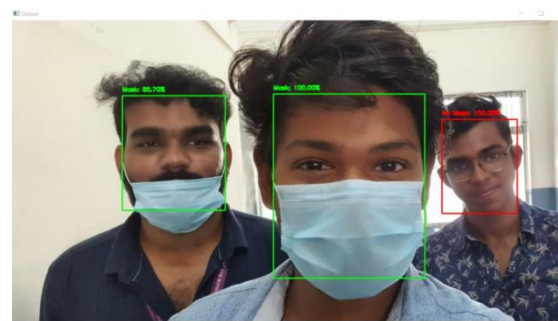
belongs to the ResNet model family. There are a total of 48 convolutional layers with the largest pooling layer and the average pooling layer. The loss is near to zero, and the accuracy is close to 100 percent, according to training and verification accuracy and attenuation. The confusion matrix after testing is depicted in this diagram. When you look at the parameters in Table 3 of the ResNet50 model, you'll notice that all of them are terminated except "support." When the ResNet 50 is compared to the other four models, it is clear that ResNet is the most advanced.

Description of architecture diagram



An architectural diagram is a graphical representation of a set of architectural concepts, including their principles, elements, and components. The architecture diagram contains images of people wearing and not wearing masks used in deep learning and tensors such as mobilenetv2. A dropout value of 0.5 is performed with the activation functions 128OutChannel and Softmax activation. Face images are extracted and loaded from SSD and resnet10, processed to load face mask detection, and converted for verification.

Implementation



Future Enhancements

- First, the proposed method can be integrated into all high-resolution video surveillance devices and is not limited to mask detection only.
- Second, the model can be extended to recognize facial features using face masks for biometric purposes.
- Face mask detection helps screen station and airport passengers, game spectators, and people who come primarily to hospitals and other public places.

Conclusion

Because of the pressing necessity to handle COVID19, real-time masks and social distance detection are becoming more relevant and important. First, this article looked at a number of studies on the COVID 19 epidemic. The essential principle of the deep CNN model was then described. We next used face mask datasets to mimic the training and testing of the most commonly used deep pre-trained CNN models (ResNet, Tensorflow, MobileNet, and DeepLearning). Finally, the top models will be put to the test on a Raspberry Pi board with a webcam embedded vision system. To automate and hurt or maintain a more masked face detection process, efficient real-time deep learning approaches are combined with social distance tasks. People are separated by a distance.

References

- [1] Shilpa Sethi, Mamta Kathuria, Trilok Kaushik, "Face mask detection using deep learning: An approach to reduce risk of Coronavirus spread", *Journal of Biomedical Informatics*, Volume 120, 2021, 103848, ISSN 1532-0464,
- [2] M. M. Rahman, M. M. H. Manik, M. M. Islam, S. Mahmud and J. -H. Kim, "An Automated System to Limit COVID-19 Using Facial Mask Detection in Smart City Network," *2020 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS)*, 2020, pp. 1-5, doi: 10.1109/IEMTRONICS51293.2020.9216386.
- [3] A. Das, M. Wasif Ansari and R. Basak, "Covid-19 Face Mask Detection Using TensorFlow, Keras and OpenCV," *2020 IEEE 17th India Council*

International Conference (INDICON), 2020, pp. 1-5, doi: 10.1109/INDICON49873.2020.9342585.

[4] Sethi S, Kathuria M, Kaushik T. Face mask detection using deep learning: An approach to reduce risk of Coronavirus spread. *J Biomed Inform.* 2021;120:103848.

[5] Gagandeep Kaur, Ritesh Sinha, Puneet Kumar Tiwari, Srijan Kumar Yadav, Prabhaskar Pandey, Rohit Raj, Anshu Vashisth, Manik Rakhra, Face mask recognition system using CNN model, *Neuroscience Informatics*, Volume 2, Issue 3, 2022, 100035, ISSN 2772-5286,

[6] J. Prinosil and O. Maly, "Detecting Faces With Face Masks," *2021 44th International Conference on Telecommunications and Signal Processing (TSP)*, 2021, pp. 259-262, doi: 10.1109/TSP52935.2021.9522677.

[7] P. Hofer, M. Roland, P. Schwarz, M. Schwaighofer and R. Mayrhofer, "Importance of different facial parts for face detection networks," *2021 IEEE International Workshop on Biometrics and Forensics (IWBF)*, 2021, pp. 1-6, doi: 10.1109/IWBF50991.2021.9465087.

IDENTIFYING IMAGES IN MULTIFRAME SEGMENTATION IMAGE CLASSIFICATION USING SVM MODEL IMPLEMENTS WITH OPENCV

¹Mrs.S.Sindhu, ²Mrs.J.Shyamala Devi, ³Mr.B N Swaminathan, ⁴Mr.Sanjay kumar, ⁵Mr.M Sanjai

^{1,2} Assistant Professor, SRM Institute of Science and Technology, Ramapuram, Chennai.

^{3,4,5} Student, SRM Institute of Science and Technology, Ramapuram, Chennai.

ABSTRACT

Image classification is vital field of research in computer vision. Image classification is a supervised learning problem: define a set of target classes (objects to identify in images), and train a model to recognize them using labeled example photos. Early computer vision models relied on raw pixel data as the input to the model. However, raw pixel data alone doesn't provide a sufficiently stable representation to encompass the myriad variations of an object as captured in an image. The position of the object, background behind the object, ambient lighting, camera angle, and camera focus all can produce fluctuation in raw pixel data; these differences are significant enough that they cannot be corrected for by taking weighted averages of pixel RGB values. We identify a person in a photo with a face. Hence we will use opencv and a technique called HAAR cascades to detect if a face and two eyes are clearly visible or not. We crop images and apply wavelet transform to extract meaning features that can help with image identification. We create a SVM model then use GridSearchCV. We will write a flask server that will use the trained model and perform image classification. We will build a website for our project. This website has an area where someone can drag and drop an image of a person and it will identify that person, we will use HTML/CSS/Javascript for this project. JQuery is used to make http calls to python flask backend.

Keywords: *HAAR cascading, SVM, OpenCv, HSV and RGB color code, Noise removal and etc.*

I. Introduction

The practice of identifying and labeling groups of pixels or vectors inside an image using specified rules is known as image classification. Although this is not a particularly challenging challenge for people[12], it has proven to be a particularly difficult problem for machines[5]. Variable and even uncontrollable imaging settings, as well as difficult-to-describe and complex angles in an image, are major causes of difficulty.

In our project, we used SVM. The Support Vector Machine,[1] or SVM, is a popular Supervised Learning technique that may be used to solve both classification and regression issues[8][9]. However, it is mostly utilized in Machine Learning for Classification difficulties.

The SVM algorithm's purpose is to find the optimum line or decision boundary[8] for categorizing n-dimensional space into classes so that additional data points can be readily placed in the correct category in the future.

A hyper plane is the name for the optimal choice boundary. The extreme points/vectors that assist create the hyper plane are chosen via SVM[3][12]. Support vectors are the extreme instances, and the algorithm is called a Support Vector Machine.

II .LITERATURE REVIEW

The traditional face detection approach is based mostly on the face's structural traits and color characteristics. Traditional face recognition algorithms extract landmarks, or features, from an image of the subject's face to identify facial traits[18].

The relative position, size, and/or shape of the eyes, nose, cheekbones, and jaw may be analyzed by an algorithm[11]. These features are then utilized to find other photos that have similar features[21]. These algorithms can be difficult to implement and demand a lot of computing power, therefore they may be slow. They can also be inaccurate when the faces have strong emotional expressions, as the size and position of the landmarks can be drastically altered.

III. Main Objectives of Proposed work

- Because of the limited hardware available, processing large amounts of data is difficult.
- Difficulty in interpreting the model due to its ambiguous character, which prevents it from being used in a variety of situations.
- Development takes longer, and as a result, flexibility suffers as a result of the lengthier development time.

IV. Proposed Work

Image classification is a hot research topic in the field of computer vision, as well as the fundamental image classification system in other image application domains, and it is usually separated into the following parts: Pre-processing of data: Data collection and cleaning, feature engineering and model training, website development, and Flask server

Any format of the digital image can be made available (improved image, X-Ray, photo negative, etc). It aids in the enhancement of images for human understanding. Images can be analyzed and information retrieved for machine interpretation. The density and contrast of the pixels in the image can be adjusted to any desired level. Images can be simply stored and accessed. It makes it simple to send photos to third-party suppliers via the internet.

V. Project Architecture

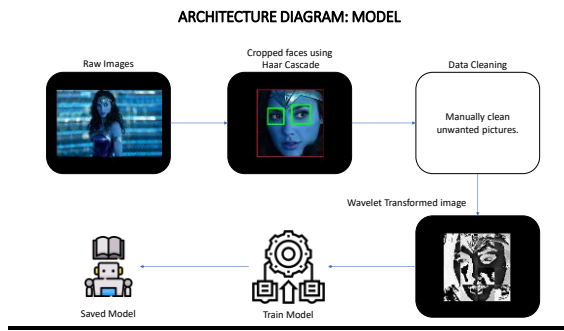


Fig 1 Architecture of the proposed work

To automate the download of photos from Google, use Python and web scraping.

Cleaning Data: We must clean photos that we have downloaded from Google or that we have taken so that they may be used to train our classifier. With the help of a face, we can recognize someone in a photograph. As a result, we'll utilize openCV and a technique known as HAAR cascades to determine whether or not a face and two eyes are clearly apparent. If they are, the image is kept; otherwise, it is discarded.

Feature Engineering: We'll use cropped images and the wavelet transforms to extract meaningful features that can aid image recognition. Concepts such as time vs. frequency domain, Fourier transform, expressing images as frequency, and so on can be used. We'll make our X using the wavelet transform and a raw pixel image, and our Y will be class labels. Model training will be done with these X and Y.

VI .Data Flow Diagram

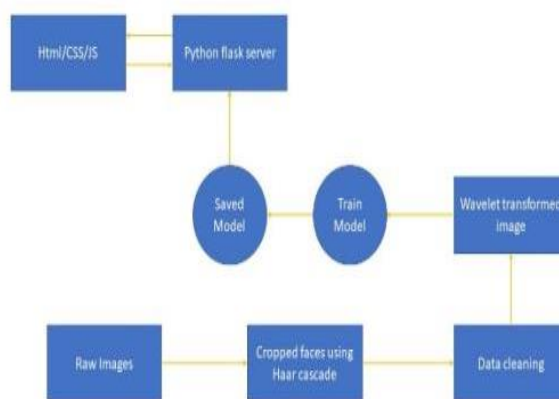


Fig 2 Data flow diagram for the proposed work

VII .Dataset

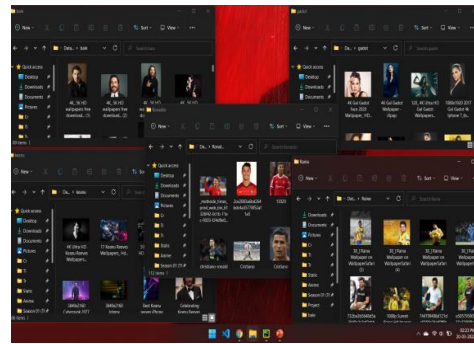


Fig 3 Dataset sample review

This system does not see data in the same manner that humans do. As a result, it's critical to pre-process the data by cleaning and finishing it, as well as annotate it with useful tags, after it's been collected. readable by a computer. In our project we use 5 celebrities as our dataset.

VIII. Implementations

Step1 Way of data collection: Manually download photos from Google Images or other sources. To automate the download of photos from Google, use Python and web scraping.

Step 2 Data Cleaning: We must clean photographs that we have downloaded from Google or that we have taken so that they may be used to train our classifier. With the help of a face, we can recognize someone in a photograph. As a result, we'll utilize opencv and a technique known as haar cascades to determine whether or not a face and two eyes are clearly apparent. If they are, the image is kept; otherwise, it is discarded.

Step 3 Feature Engineering: Using cropped photos and the wavelet transform, we will extract meaningful features that can aid in image recognition. Concepts such as time vs. frequency domain, fourier transform, expressing images as frequency, and so on can be used. We'll make our X using the wavelet transform and a raw pixel image, and our Y will be class labels. Model training will be done with these X and Y.

IX Screen shots

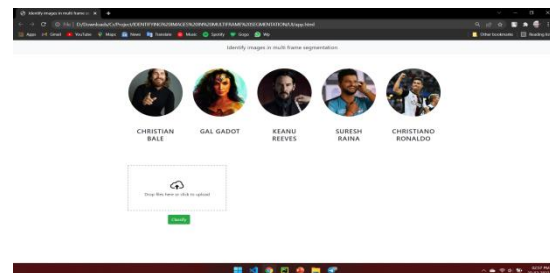


Fig 4 screen shot 1 for the home page

The home page of an application to load data from an external site is shown in the diagram above.

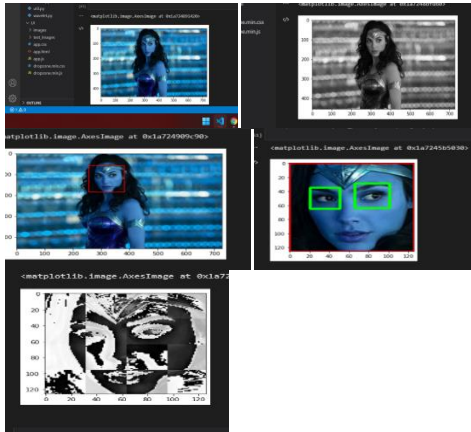


Fig 5 image identification from the trained models.

The image is first converted to grayscale, and then the colour contrast is adjusted to prevent the colour code problem. Using the haar cascading procedure, locate the region of the face from the frame set after the correction. Using the CNN method, locate the eye from the face region.

X. Result and Analysis

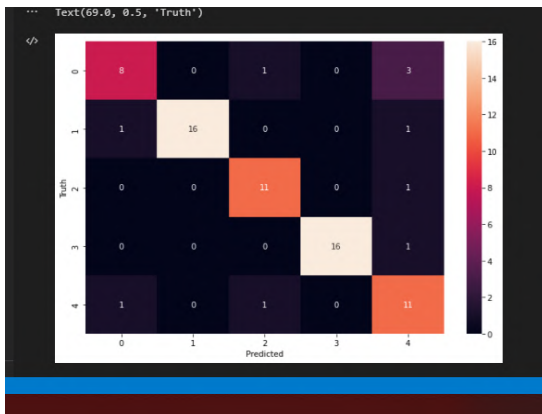


Fig 6 correlation co-efficient value for the above calculated result.

The image can be recognized because of the large region area. Then, from the given image, extract the data. After that, find the correlation to determine the data's accuracy. As a result of an issue, the maximum score will be considered.

XI. CONCLUSION

The techniques used in this paper resulted in an increase in classification accuracy, providing positive outcomes. This work can be expanded in the future to include improved methodologies. Deep learning can be used to improve project accuracy and progress.

XII References

1Sandeep Kumar¹, Zeeshan Khan², Anurag jain³. Content Based ImageClassification using Machine Learning Approach
 2JoséM. Peña^{1,2,*}, Pedro A. Gutiérrez³, César Hervás- Martínez³, Johan Six⁴, Richard E.Plant² and Francisca López- Granados¹. Object-Based Image Classification of Summer Crops with

Machine Learning Methods.

.N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection", Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR05), pp. 886-893, 2005.

3.R. Li, D. Gu, Q. Liu, Z. Long and H. Hu, "Semantic scene mapping with spatio-temporal deep neural network for robotic applications", Cognit. Comput., vol. 10, pp. 260-271, Apr. 2018.

4.I. Kostavelis and A. Gasteratos, "Learning spatially semantic representations for cognitive robot navigation", Robot. Auto. Syst., vol. 61, no. 12, pp. 1460-1475, Dec. 2013.

5.A. Krizhevsky, I. Sutskever and G. Hinton, "Imagenet classification with deep convolutional neural networks", Proc. NIPS, pp. 1106-1114, 2012.

6.L.-J. Li, H. Su, Y. Lim and L. Fei-Fei, "Object bank: An object-level image representation for high-level visual recognition", Int. J. Comput. Vis., vol. 107, no. 1, pp. 20-39, Mar. 2014.

7.D. Lin, S. Fidler and R. Urtasun, "Holistic scene understanding for 3D object detection with RGBD cameras", Proc. IEEE Int. Conf. Comput. Vis., pp. 1417-1424, Dec. 2013.

8.D. G. Lowe, "Object recognition from local scale-invariant features", Proc. 7th IEEE Int. Conf. Comput. Vis., pp. 1150-1157, 1999.

9.M. Naseer, S. Khan and F. Porikli, "Indoor scene understanding in 2.5/3D for autonomous agents: A survey", IEEE Access, vol. 7, pp. 1859-1887, 2019.

10.S. O'Hara and B. A. Draper, "Introduction to the bag of features paradigm for image classification and retrieval", arXiv:1101.3354, 2011, [online]

11.A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope", Int. J. Comput. Vis., vol. 42, no. 3, pp. 145-175, 2001.

12.A. Pronobis and P. Jensfelt, "Large-scale semantic mapping and reasoning with heterogeneous modalities", Proc. IEEE Int. Conf. Robot. Autom., pp. 3515-3522, May 2012.

13.A. Ranganathan and F. Dellaert, "Semantic modeling of places using objects", Robotics: Science and Systems III, pp. 1-8, 2007.

14.J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: Unified real-time object detection", Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 779-788, Jun. 2016.

15.S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks", IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137-1149, Jun. 2017.

16.N. Silberman, D. Hoiem, P. Kohli and R. Fergus, "Indoor segmentation and support inference from RGBD images", Proc. ECCV, pp. 746-760, 2012.

- 17.J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers and A. W. M. Smeulders, "Selective search for object recognition", *Int. J. Comput. Vis.*, vol. 104, no. 2, pp. 154-171, Sep. 2013.
- 18.P. A. Viola and M. J. Jones, "Rapid object detection using a boosted cascade of simple features", *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, vol. 1, pp. 511-518, Dec. 2001.
- 19.H. Zender, O. Martínez Mozos, P. Jensfelt, G.-J.-M. Kruijff and W. Burgard, "Conceptual spatial representations for indoor mobile robots", *Robot. Auto. Syst.*, vol. 56, no. 6, pp. 493-502, Jun. 2008.
- 20.B. Zhou, A. Lapedriza, A. Khosla, A. Oliva and A. Torralba, "Places: A 10 million image database for scene recognition", *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452-1464, Jun. 2018.
- 21.X. Li, "Object detection and scene understanding based on fusing multi-view RGB-D frames", 2019.
- 22.M. Zhai, X. Xiang, R. Zhang, N. Lv and A. El Saddik, "Optical flow estimation using dual self-attention pyramid networks", *IEEE Trans. Circuits Syst. Video Technol.*, 2019.
- 23.M. Zhai, X. Xiang, R. Zhang, N. Lv and A. E. Saddik, "Optical flow estimation using channel attention mechanism and dilated convolutional neural networks", *Neurocomputing*, vol. 368, pp. 124-132, Nov. 2019.
- 24.Z. Yang, P. Wang, Y. Wang, W. Xu and R. Nevatia, "LEGO: Learning edge with geometry all at once by watching videos", *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 225-234, Jun. 2018.
- 25.R. Mahjourian, M. Wicke and A. Angelova, "Unsupervised learning of depth and ego-motion from monocular video using 3D geometric constraints", *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 5667-5675, Jun. 2018.
- 26.J. McCormac, A. Handa, A. Davison and S. Leutenegger, "SemanticFusion: Dense 3D semantic mapping with convolutional neural networks", *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, pp. 4628-4635, May 2017.
- 27.K. He, X. Zhang, S. Ren and J. Sun, "Deep residual learning for image recognition", *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 770-778, Jun. 2016.
- 28.X. Song, L. Herranz and S. Jiang, "Depth CNNs for RGB-D scene recognition: Learning from scratch better than transferring from RGB-CNNs", *Proc. AAAI Conf. Artif. Intell.*, pp. 4271-4277, 2017.
- 29.X. Song, S. Jiang and L. Herranz, "Combining models from multiple sources for RGB-D scene recognition", *Proc. Twenty-Sixth Int. Joint Conf. Artif. Intell.*, pp. 4523-4529, Aug. 2017.
- 30.Y. Li, Z. Zhang, Y. Cheng, L. Wang and T. Tan, "MAPNet: Multi-modal attentive pooling network for RGB-D indoor scene classification", *Pattern Recognit.*, vol. 90, pp. 436-449, Jun. 2019.
- 31.X. Song, S. Jiang, L. Herranz and C. Chen, "Learning effective RGB-D representations for scene recognition", *IEEE Trans. Image Process.*, vol. 28, no. 2, pp. 980-993, Feb. 2019.
- 32.A. Ayub and A. Wagner, "CBCL: Brain-inspired model for RGB-D indoor scene classification", *arXiv:1911.00155*, 2019..

AN ENSEMBLE APPROACH FOR DOCUMENT SUMMARIZATION

Vetrivel Panneerselvam, Srushti Gajbhiye, Selvakuberan Karuppasamy and Subhashini Lakshminarayanan
Data and AI, Advanced Technology Centers in India, Accenture
Email: {v.h.panneerselvam, srushti.gajbhiye,
s.b.karuppasamy, s.j.lakshminarayanan}@accenture.com

Abstract—With the ever-growing data on the internet, there is a huge number of unstructured documents created on a daily basis. They can range from short two-page articles to eight page research papers to full-length theses, containing both text and table content. It is essential to be able to club and process them and generate actionable insights from them. This paper proposes an ensemble summarization model for PDF and Word (.doc and .docx) documents to generate section-wise, page-wise or document-wise summaries, on user demand. It minimizes perusing time, reduces bias and redundant manual effort. We have tested our hypothesis and got promising results on metrics like BLEU score and ROUGE scores in comparison to the four state-of-the-art models.

Index Terms—*Hybrid Document Summarization, Ensemble Model, Abstractive Summarization, Text and Table Extraction, Summary Curation.*

I. INTRODUCTION

Automatic summarization is the process of making a set of data short, computationally to create a subset that represents the very important and most relevant information in the original content, while maintaining the grammar, context and flow of the original content. It is a crucial asset for teams having limited time to go through a large number of documents and generate analysis and insights.

For instance, in insurance and legal domain, there are lots of documents written in complicated language having many jargons and tabular data. It is imperative for the user to be able to understand every resource available and highlight portions relevant to the user. Without the help of an agent or chartered accountant, this task seems tough for a common man. In long documents ranging more than ten pages, there are many sections covering a variety of topics. Often conventional summarization systems will result in a

summary which only covers the most significant portion. This paper proposes an ensemble approach which allows user to generate summary for each chosen section thus allowing personalization and flexibility. It also allows user to generate summary page-wise, which enables user to study a long document in a short time. The document summary of the input document(s) can be utilized as the crux of main summary while the page-wise and section-wise summaries can cater to the custom demands of each user. They provide outlines which aid research investigation and are widely used to create index from lengthy documents. Personalized summaries are helpful in question-answering systems systems. Summaries generated in an automated manner tend to have less bias when put in comparison with human-written summaries.

II. LITERATURE REVIEW

Goldstein, Jade et al. [1] discussed an approach to multidocument summarization that builds on previous work in single-document summarization by using additional, available information about the document set as a whole, the relationships between the documents, as well as individual documents. Bagalkotkar, Anusha et al. [2] presented a novel technique for generating the summarization of domain specific text from a single web document by using statistical NLP techniques on the text in a reference corpus and on the web document. The summarizer proposed generates a summary based on the calculated Sentence Weight (SW), the rank of a sentence in the document's content, the number of terms and the number of words in a sentence and using term frequency in the input corpus.

Prudhvi, Kota et al. [3] talked about basic extractive summarization as an unsupervised learning approach where cosine similarity technique is used to find the similarity between sentences. To generate rank based on similarity,

text rank algorithm is used and sentences with top rank are placed in summarized text.

A comprehensive survey of last decade's text summarization - extractive approaches is presented by Gambhir, Mahak et al. [4]. Munot, Nikita et al. [5] also presented taxonomy of summarization systems and statistical and linguistic approaches for summarization. It discusses different types of summarization methods used for summarizing a document and advantages and disadvantages of each method.

Gajbhiye, Srushti et al. [6] proposed a way to process tabular data from PDF documents, convert it into textual content, summarize it and generated actionable insights. This approach converts tabular content into a form which is much better suit-ed for machine consumption. Gholamrezazadeh, Saeedeh et al. [7] defined the most important criteria for a summary which can be generated by a system while presenting a taxonomy of these systems.

The concept of lexical chain was first introduced by Morris, Jane et. al. [8]. Ko, Youngjoong et al. [9] explained how lexical chains exploit the cohesion among an arbitrary number of related words. Lexical chains are created by grouping set of words that are semantically related.

Sewing all this together and building on top of it, we propose an approach to extract text from documents and summarize it using a hybrid model which can give out a variety of summaries, depending on user's demand.

III. SHORTCOMINGS OF CONVENTIONAL METHODS

To be able to generate automatic summary from documents, it is essential to decide and specify the most important parts of the original text in the document to preserve the originality and readability of the summary in terms of coherence, context, and grammar. Some limitations of the conventional methods for summarization are as follows:

1. Single model is used to generate the summary.
2. Inconsistencies in the summary.
3. Lack of balance in the generated summary.
4. Context of the document may not be entirely preserved in the final summary.

5. Identification and interpretation of important and relevant content in the summary.

IV. PROPOSED APPROACH

We propose an ensemble abstractive document summarization technique, described in Fig. 1, which is a combination of four state-of-the-art models to generate the final summary. Our technique extracts text and tables from PDF and Word documents, applies text cleaning and pre-processing and summarizes pre-processed text in one of the three ways, as demanded by user (page-wise, section-wise, document-wise).

The proposed architecture has four components:

- a. Document Upload Module
- b. Text and Table Extraction
- c. Hybrid Summarization Module
- d. Summary Curation Module

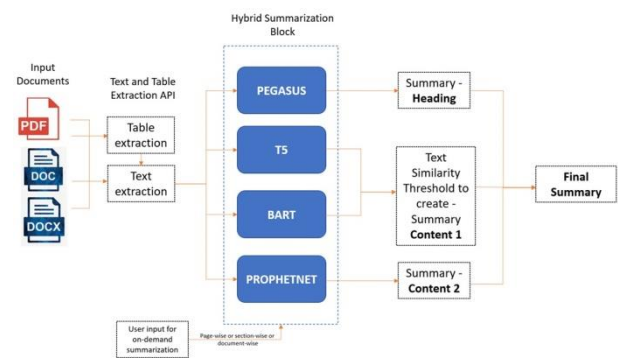


Fig. 1. Ensemble Document Summarization Model Architecture

A. Document Upload Module

Most documents are present in the form of unstructured data in .pdf, .doc and .docx file formats. This module allows user to upload any domain documents ranging from insurance, legal, healthcare, wealth management to annual reports. Once uploaded, it will be passed to the text and table extraction module, to extract all the contents from the text as well as tables.

B. Text and Table Extraction

Based on the uploaded document format, this module uses fitz [10] (for PDF), docx2txt [11] (for .docx word) and antiword [12] (for .doc word) to derive content from it. These packages extract text from the documents and custom

code extracts text from tables and converts it into legible text content. The extraction is done as chosen by user, in form of pages, sections and whole document. It is stored in key-value pairs and is used to generate any type of summary, as demanded by the user.

We have curated a dataset consisting of 35 Documents (12 .pdf, 12 .doc and 11 .docx) ranging from 6 to 15 pages with different types of tables, a variety of images and subsections, as described in Table I. Text and Table extraction module works well in capturing almost all the text in the data present in these complex tables and subsections. Sub-sections are usually hard to capture under the correct section in text documents. This module as part of our proposed model works on identifying the sub-sections in minute-level of detail and captures them precisely under each section. This enhances the accuracy of the final summary when user searches for section-wise summary. It gives great results as both sections and subsections are identified properly by the module.

TABLE I
DOMAIN-WISE DATASET DETAILS

Domain	Type of document	No. of documents	Avg . no. of pages	Tables	Images	Sub-Sections
Insurance	.pdf	3	6	✓		✓
	.docx/.	5	10	✓	✓	✓
	doc				×	
Finance	.pdf	2	8	✓	✓	×
	.docx/.	7	16	✓	✓	✓
	doc					
Technology	.pdf	3	15	✓	×	×
	.docx/.	4	10	✓	✓	✓
	doc					
Healthcare	.pdf	2	7		✓	✓
	.docx/.	5	9	✓	✓	✓
	doc			×		

Telecom	.pdf	2	12		×	
	.docx/.	5	9	✓	×	✓
	doc			×		×

C. Hybrid Summarization Module

This ensemble model is a combination of four state-of-the-art models – PEGASUS, T5, BART, ProphetNet in a unique manner making it a major novelty of the proposed approach. The inconsistencies present in the summary with existing methods is avoided as the hybrid model generates the following:

1. “Title” for the summary
2. “Main summary”
3. “End note” of the summary

Lack of balance is handled by our proposed model as it gives a full picture of the content in the document which will help the stakeholders better understand the content and context of the document in a lucid manner. The context of the whole document is captured well using our proposed method which is an improvement over conventional methods.

The main abstractive summary comes from the two state-of-the-art models of the hybrid architecture. Once both models generate a summary, they will be compared and incorporated to generate a coherent summary. The end note is generated by combining individual summaries from three state-of-the-art models and refining it to a few concluding statements. Identification and interpretation of important and relevant content in the final summary is one of the major advantages of our ensemble model over the single models. It saves huge amount of computation and running time without requiring additional or external data for training.

D. Summary Curation Module

Once the title, main summary, and end note are generated by all models of the ensemble model, it is then used to generate a final lucid summary, which gives a bird’s-eye view of the uploaded document(s) to the end user yet retaining the context.

V. BENEFITS OF PROPOSED APPROACH

The key concerns with the existing methods are inconsistencies, lack of balance, single model summary and context preservation. Our focus is to improve all of the existing concerns using our proposed model. Our ensemble document summarization model has the following major benefits over the previous methods:

1. It readily applies abstractive summarization on the documents available in various formats like PDF (.pdf) and Word (.doc and .docx).
2. The proposed algorithm can generate section-wise, page-wise or document-wise summary, as demanded by user.
3. Hybrid summarization is employed to generate a summary from a combination of four state-of-art models instead of one.
4. Tables from documents can also be extracted and will be part of the final summary.
5. Model retraining is not required while processing documents belonging to any domains.

VI. EXPERIMENTAL RESULTS

Performance of our proposed hybrid model supersedes four state-of-the-art models in both ROUGE and BLEU scores, as described in Table II. Its ROUGE-1 score is 169% more than the average of all four SOTA models, while it is 173% more on ROUGE-2 and 180% more on ROUGE-1 scores. The proposed hybrid model scores an impressive 30.64, over seven times more than the average of all four SOTA models. We used BLEU score [13] and ROUGE scores [14] for the evaluation, considering 35 documents (12 .pdf, 12 .doc and 11 .docx) ranging from six to fifteen pages as input.

The proposed model shows great improvements in all ROUGE scores [14], as the generated summaries match with the ground truth summaries created by-hand by five different subject matter experts for all the 35 documents. The proposed ensemble model provides improved context, maintains accurate grammar, reduces bias, minimizes perusing time and redundant manual effort, as can be observed in Fig. 2.

TABLE II
BLEU AND ROUGE SCORES OF PROPOSED HYBRID MODEL AND FOUR SOTA MODELS

Model Name	ROUGE-1	ROUGE-2	ROUGE-1	BLEU
Model 1 - PEGASUS	0.1656	0.1317	0.1656	0.0746
Model 2 - T5	0.3737	0.2755	0.3737	9.2062
Model 3 - BART	0.3692	0.2901	0.2769	7.7577
Model 4 - ProphetNet	0.1176	0.05	0.1176	0.0751
Model 5 – Hybrid Model (Proposed)	0.4341	0.3234	0.4219	30.64

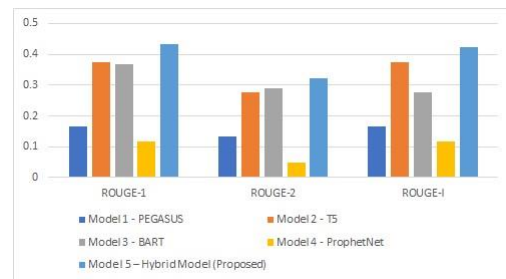


Fig. 2. ROUGE Score Comparison

VII. CONCLUSION AND FUTURE WORK

The summaries produced by our proposed ensemble model gains highly on the usage of the single-model summaries, in terms of effectiveness and robustness. The business need of deriving insights effectively from any type of documents in a short turnaround time is also satisfied. This model provides abstractive summarization of three different kinds which brings out the highlights as needed by user ; page-wise, section-wise and document-wise.

We have started working on finetuning the models and improving the overall efficiency of the proposed system. We

plan to work on other type of documents such as .pptx and images wherein the normal text extraction and image-to-text conversion, can also be performed using an OCR engine.

REFERENCES

- [1] J. Stewart, V. Mittal, J. Carbonell, and M. Kantrowitz, "Multi-document summarization by sentence extraction," *NAACL-ANLP 2000 Workshop on Automatic Summarization*, 05 2002.
- [2] A. Bagalkotkar, A. Kandelwal, S. Pandey, and S. S. Kamath, "A novel technique for efficient text document summarization as a service," in *Proceedings of the 2013 Third International Conference on Advances in Computing and Communications*, ser. ICACC '13. USA: IEEE Computer Society, 2013, p. 50–53. [Online]. Available: <https://doi.org/10.1109/ICACC.2013.17>
- [3] K. Prudhvi, A. Chowdary, P. Reddy, and P. Prasanna, *Text Summarization Using Natural Language Processing*, 01 2021, pp. 535–547.
- [4] M. Gambhir and V. Gupta, "Recent automatic text summarization techniques: a survey," *Artificial Intelligence Review*, vol. 47, pp. 1–66, 2016.
- [5] N. Munot and S. Govilkar, "Comparative study of text summarization methods," *International Journal of Computer Applications*, vol. 102, pp. 33–37, 09 2014.
- [6] S. Gajbhiye and M. Lopes, "Template-based nlg for tabular data using bert," 02 2021, pp. 1–5.
- [7] S. Gholamrezazadeh, M. A. Salehi, and B. Gholamzadeh, "A comprehensive survey on text summarization systems," *2009 2nd International Conference on Computer Science and its Applications*, pp. 1–6, 2009.
- [8] J. Morris and G. Hirst, "Lexical cohesion computed by thesaural relations as an indicator of the structure of text," *Computational Linguistics*, vol. 17, pp. 21–48, 01 1991.
- [9] Y. Ko and J. Seo, "An effective sentence-extraction technique using contextual information and statistical approaches for text summarization," *Pattern Recognit. Lett.*, vol. 29, pp. 1366–1371, 2008.
- [10] Pymupdf documentation. [Online]. Available: <https://pymupdf.readthedocs.io/en/latest/>
- [11] Project description of docx2txt package. [Online]. Available: <https://pypi.org/project/docx2txt/>
- [12] Antiword documentation. [Online]. Available: <https://cran.r-project.org/web/packages/antiword/index.html>
- [13] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002.
- [14] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *ACL 2004*, 2004.

EMOTION AND SENTIMENT CLASSIFICATION USING TRANSFORMER MODELS

Gem Rose Kuriakose, Prabhitha Nagarajan, Selvakuberan Karuppasamy and Subhashini Lakshminarayanan

Accenture, Advanced Technology Centers in India, India
 {gem.rose.kuriakose, prabhitha.nagarajan, s.b.karuppasamy, and
 s.j.lakshminarayanan}@accenture.com

Abstract— Emotions play a vital aspect in human life. All our thoughts and actions are reflected through our emotions. We primarily use words to describe those emotions and it determines how we act and behave. So, there is a need to analyze these emotions to understand them better. Usually, the data is in the form of text, audio, video, image etc. of human expressions. Currently, in our experiment, we have considered the data which is in textual form. It is a challenging task to compute the emotion. In past much research has been done in this area still many improvements are possible. With the advent of recent Deep Learning and state-of-art NLP models, we need to consider which model works best for the dataset. This paper summarizes the comparative analysis of the experiments done in the area of emotion analysis in text information. For that purpose, eight well-known state-of-the-art models have been picked up and performed several experiments across different datasets, whose results are shown and commented upon, leading to some interesting conclusions about the capabilities of each analyzed algorithm. By training and evaluating the 8 Transformer models on multiple datasets, we found that few models outperformed the other models on the sentiment and emotion classification tasks by achieving a very good accuracy and F1-score.

Keywords—*sentiment analysis, opinion mining, emotion analysis.*

I. INTRODUCTION

A. Background

Emotions can be an indispensable part of the way you think and act. The emotions you feel every day can propel you to make a move and impact the choices you make regarding your life, both enormous and little. Mostly, this information is available on social media platforms, email data, call-center recording, customer surveys, web blogs and browser history, customer rating and review sections. As these data are unstructured and the size of the data is huge it becomes hard to convert them into valuable insights. As a result, organizations are losing the opportunity to understand market needs, competitive research, brand reputation, employee sentiment within an organization, and product reception in the case of a newly launched product. With the advent of more complex deep learning algorithms, it might be possible to extract emotions from text data.

B. Introduction

There is a lot of importance in analyzing the sentiment and emotions of various text data, for instance knowing the customer, building healthy relationships with them, and improving the business based on customer feedback but with the tremendous increase in the number of deep learning algorithms and data, the complexity of the work has

increased. Following this thought, our paper aims to present a detailed evaluation of the 8 Transformer models focused on different emotion datasets. And on the other hand, this work will analyze the results obtained from Sentiment and Emotional Analysis. Consequently, we use three data collections in the field of classification analysis. The effort made from this paper will benefit the user/researcher to have sufficient knowledge about the different capabilities provided by each Transformer model, and from this, the user/researcher can select the most relevant model to be included in his/her application.

II. LITERATURE SURVEY

Bidirectional Encoder Representations from Transformers [1] is pre-trained on a large corpus of unlabelled text extracted from the Books and English Wikipedia. It is one of the state-of-the-art algorithms which has surpassed human performance in classification. During the training phase, it learns information from both directions of the context at once. It is then fine-tuned to perform specific NLP tasks by adding an extra output layer.

Robustly Optimized BERT Pre-training Approach [2] is trained on an unannotated text drawn from the web. It is a powerfully enhanced technique for pretraining natural language processing framework that enhances BERT. It predicts intentionally hidden sections of text within the dataset. Due to the dynamic masking feature each time the input is fed as a sequence to the model the masked token changes because of which we can get better performance

During the pre-training phase, we use a technique called distillation based on DistilBERT (SANH et al., 2020) [3] which improves the inference speed of BERT.

XLNet [4] is a profound model to understand the language context better. We have two versions of XLNET base and large. XLNET has made it one stride further by predicting each word in a sequence using all possible permutations. This model has a better performance but at the cost of training speed and inference on multi-class classification.

ALBERT [5] is trained with the focus to have less memory consumption and increase the training speed of BERT model. Due to its lightweight nature, this is more suitable for memory constraint applications. ALBERT's parameters are shared across multiple layers to increase efficiency. This model tries to check for coherence loss between sentences.

ELECTRA [6] outperforms existing techniques like RoBERTa and XLNet on the GLUE benchmark when using less than a quarter of their computing. This uses generator and discriminator models. It uses a replaced token detection technique where we predict whether each word or token is an original input or a generated sample by the model. It has

fewer parameters which make it possible to run on a single GPU.

DeBERTa [7] has disentangled attention which is pretrained on large amounts of raw text corpora. It is achieved using two techniques like disentangled attention mechanism and an enhanced mask decoder.

GPT-2 [8] stands for Generative Pre-trained Transformer 2 which is trained to predict the next word, given all of the previous words within some text in 40GB of Internet text. This model learns the inner representation of the language which can be used to extract features useful for downstream NLP tasks.

Munika et al. [9] has discussed the Fine-Grained Sentiment classification using SST-5. The model was trained using BERT embeddings which does not have any complex architecture. This has outperformed in terms of accuracy on Stanford Sentiment Tree-Bank datasets at the root and total node levels.

III. COMPARATIVE ANALYSIS

A. Datasets

We tested our models on three datasets which span a range of fine-grained emotion classes (2, 6, 28 classes).

- Stanford Sentiment Treebank (SST) is one of the most popular publicly available datasets for fine-grained sentiment classification tasks. SST-2 is taken from GLUE benchmark [10]. These sentences are based on movie reviews and their sentiments are manually annotated. In Figure 1 we see a more balanced dataset.



Figure 1: Label distribution of SST-2 dataset

- CARER (Saravia, Elvis et al., 2018) [11] is an emotion dataset that consists of 6 emotions. Figure 2 shows that the emotion “joy” is 9 times more frequent than “surprise”. We see a large imbalance in the distribution of emotions in the dataset.

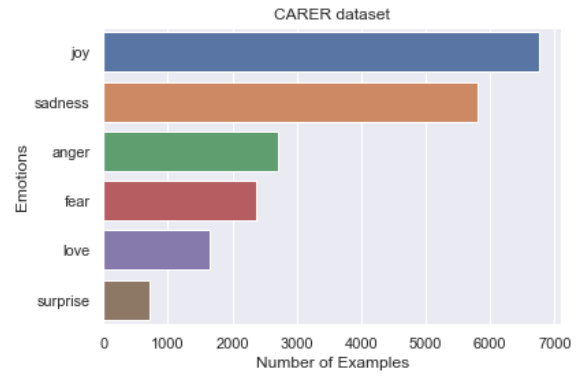


Figure 2: CARER Emotion dataset distribution

- GoEmotions [12] dataset which contains Reddit posts annotated for 28 emotions. It includes 12 positive, 11 negative, 4 ambiguous emotion categories and 1 “neutral”, making it widely apt for conversation understanding tasks that require a fine-drawn differentiation between emotion expressions. The original data corpus contains 58k carefully curated comments extracted from Reddit, with human annotations to 27 emotion categories or Neutral. Figure 3 observed a large imbalance in terms of emotion frequencies. Emotions like admiration, approval, gratitude are more frequent than emotions like grief and pride.

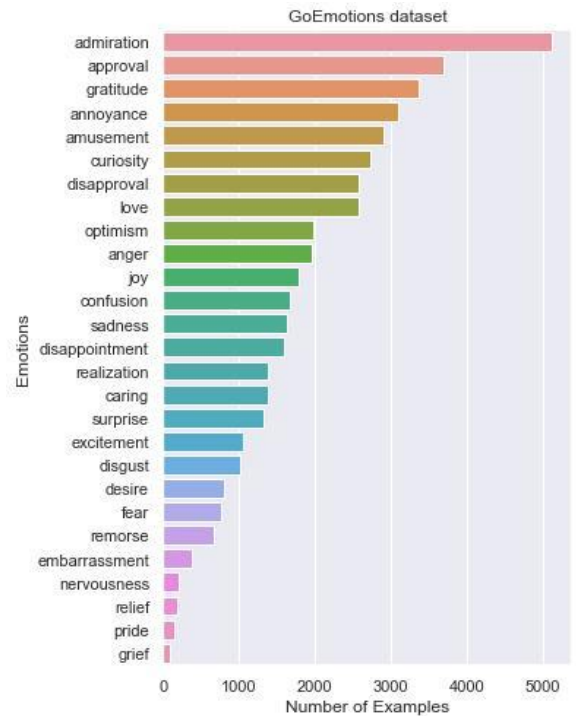


Figure 3: GoEmotion dataset distribution

Sentiment	BERT	RoBERTa	DistilBERT	XLNet	ALBERT	ELECTRA	DeBERTa	GPT-2
Negative	0.929	0.951	0.908	0.933	0.882	0.910	0.928	0.937
Positive	0.931	0.953	0.915	0.936	0.889	0.902	0.933	0.940

Table 1: Results of different models for SST-2 Sentiments

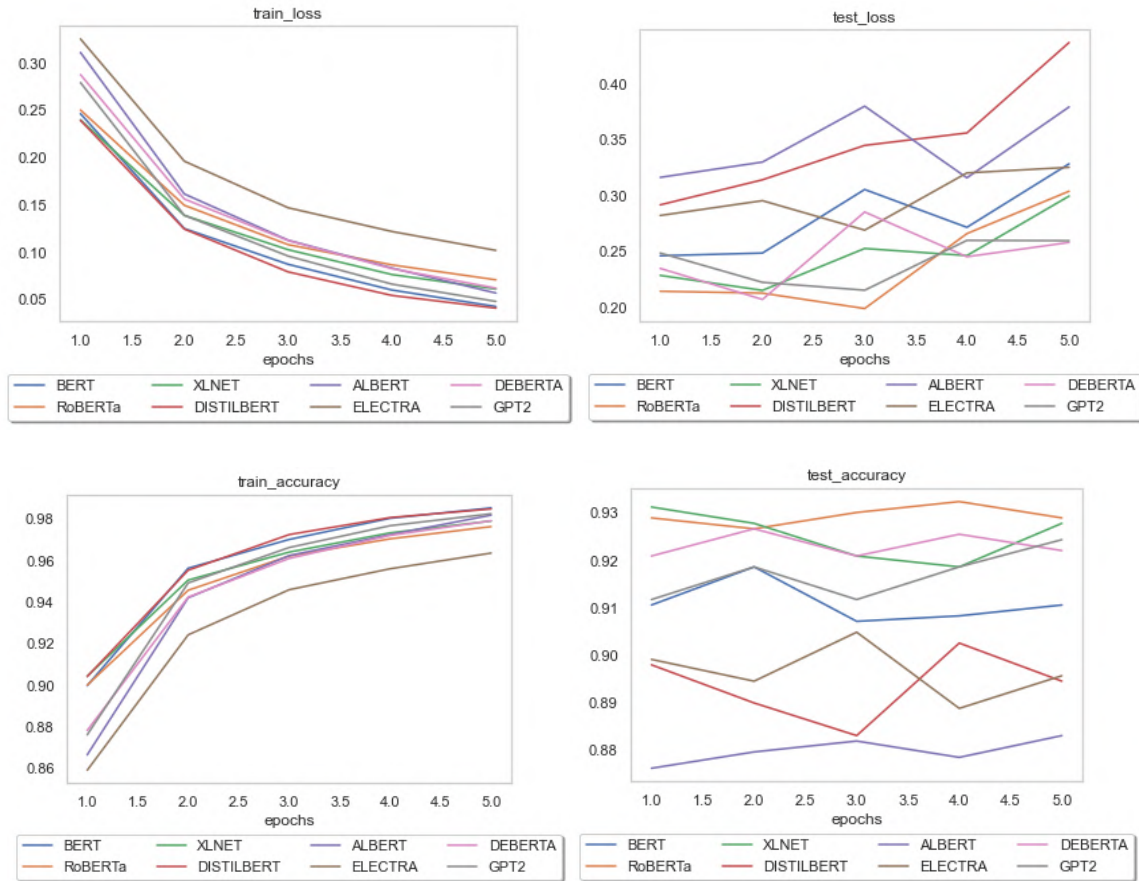


Figure 4: Model Comparison on SST-2 Dataset for each epoch based on Loss and Accuracy

B. Task Definition

This paper consists of 3 sub-tasks where all 8 Transformer models are trained for three different datasets. Table 2 shows the split-up of training, validation, and test data.

Subtask A: This is a binary classification task. Here we predict the two-class sentiment of a given sentence based on the SST-2 dataset. The task is to predict the two-class (POSITIVE/NEGATIVE) sentiment of a given sentence and this is used only sentence-level labels. The model is trained for 5 epochs.

a) Subtask B: Using the CARER dataset we predict the multi-class emotion of the sentences. The output decides whether the emotion is ANGER, FEAR, JOY, LOVE, SADNESS, AND SURPRISE. Here the model is trained for 10 epochs.

b) Subtask C: In this subtask, we deal with the GoEmotion dataset which includes more significant categories like positive, negative, and ambiguous emotions. This is suitable for any domain-specific tasks that require an understanding of emotions. In this method, we used sequences which have a maximum length of 40 in training and evaluation datasets and run over 10 epochs.

	<i>Train Data</i>	<i>Validation Data</i>	<i>Test Data</i>
SST-2	67K	1K	2K
CARER- 6 class	16K	2K	2K
GoEmotions-28 class	43K	5K	5K

Table 2: Dataset Comparison based on the emotion and sentiment class

C. Methodology

We decided to adopt an architecture like [9] with different transformer models using various datasets. Our experiment is based on comparative performance. Firstly, we perform pre-processing of the text dataset making it suitable for the different transformer model formatting. We then pass the text through the last layer of the transformer model, then fine-tune it by applying a dropout layer with the ratio of 0.1 and pass the data to a fully connected softmax layer which outputs the probabilities of all the classes based on the dataset $\{0,1\}$ (SST-2), $\{0-5\}$ (CARER), $\{0-27\}$ (GoEmotions). Between experiments of the 8 Transformer models, we add an additional output layer to the pre-trained model with the same dropout value and add a softmax layer to do the classification. We ran multiple epochs through the training data using ADAMW optimizer and kept the beta to the default values of 0.9 and 0.99.

Finally, we compared the loss and accuracy on both the training and test data over multiple epochs for all the eight transformer models against the published results of the original paper. Later, we analysed the trade-off between

Dataset	Data Split-Up
---------	---------------

different models both in terms of accuracy/F1-score and the time for training on the three chosen datasets.

Training time taken per epoch is listed in a tabular form (TABLE-2,3,4) with a batch size of 32 on a Tesla P100-PCIE-16GB. All models which we trained are implemented using the Huggingface library. This practice lets us analyse and compare the performance of the model.

D. Parameter Tuning

We fine tuned various hyper-parameters like learning rate, batch size, number of epochs, weight decay for all pre-trained transformer models to get the best accuracy/F1-score.

Different hyper parameter settings for all models are below:

- Batch Size considered 16, 32
- Learning rate performed with 5e-5, 3e-5, 2e-5
- Epochs run for 5, 10, 30
- Weight decay: (0, 0.01, 0.3)

Upon investigating the performance of the models, we observed that after a few epochs of training, the training loss gradually decreased but the test loss remained constant, and the model began to overfit on the training data, so we limited the number of epochs.

Emotion	BERT	RoBERTa	DistilBERT	XLNet	ALBERT	ELECTRA	DeBERTa	GPT-2
Sadness	0.968	0.972	0.964	0.962	0.961	0.962	0.967	0.967
Joy	0.943	0.949	0.944	0.948	0.943	0.952	0.954	0.948
Love	0.806	0.799	0.786	0.832	0.803	0.841	0.844	0.831
Anger	0.925	0.928	0.921	0.926	0.917	0.916	0.923	0.925
Fear	0.893	0.888	0.890	0.875	0.892	0.873	0.881	0.894
Surprise	0.720	0.762	0.770	0.754	0.707	0.761	0.797	0.727

Table 3: Results of different models for CARER Taxonomy

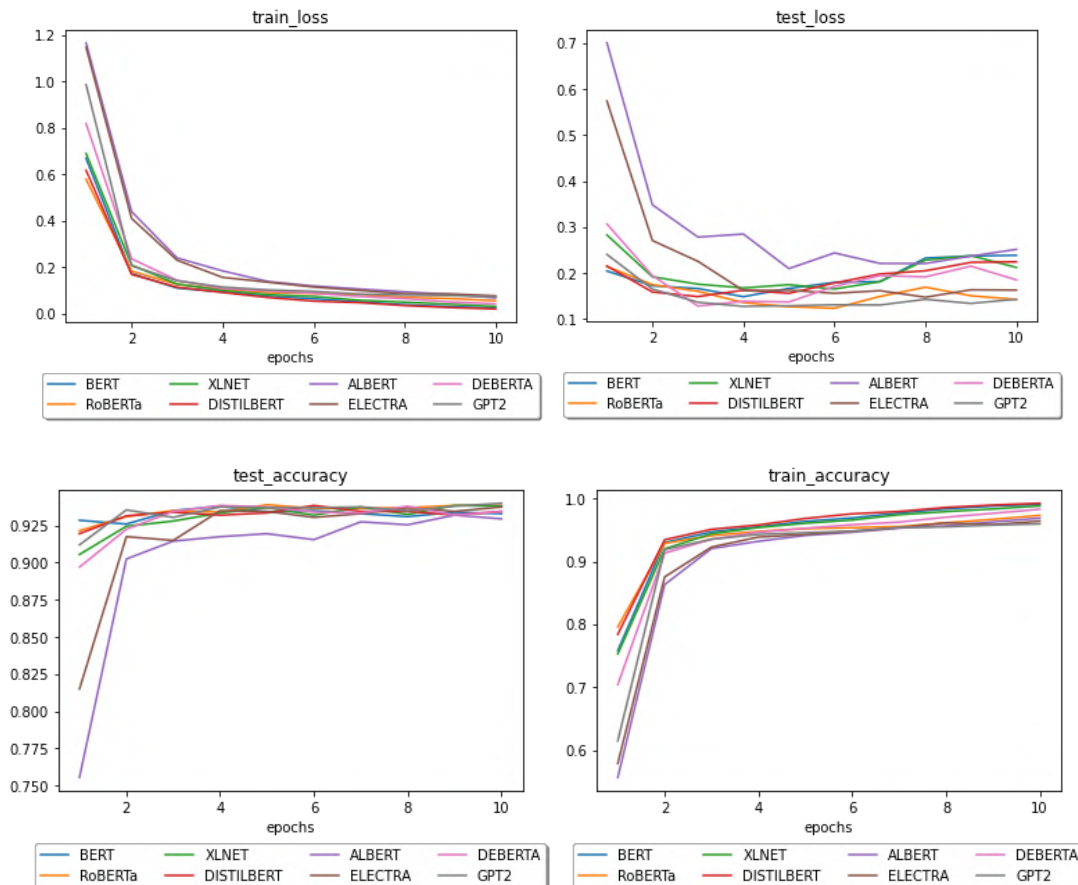


Figure 5: Model Comparison on CARER Dataset for each epoch based on Loss and Accuracy

IV. RESULTS AND DISCUSSION

This section describes the evaluation measures for our three subtasks. We computed four standard metrics to measure the models' performance for each class: accuracy, F1-score, precision and recall. We used accuracy to compare the performance among all the models for the sentiment

classification task (SST-2) and applied F1-score (macro) for the remaining two multi-class emotion classification tasks (CARER, GoEmotions).

A. Subtask A

Model	Training Time (per 5 epoch)	Accuracy on Test Set.
BERT	06:07	93.00%

Model	Training Time (per 5 epoch)	Accuracy on Test Set.
RoBERTa	06:21	95.20%
DistilBERT	03:18	91.20%
XLNet	07:34	93.50%
ALBERT	06:02	88.60%
ELECTRA	02:27	90.60%
DeBERTa	07:56	93.10%
GPT-2	20:51	93.80%

Table 4. Experiment results for classification task on SST-2

Table 4 shows the result of our eight models (BERT, RoBERTa, DistilBERT, XLNet, ALBERT, ELECTRA, DeBERTa and GPT-2) best test accuracy and the training time taken to run each epoch for the sentiment analysis task. All our eight models achieved the best results on the Stanford Sentiment Treebank dataset. RoBERTa achieved the highest test accuracy of 95% after 5 epochs and outperformed the other models, while ALBERT achieved the lowest accuracy of 88%. The models BERT, XLNet, DeBERTa, GPT-2 achieved an accuracy of approximately 93%. But if we compare these models, GPT-2 took almost thrice the running time for each epoch to reach that accuracy. The light transformer models DistilBERT and ELECTRA achieved an accuracy of 90-91% and they required substantially less pre-training resources compared to the other models.

Table 1 gives details on the eight models' performance (F1-Score). Except for DistilBERT, all other models achieved similar performance on both the sentiments (Negative, Positive). From Figure 2 we understand that the model is overfitting on the training set.

B. Subtask B

Model	Training Time (per 10 epoch)	Macro-F1
BERT	01:54	0.876
RoBERTa	01:56	0.883
DistilBERT	01:00	0.879
XLNet	02:25	0.883
ALBERT	01:54	0.87
ELECTRA	00:43	0.884
DeBERTa	04:05	0.894
GPT-2	06:53	0.882

Table 5: Experiment results for classification task on CARER

Table 5 summarizes our eight models (BERT, RoBERTa, DistilBERT, XLNet, ALBERT, ELECTRA, DeBERTa and GPT-2) best test accuracy and the training time taken to run each epoch for the multi-class emotion classification tasks on the CARER (6-emotions) dataset. We observed that all the models achieved better and almost the same performance on

Emotion	BERT	RoBERTa	DistilBERT	XLNet	ALBERT	ELECTRA	DeBERTa	GPT-2
Admiration	0.608	0.592	0.578	0.59	0.541	0.582	0.567	0.587
Amusement	0.702	0.691	0.695	0.676	0.683	0.698	0.67	0.694
Anger	0.407	0.411	0.408	0.413	0.36	0.388	0.406	0.388

the CARER emotion dataset. DeBERTa achieved the highest F1-Score (macro-average) of 0.894. DistilBERT and ELECTRA have the fastest training time. Also, when we increased the number of epochs from 5 to 10, there was 2% increase in the F1-Score while the accuracy value remained almost constant. Similarly, we noticed an improvement of 1-2% in the F1-score when we increased the batch size from 16 to 32.

Table 3 reveals the eight models' performance (F1-score). The result insights that the emotions sadness, joy, and anger have a high F1-score, while surprise has the least F1-score across all the eight Transformer models. This is because of the non-uniform distribution of the training data across different classes. This can be overcome by evenly balancing the data across the classes by performing either oversampling or undersampling. As per the plot based on Figure 5, we observe that a good trade-off is achieved between training and test accuracy.

C. Subtask C

Model	Training Time (per epoch)	Macro-F1
BERT	04:04	0.41
RoBERTa	04:09	0.436
DistilBERT	02:10	0.402
XLNet	04:51	0.426
ALBERT	03:50	0.383
ELECTRA	01:51	0.423
DeBERTa	05:10	0.413
GPT-2	13.11	0.434

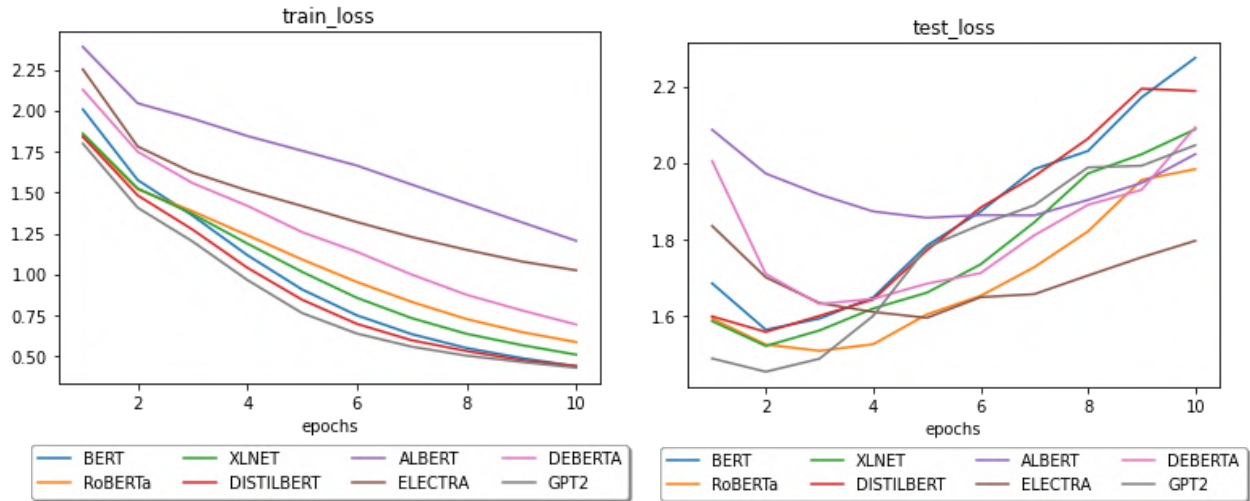
Table 6: Experiment results for classification task on GeoEmotion

Evaluation from Table 6 summarizes our eight models (BERT, RoBERTa, DistilBERT, XLNet, ALBERT, ELECTRA, DeBERTa and GPT-2) best test accuracy and the training time taken to run each epoch for the multi-class emotion classification tasks on the GoEmotions (28-emotions) dataset. RoBERTa achieved the highest F1-score (macro-average) of 0.436. ALBERT which is a Lite version of BERT achieved the least F1-score of 0.383 compared to other models. The training time of the model is an important factor when we increase the number of classes in a classification task. DistilBERT and ELECTRA achieved close to the best results with the fastest training time. GPT-2 has the longest training time with a good F1-score.

Table 7 shows the eight models' performance (F1-Score). The models obtained have the best results for most of the classes and did not achieve the worst performance for any class of emotions. The emotions grief, pride and relief have least F1-Score because of a smaller number of training data points. If we increase the data points of those classes, then our models will achieve better performance in those classes. In Figure 6 plots we observe that the test loss and accuracy does not improve after a few epochs.

Annoyance	0.254	0.307	0.263	0.271	0.204	0.253	0.281	0.271
Approval	0.337	0.355	0.265	0.316	0.261	0.331	0.323	0.349
Caring	0.346	0.319	0.309	0.344	0.231	0.374	0.322	0.382
Confusion	0.31	0.357	0.363	0.343	0.289	0.366	0.325	0.348
Curiosity	0.38	0.37	0.361	0.399	0.33	0.427	0.419	0.389
Desire	0.443	0.463	0.403	0.479	0.414	0.411	0.435	0.446
Disappointment	0.255	0.282	0.211	0.22	0.199	0.255	0.215	0.304
Disapproval	0.332	0.321	0.317	0.324	0.3	0.333	0.282	0.32
Disgust	0.448	0.441	0.42	0.395	0.414	0.397	0.421	0.472
Embarrassment	0.364	0.448	0.389	0.492	0.353	0.4	0.338	0.448
Excitement	0.305	0.313	0.323	0.318	0.273	0.328	0.378	0.343
Fear	0.607	0.607	0.548	0.607	0.559	0.639	0.623	0.619
Gratitude	0.786	0.788	0.795	0.798	0.81	0.809	0.791	0.789
Grief	0.2	0.143	0	0.308	0	0	0	0.222
Joy	0.473	0.519	0.467	0.494	0.472	0.506	0.496	0.486
Love	0.672	0.661	0.686	0.665	0.684	0.69	0.685	0.693
Nervousness	0.298	0.383	0.195	0.333	0.298	0.435	0.385	0.343
Optimism	0.45	0.447	0.458	0.455	0.45	0.473	0.406	0.448
Pride	0.182	0.414	0.357	0.19	0.385	0.182	0.353	0.211
Realization	0.189	0.163	0.192	0.167	0.117	0.224	0.178	0.193
Relief	0.1	0.381	0.222	0.261	0	0.25	0.16	0.375
Remorse	0.533	0.544	0.5	0.583	0.637	0.58	0.556	0.522
Sadness	0.459	0.475	0.498	0.478	0.456	0.483	0.5	0.461
Surprise	0.445	0.432	0.44	0.44	0.401	0.469	0.482	0.457
Neutral	0.595	0.592	0.58	0.564	0.591	0.566	0.563	0.605

Table 7: Results of different models for GoEmotion Taxonomy



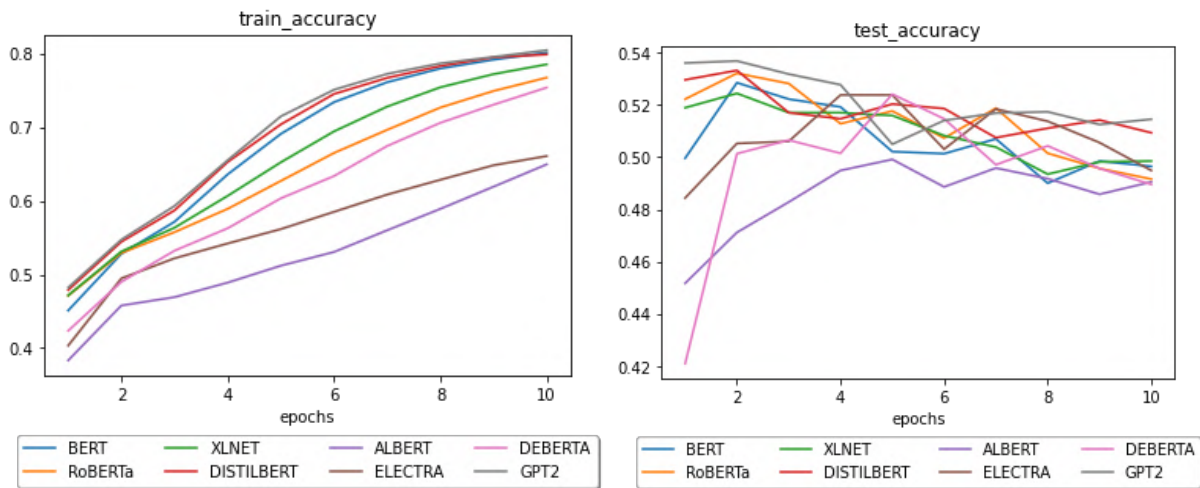


Figure 6: Model Comparison on GoEmotions Dataset for each epoch based on Loss and Accuracy

V. CONCLUSION

In this research, we used 8 transformer models which are BERT, RoBERTa, DistilBERT, XLNet, ALBERT, ELECTRA, DeBERTa and GPT-2. These are applied to SST-2 dataset, CARER dataset and Go Emotions dataset for sentiment/emotion classification. We fine tuned our models with different hyper parameters and evaluated the performance of the models using metrics like accuracy, precision, recall, and F1-score.

By training and evaluating the 8 Transformer models on multiple datasets, we found that RoBERTa outperformed the other models on the sentiment and emotion classification tasks by achieving an accuracy of 95.2% on the SST-2 sentiment dataset and macro averaged F1-score of 0.436 on 28-class Go Emotions dataset. DeBERTa outperformed the other models on the 6-class CARER emotion dataset and achieved the macro averaged F1-score of 0.894. DistilBERT and ELECTRA models had the fastest training time and smallest memory requirements and they achieved good accuracy/F1-score on SST-2 and CARER dataset. ALBERT achieved the least accuracy/F1-Score on all the tasks compared to other models. The decoder transformer model GPT-2 attained results with good accuracy/F1-score like other models but it took a larger training time due to the huge number of parameters.

When we ran the XLNet model for multiple epochs we observed that the XLNet-large-cased model performs best in sentiment analysis however it did not give good results for the emotion datasets (CARER and GoEmotions). In multi-class classification XLNet large was able to correctly predict only two labels in the dataset. So, we replaced the XLNet-large-cased model with the XLNet-base-cased which showed significant improvement in the F1 score in both the datasets.

Our future work will be about exploring different areas of the multi-class emotion classification with a more fine-grained dataset. For example, when we look at the performance metrics of the 28-emotion classification, we still have much to improve; we can develop a hybrid model and

fine-tune the hyper parameters in the existing models to achieve the best results.

REFERENCES

- [1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In North American Association for Computational Linguistics (NAACL).
- [2] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692, 2019.
- [3] Victor SANH, Lysandre DEBUT, Julien CHAUMOND, Thomas WOLF. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. EMC²: 5th Edition Co-located with NeurIPS'19. arXiv:1910.01108v4 [cs.CL]
- [4] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V Le. XLNet: Generalized autoregressive pretraining for language understanding. In NeurIPS, 2019.
- [5] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite bert for self-supervised learning of language representations. arXiv preprint arXiv:1909.11942, 2019.
- [6] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: Pre-training text encoders as discriminators rather than generators. In ICLR, 2020.
- [7] Pengcheng He, Xiaodong Liu, Jianfeng Gao, Weizhu Chen. DEBERTA: DECODING-ENHANCED BERT WITH DISENTANGLED ATTENTION. Published as a conference paper at ICLR 2021. arXiv:2006.03654v6 [cs.CL]
- [8] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever. Language Models are Unsupervised Multitask Learners. <https://openai.com/blog/better-language-models/>
- [9] Munikar M., Shakya S. & Shrestha A. (2019) Fine-grained Sentiment Classification using BERT. arXiv:1910.03474[cs.CL]
- [10] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In International Conference on Learning Representations (ICLR).
- [11] Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, Yi-Shin Che, CARER: Contextualized Affect Representations for Emotion Recognition in Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing.
- [12] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A dataset of fine-grained emotions," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Jul. 2020

SPAM DETECTION USING DEEP LEARNING TECHNIQUE

K Ranjith Reddy

Research Scholar

Dept. of Computer Science and Engineering

Madhav University, Rajasthan

Email: ranjith.kssr5@gmail.com

Dr. Sanjay Chaudhary

Professor

Dept Computer Science and Engineering

Madhav University, Rajasthan

Email: schaudhary0020@gmail.com

Abstract—Spammers and spammers abound on social networks. Despite the fact that social media platforms have implemented a number of techniques to prevent spam from spreading, tight information review mechanisms have given rise to more sophisticated spammers. In this paper, we present a spam detection approach based on the self-attention Bi-LSTM neural network model in combination with ALBERT, which is word vector model. We use ALBERT to convert text from social networks into word vectors, which we then feed into the Bi-LSTM layer. The final feature vector is created after feature extraction and combining it with the self-attention layer's information focus. Finally, the result is classified by the SoftMax classifier.

Keywords—*Feature Selection, Spam detection, Machine Learning, LSTM, Deep Neural Networks.*

I. INTRODUCTION

Online social networks[1,2] have grown in popularity as a result of the support of wireless connection and computers. Users of online social network platforms (OSNs) have access to simple communication and interactive tools that allow them to share a variety of information pertaining to their lives and jobs, such as text, photographs, and videos, in real time. However, OSNs have attracted a large number of spammers due to the large number of active users and the easy conditions for uploading information. Spammers have wreaked havoc on ordinary people and platforms by releasing spam. By analyzing sources, spams that previously had little impact on social networks can now leverage Online social networks to have a massive spread impact. Spammers may simply and precisely deliver spam to possible target users and boost dispersion via OSNs, which provide all basic user information and give follow-up options. Like as an example, messages including a URL in social platforms [3,4] may reach tens of thousands of people in a matter of seconds.

Spam is a sort of data effectively sent by a spammer and its motivation is to beguile, spread wrong information, and publicize for benefits. Spams [5,6] will create issues, for example, asset occupation, expanded correspondence time, and data transfer capacity squander. It mainly showed us that 15% of the spams contain connections to vindictive substance, sexual entertainment, or malware. Albeit a great deal of related exploration has been done, there are as yet countless spams on informal organizations. Also the spammers that make and send spams will camouflage spam by noticing the stage sifting procedures.

This paper mostly concentrate on a spam recognition model. ALBERT and Bi-LSTM [9,10,12] are used and their networks depend on a self-consideration system.

The commitment of this paper contribution is as follows:

- i. In perspective on the immense measure of informal organization tweet information, acquainting the ALBERT model with implanting the text to work on the effectiveness of model order while guaranteeing the precision of the model.
- ii. Aiming at the circumstance where spammer conceals the spam cover in the typical text, acquainting the Bi-LSTM with the spam tweet acknowledgment technique, which can completely consider the specific situation and semantics to catch the center of tweets text.
- iii. Give a lot of viable data, acquainting the self-consideration instrument with the Bi-LSTM model to additionally get catchphrases that influence the order aftereffects of tweet attributes.

II. LITERATURE SURVEY

In order to cluster or classify text documents [7,8,11] by applying machine learning techniques, documents should first be preprocessed. At present, there are already many techniques proposed in the field of Spam Detection in Social Networks [13,14,15]. This section is dedicated to all the research that has been done in this field which is used for the project reference.

In[12]Daiqi Zhou, Jun Liu and Guangxia Xu., proposed, which we have referred for the paper. In this paper they worked on building a spam detection model using Combination of Bidirectional Long-short Term Memory model with Self-attention and ALBERT or A Lite BERT(Bidirectional Encoder Representations from Transformer) as Tokenizer. They used ALBERT, a light weight word vector model of the BERT, paired with neural network model to create a spam detection system. First, they converted social network texts into word vectors using ALBERT, which they fed into Bi-LSTM layer. By first extracting features, and then combining the extracted features created. To get the result, the SoftMax classifier performs classification.

In paper [9], P.Bhuvaneshwari, A. Nagaraja Rao and Y. Harold Robinson., worked on building a non-traditional Machine Learning model for spam classification. They thought that due to some restricted

feature representation and data changes used by a spammer to avoid detection, classical machine learning approaches are unable to detect spam messages successfully. During the research process, they proposed an innovative framework based on deep learning (DL), called the self-attention-based CNN BiLSTM method, as an Machine Learning-based Alternatives to detection. With an attention mechanism, our system calculates the weightages of each word in the text and recognises the spamming clues existing in the content. It uses Convolution Neural Network (CNN) to learn sentence representation and extract high-level n-gram characteristics. Later, some sentence vectors are integrated as document feature vectors using Bidirectional lstm (Bi-LSTM) to identify spam reviews with contextual information. At the end they compared their result with results from already existing models and found that their model is better than any other models present at that time.

In this paper [10], they worked on the ALBERT model which is a version of the BERT model. When pretraining natural language representations, increasing model size frequently leads to better performance on downstream tasks. However, owing to GPU/TPU memory restrictions and lengthier training times, additional model growth becomes more difficult at some point. To overcome these issues, we describe two parameter-reduction strategies for reducing memory usage and increasing BERT training speed (Devlin et al., 2019). When compared to the original BERT, our proposed approaches produce models that scale substantially better. They also employ a self supervised loss which mainly works on modeling coherence between sentences and shows that it consistently supports downstream tasks with multi-sentence inputs.

III. MACHINE & DEEP LEARNING TECHNIQUES FOR SPAM DETECTION

ALBERT is a lite version of BERT(bidirectional encoder representations from transformers) it is mainly used in the NLP(natural language processing) field. BERT utilizes transformer structure as the principle system. It changes over the distance of two words at any situation into one, which successfully tackles the drawn out reliance issue in NLP. Training of BERT is divided in two sections: NSP (next sentence prediction) and MLM (masked language model).

ALBERT [12] structure is similar to BERT, it actually utilizes transformer and encoder layers with proper selected activation function. Given below 3

main updates which are not available in BERT but available in ALBERT.

- **Factorization of the Embedding lattice:**This is extension of BERT model, for example, ROBERTa and XLNet, the info layer embedding and secret layer embeddings have a similar size. However, the creators isolated the two inserting networks in this model. This is on the grounds that input-level inserting (E) is necessary to refine just setting autonomous adapting yet secret level implanting (H) requires setting subordinate learning. This progression prompts a decrease in boundaries by 80% with BERT.
- **Cross-layer parameter sharing:** Author of the this proposed model is also additionally proposed the boundary dividing among various layers of model to further develop proficiency and diminishing repetition. This paper suggests that past interpretations of BERT, XLNet, and ROBERTa have stack encoder layers, so the model can learn the comparison process for each layer. The creators proposed three sorts of boundary partaking in this paper:

- Just offer Feed Forward network boundary
- Just offer consideration boundaries
- Share all boundaries. Default setting utilized by creators except if expressed in any case.

The above advance prompts a 70% decrease in the general number of boundaries.

- **Inter Sentence Coherence Prediction :** Like the BERT, ALBERT additionally utilized Covered Language model in preparing. Notwithstanding, Rather than utilizing NSP (Next Sentence prediction) misfortune, ALBERT utilized another misfortune is said to be SOP (Sentence order prediction). NSP is a pairing failure of anticipating if two parts will appear sequentially in the first message, the detriment of this failure is that it checks the clarity, like the topic, to recognize the next sentence. Infact the SOP is simply seeking clarity of the proposal.

Like BERT, ALBERT [9,12] is pre-trained with several data sets such as English dictionary.

A. LSTM and Bi-LSTM

LSTM are specifically designed to eliminate this long term dependency problem. The main difference between LSTM and RNN is that the repeating module in RNNs has a very simple structure which uses a single tanh layer (supposedly), but in LSTMs, the

repeating module has a different structure which uses four different layers which interact in a very special way.

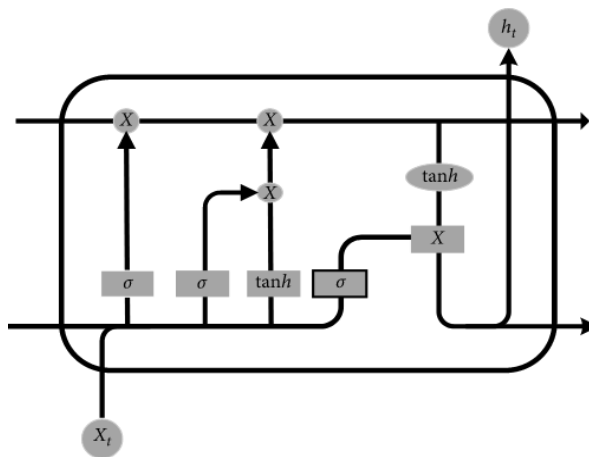
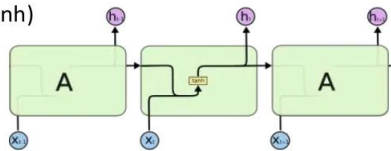


Fig 3.1 Structure of LSTM

– **RNN:** single layer (tanh)



– **LSTM:** 4 interactive layers

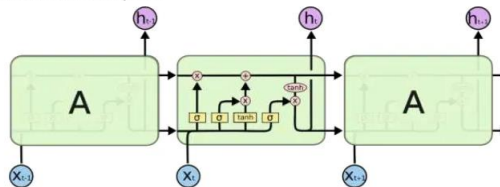


Fig 3.2 Difference between RNN and LSTM

In **Bi-LSTM** training Information travels in both forward and backward directions. Also it Remembers complex long term dependencies better. The parser always considers the text before it and ignores the text that follows. Dropout layers are applied to the output vector.

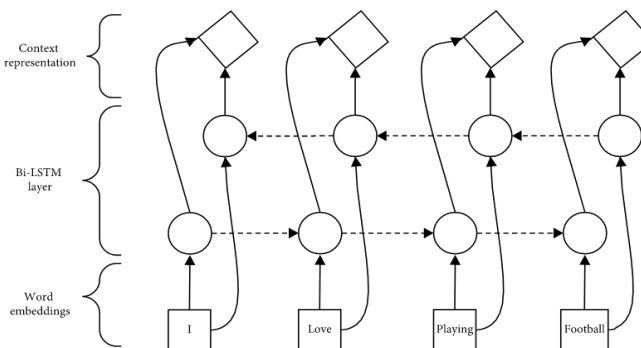


Fig 3.3 Structure of the Bi-LSTM

B. Self Attention Mechanism

At the time of spam detection, we generally confronted the circumstance about the quantity of texts being restricted, particularly for the circumstance that the substance data of various users' tweets shifts enormously, it is hard to get more compelling semantic data.. For instance, words, for example, "advancement" and "markdown" can serve to rapidly recognize the promoting expectation of the item tweeted by spammers. Simultaneously, various words assume various parts in grouping. Removing catchphrases streamlines the component extraction process. The presentation of the consideration system can build productivity and further develop the grouping precision. Contrasted and the consideration instrument, self-consideration just figures consideration inside the succession, searching for the inner association of the grouping.

The estimation equation is as per the following:

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V,$$

This is where the input is a matrix obtained from three different parameters Q,V,K. First, calculate the multiplication of the Q,K matrix and divide by $\sqrt{d_k}$. Then this is where the input is a matrix obtained from three different parameters.

IV. PROPOSED DEEP LEARNING TECHNIQUE FOR SPAM DETECTION

It merely needs to count the harmful words for standard spam detection activities. Spammers substitute dangerous terms with other ones, yet they may still send the same information, thanks to spammers' constant optimization. Short text information and word association information can be better understood by ALBERT and Bi-LSTM. We use efficient ALBERT and Bi-LSTM model for our method. For the first time, the model employs ALBERT for key attributes extraction and comprehend the original text's semantic elements. To identify spam, Bi-LSTM and self-attention are merged. The structure and processes of the model are illustrated in figure 4.1 below.

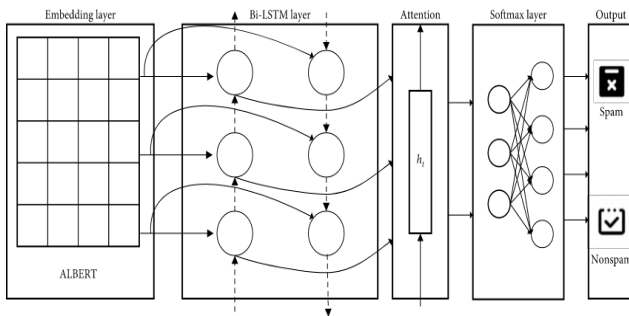


Fig 4.1: Proposal model architecture

4.1 Embedding Layer / ALBERT Layer

We must do some preprocessing procedures on the data before putting it into the model. To create a sentence with text, we need to remove some words and punctuation, such as words, parentheses, and spaces between pictograms. The embed layer then takes the preprocessed input from data and converts each word from text data into a dictionary number. The multilayer bidirectional transformer encoder will produce the feature representation.

4.2 Bidirectional LSTM with Self-Attention Layer

The used ALBERT model attempted to understand the link between different words by serialising the data. ALBERT, on the other hand, finds it difficult to comprehend a brief text. To detect spam, we use self-attention with Bidirectional-LSTM. Text feature data from the ALBERT layer is trained in this layer, and text features are input to the forward and backward LSTMs, respectively. For the final output, two text vector representations are created and combined. Finally, apply sigmoid normalization to the output to get the desired result.

4.3 Improvements / Changes in Existing Method

Some improvements that will be done the project are:

- i. Using Another Social Network Spam Dataset.
- ii. Using TensorFlow/Keras for building the model.
- iii. Adding an extra LSTM layer in the model.
- iv. Changing the hyperparameters of the model.

V. EXPERIMENT RESULTS

The dataset used in this paper is twitter spam detection dataset from kaggle. For the paper purpose we will only use Tweets and Type columns. It is then divided into a 9:1 ratio with 90% as training set and 10% testing set. Below is the number count of Spam and Quality tweets in the dataset.

```
Quality    6153
Spam       5815
Name: v1, dtype: int64
```

The model used in this project is built using TensorFlow library and It consists of 1 Embedding layer, 1 Bidirectional LSTM layer, 1 Attention layer and 1 Hidden layer.

We are using Adam Optimizer with 0.01% Learning rate with accuracy metrics.

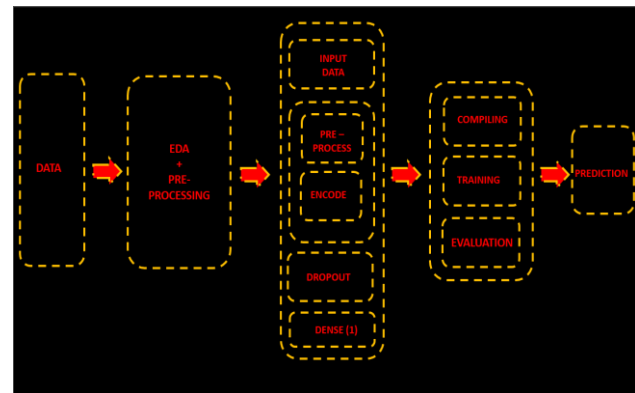


Fig 4.1 Block diagram of overall implementation

After training the model got the training accuracy of 100% and testing accuracy of 94.5% after training it for 20 epochs. Below are the accuracy and Loss plot for the same.

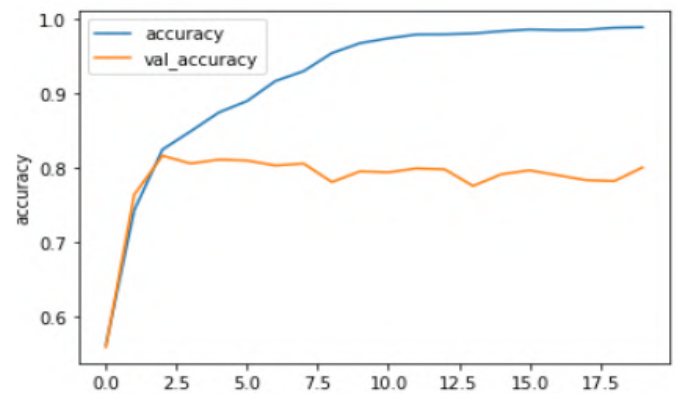


Fig 4.2 Accuracy Plot of the Model

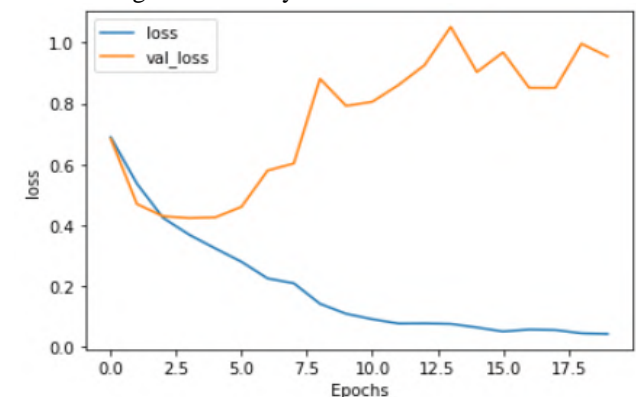


Fig 4.3 Loss Plot of the Model

VI. CONCLUSION

It is tough to spot spam on social media sites. The fundamental issue in this sector is that spammers

have been updating and iterating spam text depending on detection tactics, which has resulted in a drop in detection accuracy. To address this issue, we present a spam detection approach that takes into account all of the text's contextual information while also utilising the self attention mechanism to compensate for the text's brevity. We used Bi-LSTM to compare the differences across machine learning algorithms. Experiments suggest that Bi-LSTM is more effective at spam detection. We also used comparison tests to demonstrate the efficiency of the model's self-attention mechanism.

REFERENCES

- [1]. A. Branitskiy, D. Levshun, N. Krasilnikova et al., "Determination of young generation's sensitivity to the destructive stimuli based on the information in social networks," *Journal of Internet Services and Information Security*, vol. 9, no. 3, pp. 1–20, 2019.
- [2]. M. Kolomeets, A. Benachour, D. E. Baz et al., "Reference architecture for social networks graph analysis," *Journal of Wireless Mobile Networks, Ubiquitous Computing, and Dependable Applications*, vol. 10, no. 4, pp. 109–125, 2019.
- [3]. G. Xu, X. Wu, and J. Liu, "A community detection method based on local optimization in social networks," *IEEE Network*, vol. 34, no. 4, pp. 42–48, 2020.
- [4]. S. Liu, J. H. Yim, H. J. Lee et al., "Semantic similarity calculation method using information contents-based edge weighting," *Journal of Internet Services and Information Security*, vol. 7, no. 1, pp. 40–53, 2017.
- [5]. C. Naiyue, L. Yun and C. Hanchi, "Overlapping Community Detection Using Non-Negative Matrix Factorization with Orthogonal and Sparseness Constraints", *IEEE Access*, vol. 6, pp. 21,266-74, 2018.
- [6]. X. Deng et al., "Efficient Vector Influence Clustering Coefficient Based Directed Community Detection Method", *IEEE Access*, vol. 5, pp. 17,106-16, 2017.
- [7]. L. Ou et al., "Releasing Correlated Trajectories: Towards High Utility and Optimal Differential Privacy", *IEEE Trans. Dependable and Secure Computing*, 2018.
- [8]. J. Zhang et al., "A Feature-Hybrid Malware Variants Detection Using CNN Based Opcode Embedding and BPNN Based API Embedding", *Computers & Security*, vol. 84, pp. 376-92, 2019.
- [9]. P. Bhuvaneshwari, A. Nagaraja Rao and Y. Harold Robinson, Spam review detection using self attention based CNN and bi-directional LSTM. May 2021, *Multimedia Tools and Applications* 80(5):1-18
- [10]. Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma and Radu Soricut, ALBERT: A Lite BERT Model for Self Supervised Learning. Published as a conference paper at ICLR 2020.
- [11]. Jie Hao, Xing Wang, Baosong Yang, Longyue Wang, Jinfeng Zhang, and Zhaopeng Tu. Modeling recurrence for transformer. Proceedings of the 2019 Conference of the North, 2019.
- [12]. Daiqi Zhou, Jun Liu and Guangxia Xu., Social Network Spam Detection Based on ALBERT and Combination of Bi-LSTM with Self-Attention. vol 2021.
- [13]. Zeeshan Bin Siddique, Mudassar Ali Khan, Ikram Ud Din, Ahmad Almogren, Irfan Mohiuddin, and Shah Nazi, AI-enabled Decision Support System: Methodologies, Applications, and Advancements 2021, Volume 2021 |Article ID 6508784.
- [14]. A. Zamir, H. U. Khan, W. Mehmood, T. Iqbal, and A. U. Akram, "A feature-centric spam email detection model using diverse supervised machine learning algorithms," *The Electronic Library*, vol. 38, no. 3, 2020.
- [15]. Tian Xia and Xuemin Chen, "A Discrete Hidden Markov Model for SMS Spam Detection", *Applied Science MDPI Appl. Sci*, vol. 10, pp. 5011, 2020.

TOOLS AND TECHNIQUES FOR DETECTING SARCASM – A SURVEY

¹Ayesha Shakith, Ph.D Research Scholar,

Department of Computer Science, St. Joseph's College, Tiruchirapalli – 620005

Ayeshasm1412@gmail.com

² Dr. M. Kriushanth, Assistant Professor,

Department of Data Science, St. Joseph's College (Autonomous), Tiruchirapalli - 620002

Krishiaf11@gmail.com

Abstract:

In recent years, due to the tremendous usage of social media platforms like Twitter, Facebook etc., the demand for Sentiment Analysis has been exponentially increasing. Sentiment Analysis is one of the most exploring research area in Big Data Analytics. Like any other research area, Sentiment Analysis too has its own issues & challenges to overcome. Challenges like Multilingualism, Fake News Detection, Negation Words, Emoji's etc., prevails. This paper discusses on one of the most prominent and promising challenge called as "Sarcasm Detection". In Sentiment Analysis, this credibility is based on the Polarity (or) Accuracy of a sentence or a word. When there is an issue achieving the maximum polarity in a sentence, then it can be consider that the processed results are not accurate. So, this challenge has to be ruled out. Hence, this paper mainly discusses the various tools and techniques that helps in overcoming the challenge of "Sarcasm Detection".

Keywords:

Big Data Analytics, Sentiment Analysis, Sarcasm Detection, Machine Learning (ML), Deep Learning(DL).

1. Introduction

A Sentiment Analysis is the computationally identifying and categorizing opinion in which the writer's attitude towards a particular topic, product,

etc. With sentiment analysis tools and techniques, movie reviews, online marketing, restaurants, clothing, retail stores, airlines etc., in social media handles can quickly detect dissatisfied customers. Categorize their issues based on urgency, and prioritize their responses in order to retain the customers and also to make a new one.

Real-time sentiment analysis is more advanced during a potential crisis, allows an authority to take action before a customer's unpleasant experience goes viral in social media platforms. Sentiment analysis is also used to analyze the competitors by tracking how customers are talking about their products and find an opportunity to improve businesses. Sentiments can be classified into three types namely, **positive**, **negative**, or **neutral**.

For example:

- "I really like the new movie!" → Positive
- "I'm not sure if I like the movie" → Neutral
- "The new movie is awful!" → Negative

Detecting sentiments and emotions in a sentence from real world is highly challenging. That too challenges like sarcasm detection is quite tricky and complicated to identify when it is written in a piece of text.

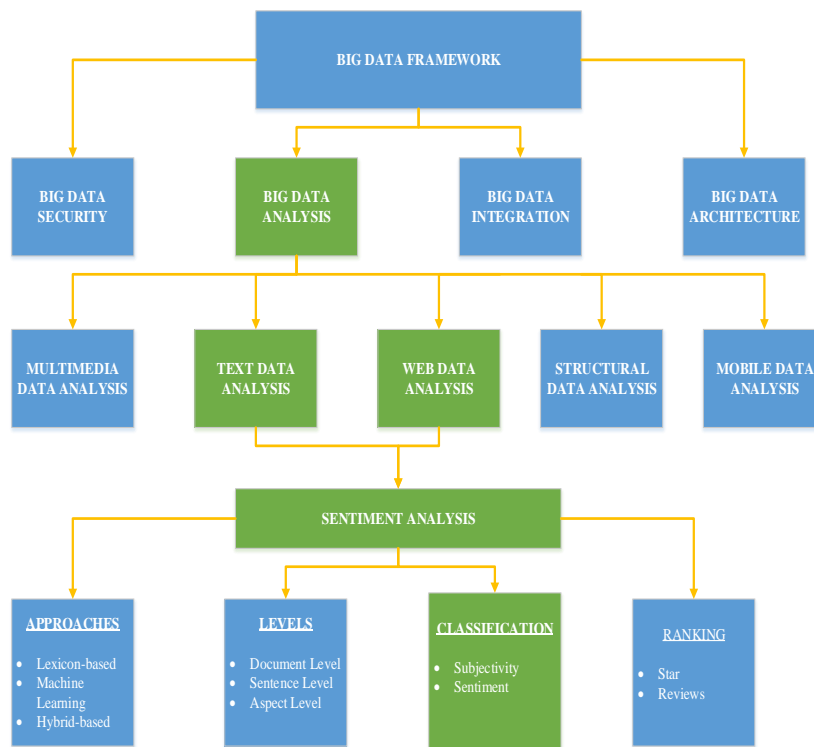


Figure: 1- Classification of Big Data Framework

2. Sarcasm Detection

Sarcasm is used to often mock or contempt in a hilarious way. Sentence said and the intention of a person contrasts. Sarcasm is a form of irony said mainly out of humor or anger. But all irony is not sarcasm. Sarcasm is all about the context and voice tone. So, that's why it is easy to detect while said verbally and complicated when it is read from a sentence. There are different types of Sarcasm like Self- Deprecating, Deadpan, Brooding, Juvenile, Literary, Obnoxious, Polite, etc

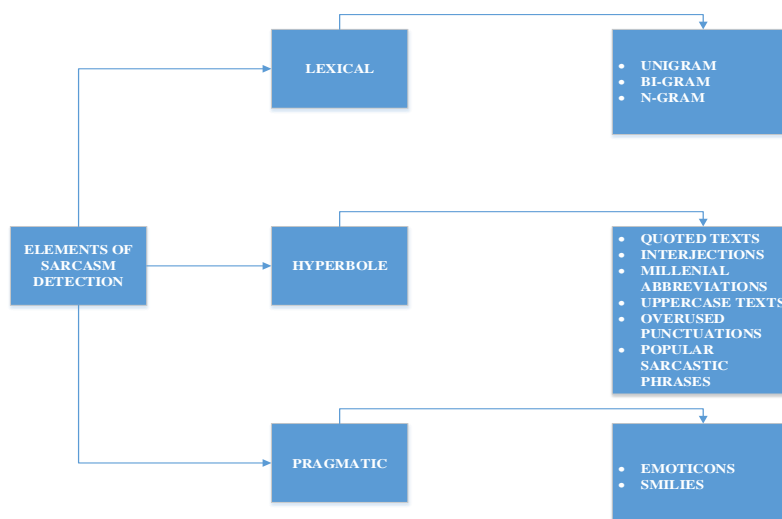


Figure: 2- Various Elements of Sarcasm

2.1 Sarcasm Identification:

Step: 1- Data Collection

In Sentiment analysis, Sarcasm Detection involves the following steps:

Most of the researchers prefer collecting dataset from twitter followed by Facebook also using by using Amazon product

Step: 2- Data Pre-processing:

Pre-processing of data is mainly carried out in three steps.

1 - Tokenization of data. (Splits sentences into words)

2 - Stemming and lemmatization. (The words are converted into present tense).

3 - Removal of stop words. (Those words where search engine has programmed to ignore)

4 -P-O-S tagging. (Classifies the words into a verb, adverb, adjective, noun etc.)

Step: 3 - Feature Extraction:

1. Term – Frequency: Number of times a particular word appears in the document

2. Feature Presence: Whether a particular feature appears in the document or not.

3. Term Frequency–Inverse Document Frequency (TF-IDF): It is used as frequency measure of feature.

4. Weighted Term Frequency-Inverse Document Frequency: It is used in the calculation of sentiment score using slang words with the help of weighted TF-IDF

5. Delta Term Frequency-Inverse Document Frequency (TF-IDF): It makes an easy computation, understanding, and implementation.

6. N-gram: It is basically a sequence, as a unigram, bigram, trigram etc.

7. Word2vec: Word2vec produces a vector space using the large input of text

8. Pattern-Related: This is the Supervised learning method is which is towards sarcastic patterns based on parts-of-speech.

Step: 4 - Sarcasm detection

Sarcasm Detection can be done by using classifiers and rule-based methods. Several machine learning classifiers such as SVM, Naïve Bayes, AdaBoost, gradient boosting, Random forest etc. are used. Several Rule-based methods are also used.

Step: 5 - Polarity Detection: After identifying the sentiments in a sentence, it is detected whether the sentence is sarcastic or non-sarcastic

Step: 6 - Calculation of Results: In the process the result is calculated using certain metrics i.e., the accuracy, precision, recall, and F-score.

2.2 Various Tools and Techniques for Sarcasm Detection:

There are several classification techniques used by many researchers during their research. The most used classifiers, clustering technique and rule based methods are tabulated below.

Table: Tools and Techniques used for Sarcasm Detection

Classifiers	Clustering	Rule based methods
<ul style="list-style-type: none"> • Naïve Bayes • Random Forest • Decision Tree • AdaBoost • Gradient Boost • Support Vector Machines (SVM) • Maximum Entropy • Cross-Validation 	<ul style="list-style-type: none"> • K-means Clustering • Agglomerative Hierarchical Clustering 	<ul style="list-style-type: none"> • Lexical Method • Semantic Method • Syntactic Method • Pragmatic Method • Prosodic Method • Idiosyncratic

3. Related Works

This section presents the research findings related on tools and techniques that is used to detect Sarcasm. This section is categorized into two sections. Literature Review, related work to Sarcasm and Comparative Literature review in a table format.

3.1 Literature Review

Pansy et al. presented a systematic survey related to social networking sites and mainly focus on Twitter Sentiment Analysis. To identify the research gaps and future scope, this paper explored and identified the techniques in well-structured approach. To enhance the efficiency of classification, stack based ensemble, fuzzy based and neural network based classifiers are implemented. It is also analyzed that WEKA, R Studio, Python are mostly used tools by research scholars for implementation. This survey paper compares various research articles from the year 2009 to 2021. Based on the observations taken from this survey, Naïve Bayes and SVM are the most explored machine learning classifiers. Plenty of work has been conducted in the area of ensemble classification technique.

Aditi et al. presented a survey on different approaches for detecting and classifying the sarcastic texts and findings of different approaches. This article also summarized various techniques used in social media posts. There are different techniques overview to sarcasm detection using emoticons, detecting sarcasm content within images combined with texts. As a future scope, this article suggests, effective automatic sarcasm detection on stream data memes, which can be found in real time. Sarcasm detection from audio clippings is another research area to be explored.

Miguel et al. suggested to add bias mitigation mechanisms in sentiment analysis to

avoid giving more or less credibility to a news item. High performance AI systems particularly those based on Deep Learning provides good results. Also discussed on Multilingualism, fake news detection in social media content. Fake news which is slightly modified data along with an authentic news story is hard to predict. To overcome this challenge, Aspect based Sentiment Analysis and Adversarial training can be used.

Jiangnan Li et al. discussed a novel BERT-based model that effectively processes common sense knowledge. COMET model fine-tuned is utilized on ConceptNet to generate common sense knowledge candidates. Two types of knowledge selection strategies like explicit knowledge selection and implicit attention based knowledge selection has been compared. The explicit selection majority, minority and contract sentiment-based methods. Knowledge text-integration module to combine the information from both text and knowledge has been designed. To testify the effectiveness of the proposed method, experiments on three datasets has been conducted.

Sameer at al. discussed how various researchers proposed three models for sociopolitical opinion polarity classification of microblog. Researchers proposed a novel sarcasm detection model that uses ideology and fine grained opinion as features with other linguistic features to classify sarcastic opinions. The various methods used are: Preferred Reporting Items for Systematic Reviews and Meta Analysis (PRISMA), multiple studies Adapted ML Algorithms (AMLA). Other studies that developed on Semi-adapted AMLA were grouped under the customized ML algorithms (CMLA). PRISMA statement was used to provide a detailed guideline of the systematic review of the classification methods, SVM algorithm is the most used in AMLA group. CNN and SVM were found

to offer a high prediction performance. As a result, SVM, CNN-SVM were the most efficient ML algorithms to predict sarcasm on twitter.

Abel et al. proposed a novel model called Chameleon Swarm Optimization (CSO) along with machine learning based sentiment analysis on Sarcasm Detection and Classification (CSOML-SASC) model. This model involves Preprocessing, Feature Selection, Classification, Parameter optimization. CSOML-SASC technique involves TF-IDF based feature extraction model, WKELM based classification and CSO based parameter optimization. The steps involved in this pre-processing are: Extraction of single letter word, elimination of several spaces, extraction of punctuation marks, elimination of numerals, elimination of stop words and convert uppercase letters into lowercase. By using this model, the result is impressive with all metrics have maximum score with high performance.

Sangeeta et al. presented a systematic survey related to social networking sites, sentiment analysis and mainly focused on twitter sentiment

analysis. This article explored and identified the techniques and tools used in structured approach to find the research gaps and identify the future scope. As a result, it concludes the following: Most explored social networking site is Twitter. Most used Machine Learning algorithms are SVN and Naïve Bayes. Recently, stack-based ensemble, fuzzy based and neural networks are implemented to enhance the efficiency of classification. Most implemented tool is WEKA, R Studio and Python. Most used data corpus is Sentiment 140.

Jyoti et al. in this article, attempted to present detailed architecture in general for sarcasm detection, along with current techniques, ensemble learning methods, similar work performed by the researchers in context of detecting Sarcasm on Twitter and future.

3.2 Literature Review – A Comparison

This section gives a clear, comparative of analysis of some of the research articles which focuses on the tools and techniques used to detecting sarcasm.

Table: 1 Comparative Literature Review

References	Key Concepts	Machine Learning/ Deep Learning	Dataset	Techniques/ Approaches	Evaluation Metrics
Pansy Nandwani et. al., 2021	Emotion Detection and Sentiment Detection	Machine Learning and Deep Learning	Social Media	Dictionary based, Corpus based and Lexicon based	Accuracy, Precision, Recall, F-Measure, Sensitivity
Miguel A. Alonso et. al., 2021	Fake News Detection	Machine Learning	Fake News Dataset, Snopes Dataset, BS Detector Dataset	NLP, Text Analytics	Accuracy, Precision, Recall, F1, AUC
Aditi Arora, 2020	Sarcasm Detection	Machine Learning, Deep Learning	Twitter, Facebook, Redditt, Flickr	Pattern based, feature extraction, Context based pattern based, SVM, Adaboost, CNN	Accruacy, Precision, F-Measure score, F1 scores

Shameer et.al., 2021	Domain Independent Model	Machine Learning, Deep Learning,	Social Media	BoW, TF/IDF, Word Embeddings, Word 2 Vec BERT, NLP, LSTM, BLSTM, BERT, Transformer, Seq2Seq Encoding	Skip Gram
Jaeheon Kim et. al., 2022	Machine driven toxic chat detection	Deep Learning	Twitch	Neural Network Classifier, LSTM NLP	Accuracy, Precision, Recall, F1-Score
Srijita Majumdar et. al., 2021	Sarcasm Detection, Mood Retention	Machine Learning, Deep Learning,	Pre- Trained Dataset	NLP, RNN, LSTM, Word Embeddings	Precision
Prasad Wagh et. al., 2021	Sarcasm Detection, Hate Speech Detection	Machine Learning	Positive Dataset, Negative Dataset, Hate-Speech Dataset	SVM, LSTM	Accuracy, Precision, Recall, F1-Score,

4. Issues and Challenges in Sentiment

Analysis

- There are various challenges identified through various research works are listed below:
- Sarcasm Detection and Hate Speech. [PRA, 2021]
- Multilingual, Multimodal Sentiment Analysis, Sarcasm Detection, Detection of Sentiment Polarity. [POO, 2020]
- Sarcasm and Irony. [ABD, 2021]
- Generation of data by means of Informal Text. [PAN, 2021]

- To identify and differentiate between legitimate and misleading sentences are sometimes critical. [JOS, 2022]
- Misspellings and Slang words in twitter dataset. [MOH, 2019]
- Fake News Detection, Multilingualism, Explainability, Mitigation of Biases, Multimedia Element. [MIG, 2021]
- Toxic Chat Detection using Twitch. [JAE, 2021]
- Domain Independent model for Sentiment Analysis using Word Embeddings.
- Translation for word embedding in local language like Dogri, Kashmiri. [SHA, 2021].

5. Future Directions

This section is divided into two sub divisions, first is the future work related to sentiment analysis and the second one is related to Sarcasm Detection.

Sentiment Analysis:

References	Future Directions
[POO, 2020]	Many authors mainly focused on English language followed by Arabic and Chinese. So, the challenges remains for the researchers to work on Multilingual Sentiment Analysis
	Most of the work is done on twitter. The opinions available on another social media website like Facebook posts, messages, product reviews on different e-commerce websites are also has to be taken into consideration

Sarcasm Detection:

[ABD, 2021]	Developing task interaction and task interaction modules and mechanisms for Sentiment Analysis and Sarcasm Detection.
[JOS, 2022]	To investigate the interoperability of the linguistic features avoiding model agnostic approaches.
[JYO, 2021]	New datasets, function sets and consideration of different ways of Sarcasm Detection among other aspects.
	Different Deep Learning methods can be explored along with more conceptually oriented functionality.
	More study on the hyperbole function and documents syntactic dependencies could be focused.
[JAE, 2021]	NLP techniques should be more comprehensive in analyzing Social Media posts aided with encryption and slang
[ADI, 2021]	Effective automatic detection on stream data, memes, audio clips which can be found in real time

6. Conclusion

After analyzing the articles, it can be concluded that, the most explored social networking site is Twitter followed by Facebook. Twitter remains the top due to its limitation of characters up to 280 per tweet. People have taken social media as a tool to express their opinions, to give feedback, to comment. Detecting Sarcasm is a most tedious and difficult process. That is just by using only text it is quite hectic to find because sarcasm is all about the tone and the way of expression. Various tools and techniques of machine learning algorithms are used in this process. But, the most used Machine Learning algorithms are SVM and Naïve Bayes. Recently, stack-based ensemble, fuzzy based and neural networks are implemented to enhance the efficiency of classification. Most implemented tool is WEKA, R Studio and Python. Most used data corpus is Sentiment 140. Most of the researches has been done on text sentiment analysis and most of them targets to improve efficiency of classification.

References

- [1] Mohammed Ibrahim Al-Mashhadani, Kilan M. Hussein, Enas Tariq Khudir, Muhammad Ilyas, *“Sentiment Analysis using optimized Features sets in different Facebook/ Twitter Dataset domains with Big Data”*, Iraqi Journal for Computer Science and Mathematics”, 2022.
- [2] Rasul Abdel Kareem Atu, Abbas Lufti Hussein, Abbas Lufti Hussein, Nadia Majid Hussein, *“Stylistic Analysis of Sarcasm in some selected extracts of schoolteacher in Morrison’s Beloved”*, Linguistics and Culture Review, 6(S2) 1-15, 2022.
- [3] Jaeheon Kim, Donghee Yvette Wahn, Meeyoung Cha, *“Understanding and Identifying the use of emotes in toxic chat on Twitch”*, Online Social Networks and Media, 2022.
- [4] Jose Antonio Garcia- Diaz, Rafael Valencia-Garcia, *“Compilation and evaluation of the Spanish SatiCorpus 2021 for Satire identification using linguistic features and transformers”*, Complex and Intelligent Systems, 2022.
- [5] Kavitha, Suneetha Chittineni, *“Sentiment Analysis for sarcastic messages in social media”*

using deep learning techniques – An empirical study”, IT in Industry, Vol.-9, No.2, 2021.

[6] Abdelkar El Mahdaouy, Abdellah El Mekki, Nabil El Mamoun, Kabil Essefar, Ismail Berrada, Ahmed Khoumsi, “*Deep Multi-Task Model for Sarcasm Detection and Sentiment Analysis in Arabic Language*”, Proceedings of sixth Arabic natural language processing workshop, PP: 334-339, 2021.

[7] Jyoti Godara, Rajni Aron, “Sarcasm Detection on Social Network: A Review”, Annals of R.S.C.B, Volume.25, Issue.6, PP: 3761-3771, 2021.

[8] Srijita Majumdar, Debabrata Datta, Arpan Deyasi, “*Sarcasm Detection and Mood Retention using NLP Techniques*”. International Journal of Information Retrieval Research, Volume.12, Issue.1, PP: 1-23, 2021.

[9] Pansy Nandwani, Rupali Verma, “*A Review on Sentiment Analysis and Emotion Detection from Text*”, Social Network Analysis and Mining, 2021

[10] Miguel A.Alonso, David Vilares, Carlos Gomez-Rodriguez, Jesus Vilares, “*Sentiment Analysis for Fake News Detection*”, Electronics, 2021.

[11] Abel Sridharan, “*Chameleon Swarm optimization with Machine Learning based Sentiment Analysis and Sarcasm Detection and Classification Model*”, International Research Journal of Engineering IT & Scientific Research, Volume.8, Issue.10, PP:821-828, 2021.

[12] Shameer Bashir, Arvind Selwal, “*A Comprehensive survey of Sentiment Analysis: Word Embeddings approach, research challenges and opportunities*”, 2nd International Conference on IOT based control networks and Intelligent Systems, 2021.

[13] Prasad Wagh, Pratik Jaiswal, Ankit Rahangdale, Sherlin Titus, “*Sentiment Analysis using Machine Learning (The Sorting Hat)*”, Journal of Science and Technology, Vol. 06,

Special Issue, PP: 377-382: ISSN: 2456-5660, 2021

[14] Jiangnan Li, Hongliang Pan, Zheng Lin, Peng Fu, and Weiping Wang, “*Sarcasm Detection with Commonsense Knowledge*”, IEEE/ACM Transactions on Audio, Speech, and Language Processing, Vol. 29, PP: 3192- 3201, 2021.

[15] Sangeeta Rani, Nasib Singh Gill, Preeti Gulia, “*Survey of Tools and Techniques for Sentiment Analysis of Social Networking Data*”, International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 12, No. 4, PP: 222 – 232, 2021.

[16] K. Kavitha, Suneetha Chittineni, “*Sentiment Analysis for Sarcastic Messages in Social Media Using Deep Learning Techniques - An Empirical Study*”, IT in Industry, Vol. 9, No.2, ISSN (Print): 2204-0595 ISSN (Online): 2203-1731 PP: 1051 – 1052, 2021.

[17] Samer Muthana Sarsam, Hosam Al-Samarraie, Ahmed Ibrahim Alzahrani, Bianca Wright, “*Sarcasm detection using machine learning algorithms in Twitter: A systematic review*”, International Journal of Market Research, Vol. 62(5), PP: 578–598, 2020

[18] Bhumi Shah and Margil Shah, “*A Survey on Machine Learning and Deep Learning Based Approaches for Sarcasm Identification in Social Media*”, In: Kotecha K., Piuri V., Shah H., Patel R. Data Science and Intelligent Applications. Lecture Notes on Data Engineering and Communications Technologies, vol.52, Springer, PP: 247 – 259, P: ISBN-978-981-15-4473-6, E: ISBN-978-981-15-4474-3, 2020

[19] Poonam Bhatia, Dr. Rajender Nath, “*A systematic review of Levels, Methods and Tasks in Sentiment Analysis*”, NOVYI MIR Research Journal, Volume-5, Issue-11, PP: 94-115, 2020

[20] Omar Y Adwani, Marwan Al Tawil, Ammar M Huneiti, Razan H Al-Dibsi, Rawan A Sahin,

ANeer A Abu Zayed, “ Twitter Sentiment Analysis Approaches: A Survey” International Journal of Emerging Technologies in Learning, August 2020.

[21] Anwar Shathik J, Krishna Prasad K, “A Literature review on application of sentiment

analysis using machine learning techniques”, International Journal of Applied Engineering and management letters(IJAEML), ISSN: 2581-7000 Vol. 4, No.2(Aug 2020)

A REVIEW ON MACHINE LEARNING APPROACH OF VIRTUAL SCREENING AND DRUG TARGET IDENTIFICATION IN DRUG DISCOVERY

G.Hemalatha, Junior Executive, Vinayaka Mission's Research Foundation, Salem, hema_phd2020@yahoo.com¹

Dr.P.Sasikala, Professor of Department. of Computer Science and Engineering, Vinayaka Mission's

KirupanandaVariyar Engineering College, rgsasi@gmail.com²

Dr.R. Reka, Professor and Head of Department of Computer Science and Engineering, Annai Mathammal Sheela Engineering College, rekasateesh@gmail.com³

Abstract

This review provides the feasible literature of drug discovery development process integrated with Machine Learning. The drug development process is a major challenge in the pharmaceutical industry since it takes a substantial amount of time and money to move through all the phases of developing of a new drug. Virtual Screening (VS) is robust and useful and is one of the most promising in silico techniques for drug design. VS attempts to predict the best interaction mode between two molecules to form a stable complex, and it uses scoring functions to estimate the force of non-covalent interactions between a ligand and molecular target. Once the drug target is identified, several interdisciplinary areas work Machine Learning (ML) methods to get enriched drugs. This ML methods are applied in every step of the computer aided drug design, and integrating these Artificial Intelligence and ML methods results in a high success rate of hit compounds. This focuses on how these new methodologies are being used in recent years of research. Analyzing the state of the art in this field will give us an idea of where chem. informatics will be developed in the short term, the limitations it presents and the positive results it has achieved.

Keywords: *drug discovery, high throughput screening, drug target, molecular target, rate of hit compounds, chem. Informatics.*

1. INTRODUCTION

Now a days drug development is a very time-consuming and capital intensive process. A 2014 study

found that the cost of developing a prescription drug is on average \$2.87 billion [1], and has likely since increased. Getting a potential drug through development stages and clinical trial phases takes years to complete and often compounds fail before ever reaching the market. One of the more problematic processes in the drug development pipeline is the early stage of identifying potential drug leads among thousands to millions of candidate compounds. High-throughput Screening (HTS) of compounds is very time- and resource-heavy, especially when considering the low number of hits it produces. In an attempt to remedy this, many researchers choose to supplement the in HTS with Virtual Screening (VS). This in silico process is a faster and cheaper way of searching for potential leads and can be utilized to reduce the number of compounds put through HTS, thereby greatly increasing HTS's yield. Like all computational biology processes, it is important to note that VS is not a replacement for HTS, because any sort of simulation or computer approximation is never guaranteed to be accurate – rather, it is a tool to aid and be used in conjunction with experiments.

1.1. Virtual Screening

The first step in VS is the assembly of a compound database on which to conduct the screening. This may begin with pulling mass quantities from publicly available chemogenomics libraries such as ChEMBL [2], PubChem [3], or ZINC [4], each include tens of millions of compounds annotated with information about their structure, known targets, and in the case of ZINC, purchasability. It is also common for

pharmaceutical companies to use their own, inhouse compound databases – which may have come from drugs that did not pass all clinical trials – to conduct VS. Whether the initial dataset is queried from the internet or not, it must go through further filtering in order to discard infeasible compounds and lower the number of false positives. It is common for researchers to exclude compounds that are much larger than the binding site of their target, or ones that are not available for purchase within the desired timeframe. Most datasets also get filtered according to Lipinski's Rule of Five [5] or standard metrics for lead-likeness [6], which remove compounds that are unlikely to be good drugs or leads, respectively. This is a particularly important step because VS is not done in isolation, but rather for the purpose of developing a drug for production. Finding a compound that can bind to the target but cannot be properly physiologically absorbed does not accomplish this goal. Likewise, it is important to take precautionary measures to reduce the number of false positives, which take up time and resources in the hit-to-lead development and clinical trial phases only to ultimately fail. This can be done by removing compounds deemed to be pan-assay interference compounds (PAINS)

Once the dataset has been assembled, the next step is to perform the actual screening. This can be done in a structure-based or ligand-based manner, or with a combination of the two. StructureBased Virtual Screening (SBVS) involves examination of the structures of the ligand and target binding site and evaluation of the likelihood that the ligand will bind. This is most often done with docking, which involves “placing” compounds in the binding site of the target and scoring how likely they are to bind given a predetermined metric [8]. This method relies upon knowledge of each compound's structure, as well as the structure of the target. LigandBased Virtual Screening (LBVS) does not require structure information, but rather the molecular and chemical properties of known actives and the tested compounds [9]. The idea behind LBVS is that

undiscovered actives will share some chemical features with the known ones. While it seems counterintuitive to use a drug discovery method that requires already knowing viable compounds, it is possible that the preexisting compounds cause undesirable side effects, do not treat all stages of the disease, or target something that has developed resistance to them. There are advantages and disadvantages to both screening techniques. SBVS has the potential to discover actives with novel scaffolds, while LBVS is restricted to finding actives that share a limited number of predetermined chemical descriptors with known ligands. Additionally, if a target has no known actives, it is only possible to conduct SBVS. However, SBVS varies widely in accuracy due to approximations in physics, thermodynamics, and molecular positioning, and is dependent upon the use of a very accurate scoring function. In this way, LBVS is more dependable. It is therefore up to the researchers to decide which screen is more appropriate to their experiment – or how to combine the hits produced by using both screens on the same data. After obtaining hits from the VS, it is imperative to validate the results in vitro. Once one or more compounds have been experimentally verified as being able to bind to the desired target, they can undergo hit-to-lead development and clinical trials in order to hopefully be made into viable drugs. But how does one actually perform VS? It is not an option to simply perform an in silico HTS simulation with molecular dynamics, as this would be extremely computationally intensive and likely take an incredible amount of time to run. Instead, computational chemists are turning to ML in order to efficiently conduct VS.

1.2. Machine Learning

Machine learning (ML) is a subfield of Artificial Intelligence (AI), and the two are the biggest buzzwords in many technological fields today. It has led to incredible breakthroughs in image processing [11] and Natural Language Processing (NLP) [12], and is being

utilized in a number of other fields including sentiment analysis and autonomous vehicles. This versatility comes from the fact that ML constitutes generalizable methods of learning that only require large training datasets in order to perform well. The upsurge in chemical data availability makes ML viable for VS. Before learning about the applications of ML in VS, it is important to understand its general principles. While most computer programs require an input and some functions to produce an output, ML uses training inputs and outputs to generate a function, which it can then use on test inputs to produce corresponding outputs. A good ML implementation must follow the Structural Risk Minimization (SRM) principle: it strikes an ideal balance between being both generalizable to unseen testing data and not over fitting the training data. This is done by minimizing the confidence interval (which corresponds to over fitting) and minimizing the empirical risk (which is the average error for the training data). ML can be supervised or unsupervised. The former involves giving inputs already labeled with classifications and asking the computer to determine the classification pattern; the latter uses unlabeled inputs and require the computer to cluster similar data points in order to generate logical classes. Because the purpose of VS is to determine the activity of tested compounds, only supervised learning algorithms are used. Two important processes that are common to all forms of ML – and particularly to ML in VS – are dataset preparation and model validation. A good ML model learns from a thoughtfully curate training dataset and is applied to a distinct testing set. When used for VS, both sets must consist of compounds with labeled binding activity – the training set requires this in order to establish patterns that the model can learn, and the testing set requires this for evaluating the model's accuracy. Active compounds must be taken from experimental results. Inactives may also be selected this way, but it is not always the case that a chosen chemogenomics library will contain enough nonbinding compounds that have been tested against the target for which the VS is being

conducted. For this reason, many researchers opt to use decoy compounds, which are structurally similar to actives but have very different chemical features. The rationale behind this is that it is important to provide inactives that physically resemble actives to prevent the VS model from erroneously equating common structural features with activity, and that the decoys' chemical differences are sufficient to assume a high unlikelihood to bind. The most common method of obtaining decoys is through the use of directories such as DUDE . Dissimilarity between known actives and assumed inactives can be further enforced by also calculating the Tanimoto coefficient and excluding presumed inactives that are too similar to actives. The Tanimoto coefficient serves another key purpose; it can provide a measure of the dataset's diversity. Diversity is critical for the creation of good training and testing sets in order to make the resulting ML model as general as possible. For this reason, it is typical to calculate the average Tanimoto coefficient between all compounds in the dataset to ensure that it is sufficiently diverse. Once the overall dataset has been assembled, it is very likely that there will be an imbalance between the number of active and inactive compounds. This can be problematic for some ML methods. Potential ways to resolve this problem are negativeundersampling of inactives and/or positive-oversampling of actives. After all this preparation, the labelled dataset can be split into training and testing data. Most often, about 70% of the data goes to the training set and the remainder to testing [20-27], although an 80/20 split can also be used [28-30]. These splits are usually done randomly – however, it has been shown that a temporal-based split generally increases classifier accuracy. An alternative to splitting is k-fold cross-validation, in which the dataset is randomly split into k partitions of equal size. k-1 partitions are used as training data, and the final partition is the testing data. This process is repeated a total of k times, with each partition serving as the testing set exactly once. The final model is chosen based on the split that produced the lowest error.

The value of k is usually chosen to be 5 or 10. The special case when k is equal to the number of samples is called leave-one-out cross-validation. Regardless of how the testing set is separated from the training set, it is used in a process called internal validation in order to judge an ML model. This is often done by first calculating the confusion matrix of the model, which consists of the intersections between predicted actives/inactives and actual actives/inactives. The confusion matrix yields values that can help quantify the performance of any ML model: sensitivity, specificity, accuracy and the Matthews' Correlation Coefficient (MCC). In each of these equations, TP, FP, TN, and FN represent the number of true positives, false positives, true negatives, and false negatives, respectively.

The MCC is usually used to compare the performance of different models, with a perfect model having a score of 1. It is also very common to measure accuracy by way of the area under the receiver operating characteristic curve (AUC), which plots SP against $1 - SE$. Again, the closer this value is to 1, the better. An AUC of 0.5 indicates performance equivalent to random classification. ML for VS also often uses the Boltzmann-enhanced discrimination of receiver operating characteristic, an accuracy metric specifically designed to compare VS ranking methods. Now that we have reviewed the general workflows of both VS and ML, we can dive into specific ML techniques.

1. Virtual Screening Algorithms

In VS, we are targeting proteins in the human body to find novel ligands that will bind to them. VS can be divided into two classes: structure-based and ligand-based. In structure-based virtual screening, a 3D structure of the target protein is known, and the goal is to identify ligands from a database of candidates that will have better affinity with the 3D structure of the target. VS can be performed using molecular docking, a computational process where ligands are moved in 3D space to find a configuration of the target and ligand that maximizes the

scoring function. The ligands in the database are ranked according to their maximum score, and the best ones can be investigated further, e.g., by examining the mode and type of interaction that occurs. Additionally, VS techniques can be divided according to the algorithms used as follows:

- Machine Learning-based Algorithms
- Artificial neural networks (ANNs) ([Ashtawy and Mahapatra, 2018](#));
- Support vector machines ([Sengupta and Bandyopadhyay, 2014](#));
- Bayesian techniques ([Abdo et al., 2010](#));
- Decision tree ([Ho, 1998](#));
- k-nearest neighbors (kNN) ([Peterson et al., 2009](#));
- Kohonen's SOMs and counterpropagation ANNs ([Schneider et al., 2009](#));
- Ensemble methods using machine learning ([Korkmaz et al., 2015](#));
- Evolutionary Algorithms
- Genetic algorithms ([Xia et al., 2017](#));
- Differential evolution ([Friesner et al., 2004](#)), Gold ([Verdonk et al., 2003](#)), Surflex ([Spitzer and Jain, 2012](#)) and FlexX ([Hui-fang et al., 2010](#));
- Ant colony optimization ([Korb et al., 2009](#));
- Tabu search ([Baxter et al., 1998](#));
- Particle swarm optimization ([Gowthaman et al., 2015](#)) and PSOVina ([Ng et al., 2015](#));
- Local search such as Autodock Vina ([Trott and Olson, 2009](#)), SwissDock/EADock ([Grosdidier et al., 2011](#)) and GlamDock ([Tietze and Apostolakis, 2007](#));
- Exhaustive search such as eHiTS ([Zsoldos et al., 2007](#));
- Linear programming methods such as Simplex Method ([Ruiz-Carmona et al., 2014](#));
- Systematic methods such as incremental construction used by FlexX ([Rarey et al., 1996](#)), Surflex ([Spitzer and Jain, 2012](#)), and Sybyl-X ([Certara, 2016](#));
- Statistical methods
- Monte Carlo ([Harrison, 2010](#));

- Simulated annealing (SA) ([Doucet and Pelletier, 2007](#)), [Hatmal and Taha \(Hatmal and Taha, 2017\)](#));
- Conformational space annealing (CSA) ([Shin et al., 2011](#));
- Similarity-based algorithms
- Based on substructures ([Tresadern et al., 2009](#));
- Pharmacochemical ([Cruz-Monteagudo et al., 2014](#));
- Overlapping volumes ([Leach et al., 2010](#));
- Molecular interaction fields (MIFs) ([Willett, 2006](#));
- Hybrid approach ([Morris et al., 2009](#); [Haga et al., 2016](#));

After performing a VS simulation, it is necessary to verify whether the quality of the generated protein-ligand complexes can represent a complex that could be reproduced in experiments.

Vs Software Programs

There are several VS software programs using different docking algorithms that make a VS process easier for the researchers to execute by avoiding the need to have advanced knowledge of computer science and on how to implement the algorithms used in this task. In this regard, VS software can act as a possible cost reducer, since they function as filters that select from a database with thousands of molecules that are more likely to present biological activity against a target of interest. VS programs measure the affinity energy of a small molecule (ligand) to a molecular target of interest to determine the interaction energy of the resulting complex ([Carregal et al., 2017](#)).

[Table 1](#) summarizes the main characteristics of the most used software in VS. The first column contains the software used and its reference. The second column contains the type of software license: free for academic use, freeware, open-source, or commercial. The free for academic use license indicates that the software in question can be used for teaching and research in the academic world without a fee. However, it implies that the software has restrictions for commercial use. A freeware license indicates that the software is free. Thus, users can use it without a fee, and all the functions of the

program are available to be used without any restrictions. An open-source license indicates that the software source code is accessible so users can study, change, and distribute the software to anyone and for any purpose. Software developed under a commercial license indicates that it is designed and developed for a commercial purpose. Thus, in general, it is necessary to pay some licensing fee for its use. The third column indicates on which platforms the software can be used (Windows, Linux, or Mac). The next column indicates whether or not the software may consider protein flexibility during anchoring. The docking algorithm column lists the algorithms used by the software to perform the docking. The sixth column, called the scoring function, indicates which scoring functions are used by the software.

Classification Model	AR (%)	SE (%)	SP (%)	PPV (%)	NPV (%)	DR (%)	bAR (%)	FS (%)	MCC (%)	k
Discriminant Classifiers										
Linear discriminant analysis	72.69	89.80	58.47	64.23	87.34	40.74	74.14	74.89	49.90	0.467
Robust linear discriminant analysis	75.93	91.84	62.71	67.16	90.24	41.67	77.27	77.59	55.96	0.529
Quadratic discriminant analysis	69.91	87.76	55.08	61.97	84.42	39.81	71.42	72.57	44.53	0.414
Robust quadratic discriminant analysis	73.61	80.61	67.80	67.52	90.81	36.57	74.20	73.49	48.37	0.476
Mixture discriminant analysis	75.93	90.82	63.56	67.42	89.29	41.20	77.19	77.39	55.53	0.528
Flexible discriminant analysis	78.24	89.80	68.64	70.40	89.01	40.74	79.22	78.92	58.92	0.571
Nearest shrunken centroids	74.07	91.84	59.32	65.22	89.74	41.67	75.58	76.27	53.03	0.494
Decision Tree Classifiers										
Classification and regression trees	72.22	88.78	58.47	63.97	86.25	40.28	73.63	74.36	48.71	0.457
CS.0	78.24	89.80	68.64	70.40	89.01	40.74	79.22	78.92	58.92	0.571
J48	77.31	89.80	66.95	69.29	88.76	40.74	78.37	78.22	57.40	0.554
Conditional inference tree	73.61	86.73	62.71	65.89	85.06	39.35	74.72	74.89	50.19	0.482
Kernel-based Classifiers										
Support vector machines with linear kernel	76.39	87.76	66.95	68.80	86.81	39.81	77.35	77.13	55.16	0.535
Support vector machines with radial basis function kernel	77.78	90.82	66.95	69.53	89.77	41.20	78.88	78.76	58.53	0.563
Partial least squares	74.07	91.84	59.32	65.22	89.74	41.67	75.58	76.27	53.03	0.494
Least squares support vector machines with linear kernel	73.15	90.82	58.47	64.49	88.46	41.20	74.65	75.42	51.09	0.476
Least squares support vector machines with radial basis function kernel	78.70	87.76	71.19	71.67	87.50	39.81	79.47	78.90	59.05	0.578
Ensemble Classifiers										
Random forests	76.85	88.78	66.95	69.05	87.78	40.28	77.86	77.68	56.27	0.544
Bagged support vector machines	76.39	88.78	66.10	68.50	87.64	40.28	77.44	77.33	55.51	0.535
Bagged k-nearest neighbors	75.46	90.82	62.71	66.92	89.16	41.20	76.76	77.06	54.79	0.520
Other Classifiers										
Naive bayes	68.06	88.78	59.85	60.00	84.51	40.28	69.81	71.60	41.99	0.381
Neural networks	77.31	86.73	69.49	70.25	86.32	39.35	78.11	77.63	56.39	0.551
k-Nearest neighbors	76.85	90.82	65.25	68.46	89.53	41.20	78.04	78.07	57.03	0.546
Learning vector quantization	74.07	87.76	62.71	66.15	86.05	39.81	75.23	75.44	51.33	0.491

AR: Accuracy rate, SE: Sensitivity, SP: Specificity, PPV: Positive predictive value, NPV: Negative predictive value, DR: Detection rate, bAR: Balanced accuracy rate, FS: F-scores, MCC: Matthews correlation coefficient, k: Kappa statistic. Bold values indicate the top three winner algorithms in each performance measure

doi:10.1371/journal.pone.0246000.t001

To identify drug candidates for known therapeutic targets, identify the target, and the mechanism of action of active molecules discovered by phenotypic tests, anticipate potential side effects or drug-drug interactions by identifying secondary protein targets, or suggest repositioning of known drugs in new therapeutic indications. The numerical representation of molecules and proteins, since it is a crucial prerequisite for machine learning methods. We also introduced available toolkits and databases that are considered throughout the manuscript. In the second part of the manuscript, we first

reviewed methodological advances in machine learning algorithms for drug virtual screening, based on expert-based extracted descriptors or similarity measures. More precisely, we first discussed the main frameworks in drug-target interaction prediction: docking, ligand-based and the chemogenomics frameworks. Then, we reviewed data-blinded approaches whose input feature vectors are either handcrafted in an ad hoc manner, or calculated by toolkits that were designed based on chemists expertise. This allowed us to propose NN-MT, a multitask Support Vector Machine algorithm that is trained on a limited number of data points, in order to solve the computational issues of proteome-wide SVM for chemogenomics. We showed that the prediction performances of NN-MT are, at an efficient calculation cost, similar or better than various state-of-the-art methods. NN-MT was particularly efficient when predicting (protein, ligand) interactions in the most difficult double-orphan case, i.e. when no interactions are previously known for the protein and for the ligand. The NN-MT algorithm appears to be an appropriate default method providing state-of-the-art or better performances, in a wide range of prediction scenarios : proteome-wide prediction, protein family prediction, test (protein, ligand) pairs dissimilar to pairs in the 183

CONCLUSION

However, predicting drug-target interactions is still a challenging problem. Indeed, only a relatively small portion of the chemical space of molecules has been explored and assayed so that the data statistics limit the predictive power of machine learning algorithms. In this perspective, we thus emphasise that machine learning methods should be evaluated and optimised toward unexplored chemical subspaces, and should particularly focus on the double orphan prediction case. The cystic fibrosis project illustrates how chemogenomics approaches can propose realistic targets for drugs identified by phenotypic tests, and how analysis of the predicted targets can explain and predict drug-drug

interactions. At a fundamental level, it shows how chemogenomics can help identify molecular mechanisms at the origin of the disease symptoms, and explain the relations between the pathways perturbations. Most of all, it shows how target predictions provides comprehensive interpretation of biological results reinforcing the interest of the corresponding experimental results. This project is also a striking example of how chemogenomics can guide future experiments, and even in our case, suggest a whole research program in translational medicine based on drug repositioning. It also illustrates how different types of data can be used to complement chemogenomics results. In this case, independent biological data were not integrated within the prediction step, but used to help analysis of the predictions and identify the most promising biological pathways to be targeted in TNBC. It again illustrates how prediction of targets can help interpretation of biological experiments and help them "make sense". Indeed, drug virtual screening is a creative force that provides different visions and proposes other research directions.

REFERENCES

- [1] Gasteiger, J., Eds. Handbook of Chemoinformatics; Wiley-VCH: Weinheim, 2003.
- [2] Inglese, J.; Auld, D. S.; Jadhav, A.; Johnson, R. L.; Simeonov, A.; Yasgar, A.; Zheng, W.; Austin, C. P. Proc. Natl. Acad. Sci. USA, 2006, 103, 11473-11478.
- [3] Buttingsrud, B.; Ryeng, E.; King, R. D.; Alsberg, B. K. J. Comput.- Aided Mol. Des., 2006, 20, 361-373.
- [4] Srinivasan, A.; Page, D.; Camacho, R.; King, R. Mach. Learn., 2006, 64, 65-90.
- [5] Enot, D. P.; King, R. D. In Knowledge Discovery in Databases: Pkdd 2003; Proceedings, 2003, pp. 156-167.
- [6] Marchand-Geneste, N.; Watson, K. A.; Alsberg, B. K.; King, R. D. J. Med. Chem., 2002, 45, 399-409.
- [7] Sternberg, M. J. E.; Muggleton, S. H. QSAR Comb. Sci., 2003, 22, 527-532.
- [8] Mitchell, T. M. Machine Learning; McGraw-Hill: Singapore, 1997.

- [9] Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; SpringerVerlag: New York, 2001.
- [10] Witten, I. H.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*; Morgan Kaufmann: San Francisco, 2005.
- [11] Hansch, C.; Fujita, T. *J. Am. Chem. Soc.*, 1964, 86, 1616-1630.
- [12] Cramer, R. D., III; Patterson, D. E.; Bunce, J. D. *J. Am. Chem. Soc.*, 1988, 110, 5959-5967.

SURVEY ON AN EFFICIENT CREDIT CARD FRAUD DETECTION USING BIG DATA ANALYTICS WITH MACHINE LEARNING APPROACHES

MS V.SUGANTHI

*Assistant Professor,
Department of Computer Science
S.S.K.V college of Arts and Science for women,
Kanchipuram.
E-mail: ¹suganthinethaji17@gmail.com.*

Dr.J.JEBATHANGAM

*Associate Professor
Department of Computer Science
Vels Institute of Science, Technology
Advanced Studies(VISTAS)
E-mail: ²jthangam.scs@velsuniv.ac.in.*

Abstract

A credit card is a payment card provided by every bank to eligible customers (cardholders) to make day-to-day transactions. In real life, it is not easy to track the transaction done by the fraudulent because the technique used is very simple like pattern matching techniques. Due to these reasons, now banks and other transaction partners are going for the better option and imperative methods for detecting the frauds in transactions. Therefore, in this paper, there are various approaches are identified for credit card fraud detection using Big Data Analytics with the machine learning algorithm and that has been surveyed and compared with advantages and disadvantages of each method.

Keywords — *Credit Card, Fraud Detection, Fraud analysis, Imbalanced data set, machine learning techniques, big data, Detection Techniques.*

1. INTRODUCTION

A credit card is a payment card provided by every bank to eligible customers (cardholders) to make day-to-day transactions. Using the card, a cardholder can pay for goods and services without having money in their account at the particular moment and can be paid back to banks later point in time. The legitimate transactions made by the cardholder provide a pattern of his/her expenditures. If a card is stolen or accessed by some fraudsters, the transactions show an abnormal expenditure pattern, and such a transaction is called a fraudulent transaction. In today's era, the people of smart societies are paying more money using debit/credit cards while purchasing something online/of-line because it is heavy and uneasy to carry the wallet with a huge amount of money and this is becoming the basic reason for the drastic increase in the rate of fraud. In real life, it is not easy to track the transaction done by the fraudulent because the technique used is very simple like pattern matching techniques. Due to these reasons, now banks and other transaction partners are going for the better option and imperative methods for detecting the frauds in transactions. Therefore, in this paper, we explain the current fraud detection systems and what are the problems they handled, and the present machine understanding approaches are used to increase the fraud detection rate and decrease the rate of fraud activity. Also, we have examined and analyzed these techniques. The remaining of this paper gives the following points: First, it specifies the theoretical views of fraud detection systems. Second, it explains the existing

machine learning algorithms used to face the banking fraud detection issues and challenges and provides a comparative study of these approaches based on different parameters. Third, it defines the outlook of our paper.

1.1 CREDIT CARD FRAUD

Credit card fraud is a common and fast-growing issue. Such problems may be tackled with Big Data Analytics, which, together with Machine Learning, should not be overlooked. If a card is stolen or accessed by some fraudsters, the transactions show an abnormal expenditure pattern, and such a transaction is called a fraudulent transaction.

1.2 CREDIT CARD FRAUD DETECTION

New advances in electronic commerce systems and communication technologies have made the credit card the potentially most popular method of payment for both regular and online purchases; thus, there is significantly increased fraud associated with such transactions. Fraudulent credit card transactions cost firms and consumers large financial losses every year, and fraudsters continuously attempt to find new technologies and methods for committing fraudulent transactions. The detection of fraudulent transactions has become a significant factor affecting the greater utilization of electronic payment. Thus, there is a need for efficient and effective approaches for detecting fraud in credit card transactions. Class imbalance with overlap is a very challenging problem in

electronic fraud transaction detection. Fraudsters have racked their brains to make a fraud transaction as similar as a genuine one in order to avoid being found. Therefore, lots of data of fraud transactions overlap with genuine



Credit Card Fraud Detection

Fig:1 Credit Card Fraud Detection

transactions so it is hard to distinguish them. However, most attention has been focused on class imbalance rather than overlapping issues for machine-learning-based methods of fraud transaction detection.

2. BIG DATA

Big Data is a collection of large volume of data both structured and unstructured. It is a data with so large size and complexity that none of traditional data management tools can store it or process it efficiently. Big data can be analyzed for perception, that lead to better decisions and strategic business moves. With large amounts of information streaming in from countless sources, banks are faced with finding new and innovative ways to manage big data. While it's important to understand customers and boost their satisfaction, it's equally important to minimize risk and fraud while maintaining regulatory compliance.

2.1 BIG DATA CHALLENGES

The Big Data challenges are based on its V's characteristics which are enlisted as follows:

• Volume:

In credit card transactions, every financial organization deals with huge amount of credit card transactions in each and every minute or an hour. Here, huge amount refers to volume of data. With the dynamic growth of data sets, multiple classification tasks will lead to decreased accuracy. The data in the Credit Card Fraud Detection is very huge and the number of attributes that used to define the pattern of a cardholder is around 31 attributes on average. To solve those problem there are different data minimizing techniques are implemented, Principal Component Analysis (PCA) for choosing the very important variables for collecting credit card transactions [6].

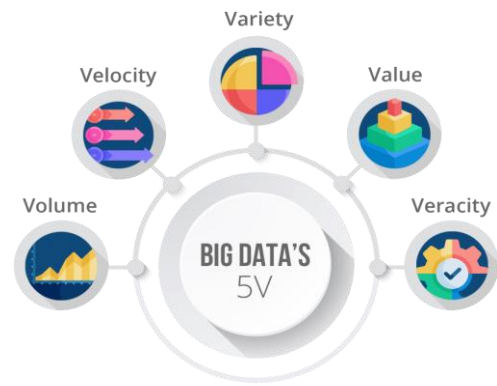


Fig:2 Big Data Challenges

• Velocity:

Velocity defines how quickly data are generated and how fast that data move. In credit card transactions, every financial organization deals with huge amount of credit card transactions in each and every minute or an hour. Here, number of transactions per hour denotes the velocity.

• Veracity:

Every bank to eligible customers (cardholders) are changed their patterns frequently. Detection Accuracy was low and it will take high computation time in existing classification techniques with imbalanced overlapped data.

• Value:

Class imbalance with overlap is a very challenging problem in electronic fraud transaction detection. Fraudsters have racked their brains to make a fraud transaction as similar as a genuine one in order to avoid being found. Therefore, lots of data of fraud transactions overlap with genuine transactions so it is hard to distinguish them. However, most attention has been focused on class imbalance rather than overlapping issues for machine-learning-based methods of fraud transaction detection.

• Variety:

Variety of data specifies various forms of credit card data sets. Banking and financial organizations are stored the variety of data in many database formats that are identified by their own unique data model. These variety of data can be also in the assortment of processes developed by fraudsters to make frauds.

2.2 BIG DATA ANALYTICS

Nowadays, number of financial sectors have been implemented Big Data Analytic (BDA) techniques in their e-commerce systems to provide service for their own customers to make the

transaction online from anywhere and whenever they are. There are various ways to implement to those systems, such as Credit card payment, Telecommunication, and Insurance method systems, that are used by eligible customers fraudsters. New advances in electronic commerce systems and communication technologies have made the credit card the potentially most popular method of payment for both regular and online purchases; thus, there is significantly increased fraud associated with such transactions. Fraudulent credit card transactions cost firms and consumers large financial losses every year, and fraudsters continuously attempt to find new technologies and methods for committing fraudulent transactions. The detection of fraudulent transactions has become a significant factor affecting the greater utilization of electronic payment. Thus, there is a need for efficient and effective approaches for detecting fraud in credit card transactions. Class imbalance with overlap is a very challenging problem in electronic fraud transaction detection. Fraudsters have racked their brains to make a fraud transaction as similar as a genuine one in order to avoid being found. Therefore, lots of data of fraud transactions overlap with genuine transactions so it is hard to distinguish them. However, most attention has been focused on class imbalance rather than overlapping issues for machine-learning-based methods of fraud transaction detection.

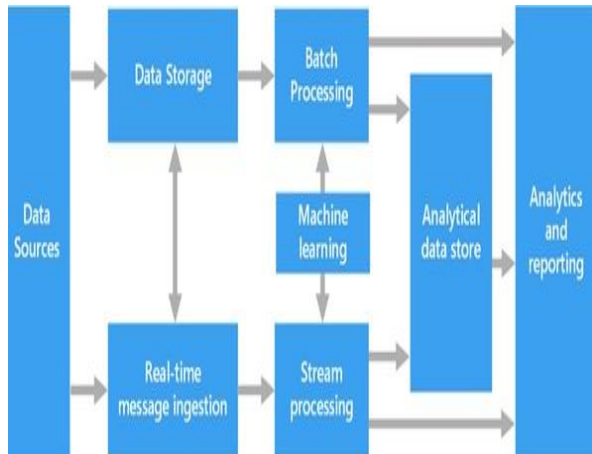


Fig:3 Big Data Analytics

3. LITERATURE SURVEY

[1] sara makki¹, zainab assaghir², yehia ta³ , rafiqul haque⁴, mohand-said hacid¹, hassan zeineddine.² “An Experimental Study with Imbalanced Classification Approaches for Credit Card Fraud Detection”, DOI 10.1109/ACCESS.2019.2927266, IEEE Access.

Fraud detection concerns a large number of financial institutions and banks as this crime costs them around \$ 67 billion per year. There are different

types of fraud: insurance fraud, credit card fraud, statement fraud, securities fraud etc. Of all of them, credit card fraud is the most common type. It is defined as an unauthorized use of a credit card account. It occurs when the cardholder and the card issuer are not aware that the card is being used by a third party. Many research works have been dedicated to the imbalance classification problem. Several solutions have been proposed in a large body of literature which, to the best of our knowledge, are built on machine learning and data mining algorithms. However, class imbalance remained an unsolved issue.

[2] John O. Awoyemi, Adebayo O. Adetunmbi, Samuel A. Oluwadare “Credit card fraud detection using Machine Learning Techniques: A Comparative Analysis” 978-1-5090-4642-3/17/\$31.00 ©2017 IEEE

Data mining technique is one notable methods used in solving credit fraud detection problem. Credit card fraud detection is the process of identifying those transactions that are fraudulent into two classes of legitimate (genuine) and fraudulent transactions. Credit card fraud detection is based on analysis of a card’s spending behaviour. Many techniques have been applied to credit card fraud detection, artificial neural network, genetic algorithm, support vector machine, decision tree and naïve bayes. The performance of bayesian and neural network is evaluated on credit card fraud data. Decision tree, neural networks and logistic regression are tested for their applicability in fraud detections. This paper evaluates two advanced data mining approaches, support vector machines and random forests. Detection of credit card fraud using decision trees and support vector machines is investigated and the results show that the proposed classifiers of decision tree approaches outperform SVM approaches in solving the problem under investigation.

[4] Asha RB , Suresh Kumar KR “Credit card fraud detection using artificial neural network”. The main objective of the research is to find a fraudulent transactions in credit card transactions. Comparison between the supervised learning and deep learning and deep learning algorithm outperformed based on accuracy. The existing systems are carried out by considering machine learning algorithms like Support Vector Machine, Naïve Bayes, k-Nearest Neighbor and so on and some of them used random dataset.

Very few have used artificial neural network for credit card fraud detection. The Proposed system uses the Artificial Neural Network to find the fraud in the credit card transactions. Performance is measured and accuracy is calculated based on prediction. And also classification algorithms such as Support vector machine and k-Nearest Neighbor are used to build a credit card fraud detection model.

4. EXISTING TECHNIQUES

The previous research methods consist of some drawbacks, which are enlisted as follows:

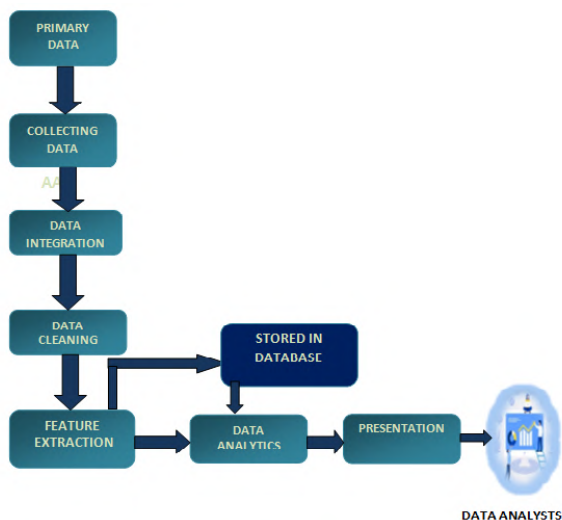
Existing system cannot effectively solve the massive Credit Card Fraud Detection problem that arises in the face of a real network application environment. With the dynamic growth of data sets, multiple classification tasks will lead to decreased accuracy.

It is a highly challenging task to create a training data set that will permit the algorithms to select the specific distinctiveness that makes a particular transaction more or less probable to be fraudulent in a cloud-IoT-based distributed environment.

More input features often make a predictive modeling task more challenging to model and are more generally referred to as the curse of dimensionality.

Detection Accuracy was low and it will take high computation time in existing classification techniques with imbalanced overlapped data.

Existing Big data technique that is Map and Reduce of HDFS framework is not good to use in the real-time data processing. The training time was more due to the back propagation technique in the existing neural network algorithm.



Number of comparative studies between the existing research papers related to the credit card fraud detection techniques:

	Fraud Detection Using Convolutional Neural Networks (2016)	Neural Network (CNN) model achieves the best performance	methods should focus on handling the issue of highly imbalanced data.
[14]	Horse Race Analysis in Credit Card Fraud—Deep Learning, Logistic Regression, and Gradient Boosted Tree (2017)	After comparing the predictive performance of 3 machine learning algorithms, we observe that Neural Network (NN) is the best one.	There is some limitations such as the less predictive power of Logistic regression, the large volume of data for GBT and the feature selection issue for NN.
[15]	An Efficient Way to Detect Credit Card Fraud Using Machine Learning Methodologies (2018)	Logistic Regression and Decision Tree have the most accurate results.	The machine learning models used in this work ignore the other performance metrics.
[16]	A Comparative Study of Machine Learning Techniques for Credit Card Fraud Detection Based on Time Variance (2018)	This paper compare between 10 machine learning algorithms without and with the "Time" feature to capture the performance differences.	Hence, the work should focus on a more efficient way to handle the high level of the data imbalance problem.

5. CONCLUSION

In today’s era, the people of smart societies are paying more money using debit/credit cards while purchasing something online/of-line because it is heavy and uneasy to carry the wallet with a huge amount of money and this is becoming the basic reason for the drastic increase in the rate of fraud. In real life, it is not easy to track the transaction done by the fraudulent because the technique used is very simple like pattern matching techniques. Due to these reasons, now banks and other transaction partners are going for the better option and imperative methods for detecting the frauds in transactions. Therefore, in this paper focused to find the most reliable machine learning algorithms such as Artificial Neural Network, Decision Tree, Naive Bayes, and Support Vector Machine using the result parameters such as Precision, Recall, F-Measure, Accuracy, Sensitivity, and Specificity.

REFERENCES

No of Ref	Title of the paper	Advantages	Limitations
[12]	Performance Evaluation of Machine Learning Algorithms for Credit Card Fraud Detection (2019).	Unsupervised techniques solve the missing value of dataset and provide best results than other techniques like Random Forest and Local Outlier Factor.	The work should be focusing on resampling techniques that reduce high imbalance rate.
[13]	Credit Card	The Convolutional	The future

[1] SARA MAKKI¹, ZAINAB ASSAGHIR², YEHIA TAHER³, RAFIQU HAQUE⁴, MOHAND-SAÏD HACID¹, HASSAN ZEINEDDINE.² “AN EXPERIMENTAL STUDY WITH IMBALANCED CLASSIFICATION APPROACHES FOR CREDIT CARD FRAUD DETECTION”, DOI 10.1109/ACCESS.2019.2927266, IEEE ACCESS.

[2] JOHN O. AWOYEMI, ADEBAYO O. ADETUNMBI, SAMUEL A. OLUWADARE “CREDIT CARD FRAUD DETECTION USING MACHINE LEARNING TECHNIQUES: A COMPARATIVE ANALYSIS” 978-1-5090-4642-3/17/\$31.00 ©2017 IEEE

[3] AISHA MOHAMMAD FAYYOMI, DERARELEYAN, AMINA ELEYAN “A SURVEY PAPER ON CREDIT CARD FRAUD DETECTION TECHNIQUES” INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH VOLUME 10, ISSUE 09, SEPTEMBER 2021 ISSN 2277-8616.

[4] ASHA RB , SURESH KUMAR KR “CREDIT CARD FRAUD DETECTION USING ARTIFICIAL NEURAL NETWORK”.

[5] EMMANUEL ILEBERI¹ , YANXIA SUN ¹ , (SENIOR MEMBER, IEEE), AND ZENGHUI WANG ² , (MEMBER, IEEE) “PERFORMANCE EVALUATION OF MACHINE LEARNING METHODS FOR CREDIT CARD FRAUD DETECTION USING SMOTE AND ADABOOST”

[6] A. ABDALLAH, M. A. MAAROF, AND A. ZAINAL, “FRAUD DETECTION SYSTEM: A SURVEY,” J. NETW. COMPUT. APPL., VOL. 68, PP. 90–113, JUN. 2016

[7] ALTYEB ALTAHER TAHA AND SHARAF JAMEEL MALEBARY, “AN INTELLIGENT APPROACH TO CREDIT CARD FRAUD DETECTION USING AN OPTIMIZED LIGHT GRADIENT BOOSTING MACHINE”, IEEE ACCESS, VOL. 8, PP. 25579-25587, 2020.

[8] Zhenchuan Li, GuanJun Liu and ChangJun Jiang, “Deep representation learning with full center loss for credit card fraud detection”, IEEE Transactions on Computational Social Systems, vol. 7, no. 2, pp. 569-579, 2020.

[9] C. REVIEWS, —A COMPARATIVE STUDY : CREDIT CARD FRAUD, I VOL. 7, NO. 19, PP. 998–1011, 2020.

[10] IBTISSAM BENCHAJI, SAMIRA DOUZI, BOUABID EL OUAHIDI AND JAAFAR JAAFARI, “ENHANCED CREDIT CARD FRAUD DETECTION BASED ON ATTENTION MECHANISM AND LSTM DEEP MODEL”, JOURNAL OF BIG DATA, VOL. 8, PP. 1-21, 2021.

[11] Y. DAI, J. YAN, X. TANG, H. ZHAO, AND M. GUO, “ONLINE CREDIT CARD FRAUD DETECTION: A HYBRID FRAMEWORK WITH BIG DATA TECHNOLOGIES,” IN 2016 IEEE TRUSTCOM/BIGDATASE/ISPA, TIANJIN, CHINA, 2016, PP. 1644– 1651.

[12] S. MITTAL AND S. TYAGI, “PERFORMANCE EVALUATION OF MACHINE LEARNING ALGORITHMS FOR CREDIT CARD FRAUD DETECTION,” IN 2019 9TH INTERNATIONAL CONFERENCE ON CLOUD COMPUTING, DATA SCIENCE & ENGINEERING (CONFLUENCE), NOIDA, INDIA, 2019, PP. 320–324.

[13] K. FU, D. CHENG, Y. TU, AND L. ZHANG, “CREDIT CARD FRAUD DETECTION USING CONVOLUTIONAL NEURAL NETWORKS,” IN NEURAL INFORMATION PROCESSING, VOL. 9949, A. HIROSE, S. OZAWA, K. DOYA, K. IKEDA, M. LEE, AND D. LIU,

EDS. CHAM: SPRINGER INTERNATIONAL PUBLISHING, 2016, PP. 483–490.

[14] G. RUSHIN, C. STANCIL, M. SUN, S. ADAMS, AND P. BELING, “HORSE RACE ANALYSIS IN CREDIT CARD FRAUD—DEEP LEARNING, LOGISTIC REGRESSION, AND GRADIENT BOOSTED TREE,” IN 2017 SYSTEMS AND INFORMATION ENGINEERING DESIGN SYMPOSIUM (SIEDS), CHARLOTTESVILLE, VA, USA, 2017, PP. 117– 121.

[15] T. CHOUDHURY, G. DANGI, T. P. SINGH, A. CHAUHAN, AND A. AGGARWAL, “AN EFFICIENT WAY TO DETECT CREDIT CARD FRAUD USING MACHINE LEARNING METHODOLOGIES,” IN 2018 SECOND INTERNATIONAL CONFERENCE ON GREEN COMPUTING AND INTERNET OF THINGS (ICGCIOT), BANGALORE, INDIA, 2018, PP. 591–597.

[16] S. RAJORA ET AL., “A COMPARATIVE STUDY OF MACHINE LEARNING TECHNIQUES FOR CREDIT CARD FRAUD DETECTION BASED ON TIME VARIANCE,” IN 2018 IEEE SYMPOSIUM SERIES ON COMPUTATIONAL INTELLIGENCE (SSCI), BANGALORE, INDIA, 2018, PP. 1958–1963.

A SYSTEMATIC STUDY ON APPLICATIONS OF DIGITAL IMAGE PROCESSING

¹Dr.D.Sasirekha, Assistant Professor, Post Graduate Department of Computer Science, Anna Adarsh College for Women, AnnaNagar, Chennai.

²Dr.A.Ambeth Raja, Head & Associate Professor, PG Department of Computer Science, Thiruthangal Nadar College, Selavayal, Chennai.

Abstract – The great tractability of the digital manner of image handling sorts a widespread variability of rectilinear and nonlinear progressions to made potential. The digital image processing procedures established have remained practical to imageries from an extensive choice of restraints. A methodical study on prominence of image handling and its claims to the arena of computer vision is conceded on view in this work. An image is well-defined as an array, or a conditions, of square pixels (components of picture) prescribed in rows and columns. Image processing is a way of transforming a spitting image into numerical method and carry out particular process on it, in demand to get a better-quality image and take out numerous supportive data from it. Image handling is one of the most favorable regions, which is smeared for eminence examination of goods where the perplexing duty lies in acknowledgement of the object and feature abstraction. This paper made challenge to afford an outline of the solicitation of image processing and their approach with limited set of rules that have stood utilized in diverse arenas of engineering which cascades under three important stages: acquirement of images, the region of interest and credentials of defects. This paper distillates on presentations of image handling in the arena of science and technology include computer vision, remote recognizing, feature mining, face recognition, forecasting, optical character recognition (OCR), finger-print discovery, optical organization, argument authenticity, microscope imaging, lane departure caution system, Non-photorealistic demonstration, medical picture handling, and morphological imaging.

Keywords – *Image, Digital Image, Acquisition, Feature extraction, OCR*

I. Introduction

Image handling commonly denotes to digital picture handling. It similarly discusses to photosensitive and referent image handling. In this work we, have offered an organized study on picture handling and its prominence to the arena of computer visualization. A spitting image comprises of sub-images frequently denoted as sections or regions of interest. Images repeatedly comprise clusters of matters every single of which is the foundation for a section. Furthest normally, image handling necessitates the imageries to be accessible in digitized method. For digitization progression, the input image is experimented on a distinct matrix and every trial or pixel is quantized by a stationary quantity of information's. This

progresses the digitized image. To show a digital picture, majorly it is transformed into a referent signal that is skimmed onto an output. Scientists use a comprehensive assortment of elementary technique of image elucidation while approving analog photographic systems [1]. This type of image handling is just constrained within the field of acquaintance of the specialist. So specialists may relate a mixture of individual information and information in image handling. In digital image processing, computer based procedures are technologically advanced to accomplish image processing practice [2]. Bearing in mind the benefits of digital image processing in contradiction of analog image handling and due to enormous quantity of algorithms accessible that can be utilized with the input data.

In digital image handling, few complications during processing such as noise conception, signal misrepresentation etc., can be diminished and detached during preprocessing procedure called signal dispensation. Later, due to the progression that occurred in digital image dispensation with assistance of supercomputers has turn out to be the developing practice of picture handling which is more adaptable, and also the economical one. Image handling has robust relation with processor vision and computer graphics. In this work, an assessment on digital image handling, realistic in numerous arena has been given with appropriate procedures.

II. Image processing algorithms used in different fields

2.1 Image processing in Computer vision

Computer vision has remained prolonged into the massive expanse of arena extending from recording fresh records into the abstraction of spitting image design and data elucidation [3]. Computer vision projects by consuming a set of rules and photosensitive instruments to arouse humanoid conception to repeatedly abstract respectable data from an item [4]. The prime determination of Computer Visualization is to produce prototypes and information abstracts and data on or after the imageries, while Spitting image Handling is roughly instigating computational conversions for pictures, such as perfecting, dissimilarity, amongst others [5]. Equated to conservative approaches that yield a extensive period and need refined laboratory study, computer visualization has remained prolonged into a outlet of simulated intellect and replicated humanoid conception. It also united with illuminating

structures to simplify picture acquisition sustained with picture exploration. In further aspect, the phases of image scrutiny are: 1) picture development, in which picture of entity is apprehended and deposited in processor; 2) picture preprocessing, whereby eminence of picture is enhanced to increase the spitting image feature; 3) picture separation, in which the entity picture is acknowledged and detached from the circumstantial, 4) picture dimension, where numerous important structures are quantized, and 5) picture elucidation, where the mined imageries are then inferred [6].

Exploration on picture handling has stood engaged to encounter appliance erudition and figuring procedures that know how to distinguish designs of progressively more assorted matters. Mechanism erudition is thoroughly correlated to computational measurements which comprised of junk straining, optical character acknowledgement, exploration locomotives and computer visualization. A conventional computer vision data flow model was shown in fig.1.

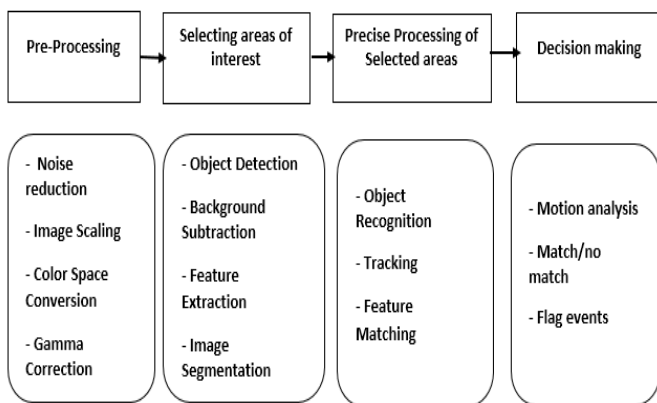


Fig.1 A conventional computer vision data flow

2.2 Image processing in Medical field

We predominantly contribute the significance to medicinal handling as the important request of picture handling. Medical imaging is the technique and method utilized to produce imageries of the humanoid form (or fragments and occupation thereof) for medicinal tenacities (medicinal measures beholding for to disclose, analyze, or scrutinize illness) or medicinal science (counting the study of physiology and customary examination). The mutual solicitations of picture handling in the ground of medicinal are Gamma ray handling, PET examination, X Ray handling, Medicinal CT, Picturing in the electromagnetic group, Imaging in the microwave band. In recent times established a content-based mammogram recovery structure as an investigative support to the radiologists in their elucidation of mammograms [7]. Brain MRI know how to be utilized to identify glioma, HIV and tumor metastasis, in the comparable way mammograms are utilized to perceive breast tumor and CT scans are engaged to identify vascular ailments. Digital dermoscopy is a broadly utilized non-invasive instrument that

associates ophthalmic exaggeration and superior lighting practices to condense an enhanced dermoscopic picture for medical analysis of malignance. Dermatologists have recurrently applied this instrument for numerous periods to examine the superficial arrangement of humanoid coating that is imperceptible to the unprotected eyes [8, 9]. Flat the covering ailments like eczema, spots, malignance, mycosis, and so, can also be acknowledged by tiny pictures. The RGB measure of the imageries are engaged into contemplation to investigate the ailments. Dissimilar color shadows using hue capacity assessment and YCbCr like blackish, reddish, bluish, whitish and grayish are utilized to discriminate the area of concern.

The disparity of the picture can be accustomed further accurately by plotting the area of Interests with standardization of the dark picture greatness. The ANN system stretches the finest presentation as it disregards the circumstantial and spectacles the mandatory share of a picture that we requisite. This image handling procedure is one and only the greatest effective methods of identifying lung malignancy [10]. For instance contemplate the discovery of cancer in brain which is established using the flowchart shown in fig.2 and the images which is shown in fig.3 [11].

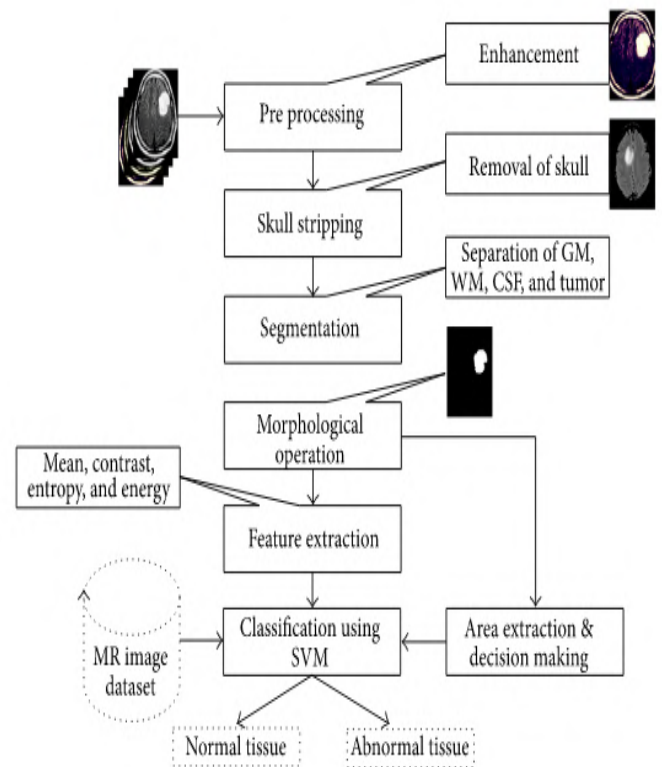


Fig.2 Flowchart for brain tumor segmentation part

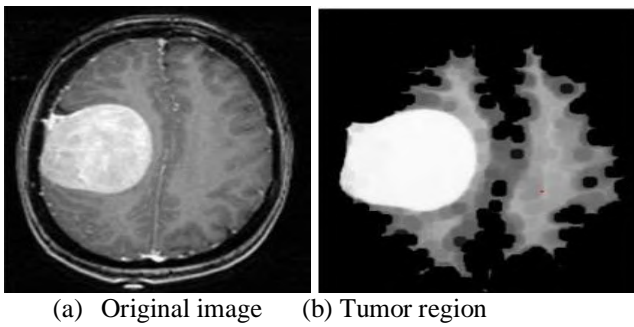


Fig.3 Detection of tumor in brain

2.3 Image processing in Biometrics

Biometrics or biometric authentication denotes to the spontaneous credentials of individuals by their performances or individualities. Biometrics is reprocessed in computer science as a category of credentials and admittance switch. It is also utilized to distinguish entities in clusters that are underneath investigation. Biometric identifiers are the incomparable, computable appearances utilized to mark and designate entities. Biometric identifiers are consistently categorized as physical versus developmental physical individualities. Functional individuality is correlated to the natural surroundings of the physique. Limited samples contains thumb print, face acknowledgement, Palm print, DNA, pointer geometry, iris acknowledgement, retina and odor or smell. Developmental physical individualities are associated to the design of presentation of an individual, comprising but not restricted to capturing beat, vocal sound and pace. Certain scientists have developed the period performance metrics to designate the concluding period of biometrics. Block diagram of biometric process was displayed in fig.4.

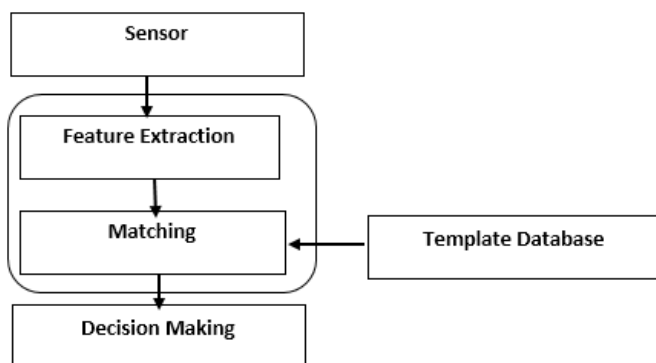


Fig.4. Block Diagram for biometric process

2.4 Image processing in Food Processing

Image handling is projected to utilize for nutrition security and criterions. Furthermore, this is a present-day procedure utilized to guarantee customer contentment frequently in a variety of food interrelated arenas. Predominantly emphases on the discovery of debasement,

guarantees virgin fruitlets, tubers and substance to consumers. Even picture handling is utilized to separate foodstuffs and crisscross & estimate nutriment fabricating utensils. This altogether ultimately penetrating to encounter excellence in life-saving goods, particularly in foodstuffs.

For example let us perceive the two dissimilar forms of popcorn imageries in the dataset: the first set embraces reflectance style imageries, and the second set contains transmittance type imageries. Reflectance imageries are analogous to consistent camera imageries. In this approach, the sunlit replicated from the seeds is apprehended. In the transmittance approach, the sunlit that permits through the popcorn seeds is seized. In many kernels, blue-eye destruction is further perceptible in transmittance-mode imageries than in reflectance-mode imageries [12]. Examples of transmittance-style and reflectance-style imageries are displayed in fig.5.

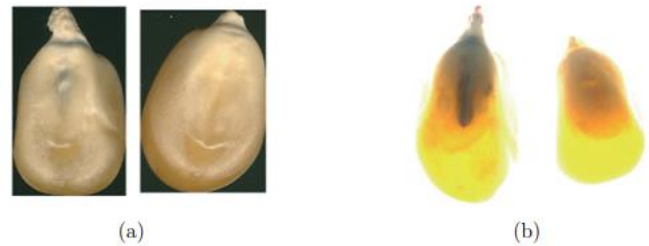


Fig.5. (a) Spoiled (left) and unspoiled (right) popcorn seed images picked up in the reflectance style. (b) Spoiled (left) and unspoiled (right) kernel imageries that were acquired in the transmittance style.

2.5 Image processing in Fingerprint Verification and signature recognition

The Fingerprint Authentication resentment is a worldwide antagonism concentrated on fingerprint authentication software valuation. A subcategory of impersonations of thumb print acquired with dissimilar sensors was presented to recorded participants are permit to change the constraints of particular procedures was shown in fig.6. Members were fascinated to compromise join up and match executable proceedings of their set of rules; the assessment was transfer on view at the controllers' conveniences using the succumbed operative collections on a seized record acquired with the matching instruments as the preparation set.

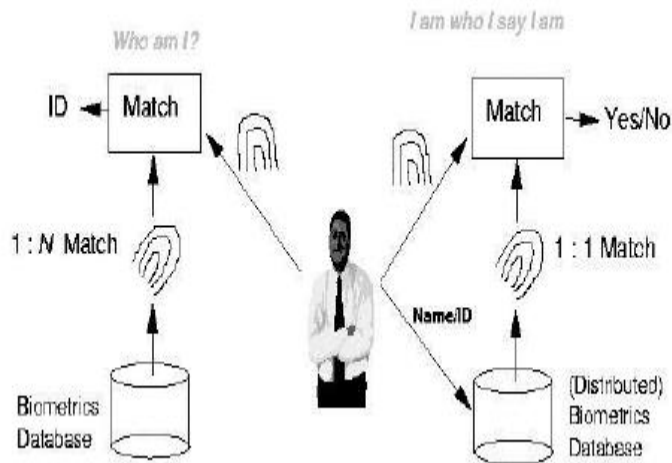


Fig.6. Processing of images in Fingerprint Verification

Signature authentication and acknowledgement is also a significant solicitation, which is to resolve, whether a signature be in the right place to a specified signer grounded on the picture of autograph and a limited trial imageries of the innovative autographs of the signer. Handwritten autographs are inaccurate in natural surroundings as their bends are not all the time strident, lines are not effortlessly conventional, and bends are not essentially even. Additionally, the typefaces can be strained in dissimilar dimensions and alignment in distinction to calligraphy which is repeatedly presumed to be transcribed on a starting point in an erect location. Consequently, a vigorous handwritten signature acknowledgement structure has to interpret for all of these influences [13].

2.6 Image processing in Satellite imaging

Remote recognizing information's encompasses of spatial, spectral and temporal tenacity. Spectral statistics is extensively through practice in remote detecting image classification. The chief feature in accurateness of powdered object is spatial tenacity. Temporal tenacity supports to produce land cover drawings for environmentally friendly forecasting, land use change discovery, metropolitan forecasting and so forth.

Image improvement progresses the image superiority and data content beforehand extra handling is conceded out. Some normally used procedures comprise contrast augmentation, spatial filtering and so forth. The rectilinear divergence improvement is greatest smeared to remote detecting imageries with Gaussian or near-Gaussian histograms where all the illumination standards commonly descent within a particular slender variety of the histogram [14]. However there emerges a state of affairs fundamentally for predictions with big land and water forms. Non-linear distinction perfections be able to be pragmatic on truncated comparison descriptions of which histogram distribution is a protruding method. Non blend grounded improvement contributes low three-dimensional data on the other hand with extraordinary computational intricacy. To overawe the

restriction of great convolution, mixture centered augmentation is utilized.

Pixel centered feature mining procedures are utilized to take out small side by side structures, which do not proceeds into the description of statistics regarding altitudinal associations. Small level topographies are removed in a straight line from the underdone, strident pixels with verge recognition being the furthestmost normally utilized procedure. Object centered methods are utilized to acquire great level topographies, which signify outlines in imageries that are perceived regardless of brightness, interpretation, alignment and measure. Great level topographies are mined contingent on the data from small level topographies. Great level feature mining is mostly utilized for computerized entity discovery and abstraction. Contingent on the altitudinal determination of the foundation image, various separation procedures are utilized. For small to standard tenacity descriptions, grouping set of rules are enhanced superior, conversely for great determination descriptions, multiresolution separation provides restored outcomes. On an extraordinary altitudinal determination image, comprehensive symmetrical structures can straightforwardly be identified, while the multispectral imageries encompass wealthier spectral figures. The competencies of the imageries can be improved if the benefits of together high spatial and spectral tenacity can be combined into one solitary image. The comprehensive topographies of such a combined picture as a result can be effortlessly renowned and will be useful for voluminous solicitations, such as built-up and environmentally friendly works. In revolution discovery, it is hard to discover the technique which is appropriate for identifying the variations which have happened. Collection of appropriate procedure for variation discovery is disturbed by numerous aspects like disparity of corporeal appearances of topographies with period, inappropriate recording of the imageries, effect of haze and so into view and it is rather a problematic duty to discover a solitary process since the natural surroundings of issue regulates which technique is top fitted. Methods like spitting image differencing and image ratioing can be utilized only at what time the variation and no transformation data is essential. If a comprehensive environment is essential, upright classification scene discovery is a respectable choice. Image-classification methods, either pixel centered or object centered, are utilized for translating a picture into a thematic chart. The formation of numerical surface and terrain models (DSMs/DTMs) from settlement imagery has come to be a corporate way to scrutinize the construction and improvement of plant life and geomorphological natural feature of earth's exterior.

In recent times, picture handling over substantial design also called quantum picture handling and solicitation of machine learning from side to side quantum calculation or quantum knowledge have delivered a novel viewpoint to great measure of image handling. Quantum picture handling utilizes quantum assets to scramble imageries for enlightening packing and period efficacy of convinced maneuvers like image

revolution. Thus quantum procedures are being extensively prolonged to enlightening alphanumeric or conservative image handling submissions and tasks. The convolution of satellite image additionally positions a difficulty in actual world organization solicitations and be able to be resolved by means of comprising computationally intellectual prototypes like machine knowledge.

2.7 Image processing in Face Detection

Face discovery is reliant on supercomputer knowledge which inaugurates the dimensions and localities of humanoid faces in inconsistent (digital) imageries. It discovers facial structures and disregard such as constructions, physiques and vegetation. Face acknowledgement can be perceived as a further mutual situation of face adaptation. In face adaptation, the procedure is to discover the locations and dimensions of a recognized quantity of faces. In face acknowledgement, one does not have this accompanying in advancement.

We produced representative masked-face imageries for prototypical working out over appearance discovery, the uncovering of strategic topographies, and mask reporting examination. The method of masked face picture group is revealed in fig.7.

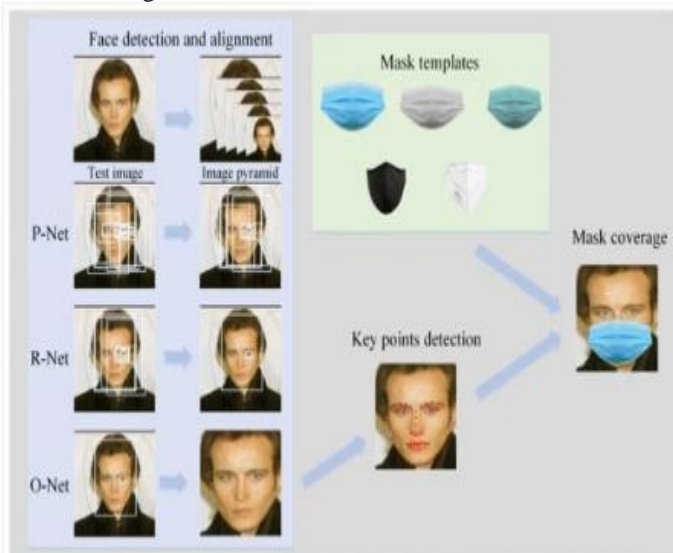


Fig.7. The practice of masked face appearance group Majorly, the face in the trial picture is identified and associated; then, the data about the jaw and the association of the snout is yield; lastly, the mask patterns are utilized to shield the look to produce the masked face picture. In Face exposure, we utilized a multi job spilled convolutional neural network (MTCNN) [15] for preprocessing, and got a spitting picture comprising merely faces. The outcome is displayed in Fig.7.

2.8 Image processing in Manufacture Industry

Image handling can be comprehended as computerized scheme to accomplish the fitness observing and

assessment classification to measure the injury happening in concretes and arrangements due to regular catastrophes [16]. This covenants with the description of crash arrangements, dimension of stress arenas exposed to dissimilar masses. The computerization in representation pointers to evaluating and associating the whole crash extent over an epoch of time, eradicating the humanoid error with high accurateness [17]. Procedures has been established to discover a crash advanced on the superficial of the material erection grounded on image developed by means of an automated robot was displayed in fig.8 [18].



Fig.8. Mobile robot for image acquisition

2.9 Image processing in Traffic monitoring

The contemporary revelation transmits to an amount of discovery headline for, customarily to the presentation of picture handling practices to road traffic flow records acquirement using images or videos. The discoveries occur in a scheme of traffic observing, the significant work of which is for acquirement of traffic data and discovery of occurrence. Added characteristically, the solicitation of picture dispensation approaches for the motor vehicle discovery, from the sequence of video imageries, as well as the acquirement of traffic information and discovery of traffic event. In a separate surface, the current expansion delivers a procedure of handling imageries acknowledged from an organization of traffic observing which is video centered. In one extra feature, the present improvement is regulated to a area of Interest i.e., for decision of an automobile which is touching and an auxiliary article is focused to a procedure of discovering diurnal or nocturnal location in observing a road traffic structure. It is the solicitation of a diversity of procedures to a road traffic observing arrangement grounded on video is also restrained inventive. Further imaginative representative of the contemporary day observing of road traffic structure is outlined in the proclaims.

2.10 Image processing in Medical Palmistry

Palmistry is a knowledge which perceives humanoid palm by dissimilar characteristics and originates inferences about natural surroundings of the individual. Subsequently from prehistoric epochs, countless societies like Indian, Chinese, Persian, Egyptian, Roman and Greek, people were used to get supervision around their current and forthcoming through means of palmistry. It comprises characteristics of humanoids like, fitness, consciousness, mental power,

standard of living and former correlated objects. Medical palmistry can be measured as unique of the subdivisions of palmistry. By means of utilizing this medical palmistry, feasible ailments can be recognized by detecting particular codes in humanoid palms such as Iceland, annoyed, vent, spot, star, square and circle. Furthermore figures of palm and limbs also perform precise significant role in such resolution building for credentials of ailments [19].

2.11 Image processing in Character Recognition

Pattern acknowledgement encompasses study commencing from image handling and starting from numerous added turfs that comprises machine knowledge. In design acknowledgement, picture handling is utilized for classifying the matters in imageries and then machine learning is utilized to sequence the structure for the variation in design. Pattern acknowledgement is utilized in supercomputer assisted analysis, acknowledgement of handwriting, acknowledgement of imageries etc. Character acknowledgement, usually known as optical character recognition. It is machine-driven or microelectronic conversion of imageries of either one typewritten or reproduced transcript (habitually apprehended by a scanner) into computer modifiable text. It is an extensive range for investigators in design acknowledgement, simulated intellect and appliance visualization [20]. Aimed at countless manuscript input tasks, character acknowledgement is the utmost charge in effect and immediate scheme obtainable.

It is normally used as an arrival of archives contact from a tiny generous of innovative numbers foundation, whether credentials, proof of purchase, bank proclamation, takings, business cards, a quantity of published accounts or mail. It is a conventional procedure of digitizing published documents such that they can be able to be by microelectronic revenues corrected, examined, supply more carefully utilized in appliance procedures such as appliance conversion and exhibited operational, manuscript to language, strategic records removal and text removal. OCR is a field of research in intellect, design and computer visualization. Premature forms mandatory to be computerized with imageries of each appeal, and performed on unique typeface at a stretch. "Intelligent" buildings with an unlimited mark of appreciativeness precision for most typestyles are now steady. Some merchantable approaches are accomplished of repeating arranged productivity that actual far bear a resemblance to the innovative perused sheet together with columns, imageries and other non-textual mechanisms [21].

2.12 Image processing in forensics

Digital image forensics (DIF) purposes at provided that trappings to upkeep visionless examination. This variety novel discipline branches from prevailing hypermedia security-related investigation areas like Watermarking and Steganography and adventures picture handling and examination apparatuses to make progress data around the

past of a picture. Twofold foremost investigation trails change below the term of Digital Image Forensics. The major one comprises approaches that challenge at responding enquiry a), by execution certain kind of airborne scrutiny to classify the scheme that apprehended the picture, or at slightest to regulate which maneuvers did not seizure it. These approaches will be composed in the succeeding beneath the mutual term of picture foundation device credentials methods. The another cluster of approaches targets as a substitute at revealing hints of semantic management (i.e. forgeries) by reviewing contradictions in ordinary image information [22].

III. Conclusion

In this work we have examined the numerous solicitations of digital picture handling in dissimilar regions. New discoveries in image handling expanse will converse the domain. Progress investigators in picture handling and synthetic intellect will comprise speech instructions, linguistic conversion, diagnosing and following individuals and identifying and pursuing individuals and possessions, detecting medicinal circumstances, executing maneuver and surgical procedure. We can furthermore practice the digital picture handling method in verdicting the incongruities of breathing organization, which hints to the discovery of covid-19. This review will support the scientists functioning on numerous fields such as picture handling, liability discovery in manufacturing Industries, and medicinal picture separation and moreover useful for the scholars of numerous fields. Ultimately probable for exploit investigation in computer assisted image handling methods are considerable and experiments to be resolved and until now to discover in each arena are unlimited. The theory and approach to appliance these method is practically analogous in all arenas which can gather to the novel investigators with widespread prospects. Finally it is recognized that extensive inventions and explores are in growth in and everywhere the computer assisted digital picture handling and this working to chief the forthcoming universally. In recent situation, machine learning is a developing investigation region with its capability to mark data determined conclusions effectively and acquire and accomplish perceptively. The excellence of the involvement picture and the complication of image topographies are certain aspects answerable for determining the picture handling procedure to be functional. Examiners are presently being prolonged into crossbred picture handling procedures to advance the toughness of the prevailing procedures. The upcoming study can be stretched to smear the illuminated methods in numerous everyday extents in distant recognizing and encompassing the enactment of quantum systems for distant recognizing submissions.

References

- [1] M. Petrik, et al, "Digital Img Handling Of Structure Response", Eng Mechs, 18th Intl Conf, Pgs: 244–245, 2012.

- [2] Anil K. Jain, A guide of "Fund of Digi Imge Procsng", 1989.
- [3] Patel, et al, "Machine vision system: a tool for quality inspection of food and agricultural products." *Jornl of food sci and tech* 49, no. 2 (2012): 123-141.
- [4] Matiacevich S, et al, "Quality Parameters of Six Cultivars of Blueberry Using Computer Vision", *Intl Jrnl of Food Sci.* 2013
- [5] Babatunde, et al, "A survey of computer-based vision systems for automatic credentials of plant species." *Jrnl of Agri Info*, no. 1 (2015): 61-71.
- [6] Mery, et al, "Automated design of a computer vision system for visual food quality evaluation." *Food and Bioprocess Tech* 6, no. 8 (2013): 2093-2108. *Img processing in Medical field*
- [7] Y. Yang, et al, "Micro calcification Classification Assisted by Content Based Image Retrieval for Breast Cancer Diagnosis", *IEEE Int Conf on Img Procng*, Vol. 5, pp. 1-4, 2007.
- [8] Wei L, et al., "A study on several machine-learning methods for classification of malignant and benign clustered microcalcifications", *IEEE Trans on Medical Imgng*, 24(3):371-380, 2015.
- [9] Oludayo O., et al "Segmentation of Melanoma Skin Lesion Using Perceptual Color Difference Saliency with Morphological Analysis" *Math Problms in Eng Vol* 2018.
- [10] D Kalyani et al, "An Improved Lung Cancer Prediction System using Image Processing", *Intl Jrnl of Recent Tech and Eng (IJRTE)* ISSN: 2277-3878, Vol.8, Issue-4, November 2019.
- [11] Arun Kumar Ray, et al, "Analysis for MRI Based Brain Tumor Detection and Feature Extraction Using Biologically Inspired BWT and SVM", *International Journal of Biomedical Imaging / 2017*
- [12] Onur Yorulmaz, et al, "IMAGE PROCESSING METHODS FOR FOOD INSPECTION", *Electrical and Electronics Eng*, Jan 2012
- [13] S. Padmappriya, et al., "Digital Image Processing Real Time Applications", *Intl Jrnl of Eng Sci Invention (IJESI)*, ISSN (Online): 2319 – 6734, pp. 46 -51, 2018.
- [14] Anju Asokan ,et al, "Image Processing Techniques for Analysis of Satellite Images for Historical Maps Classification—An Overview", *June 2020 ,Applied Sciences* 10(12):4207
- [15] Hongxia Deng, et al, "MFCosface: A Masked-Face Recognition Algorithm Based on Large Margin Cosine Loss", *Appl. Sci.* 2021, 11(16), Publd: 9 Aug 2021
- [16] Jeong Choi, et al, "Image processing technique to detect carbonation regions of concrete sprayed with phenolphthalein solution", 2017, *Const and Builng Materials*, Vol. 145, pp. 451-461.
- [17] Hesham M Shehata, et al, "Depth estimation of steel cracks using laser and image processing techniques", 2018, *Alexandria Eng Journl*, Vol. 57 (4), pp. 2713-2718.
- [18] Tomoyuki Yamaguchi, et al, "Image based crack detection for real concrete surfaces", 2008, *IEEE Jrnl on Trans on Elecl and Electron Eng*, Vol. 3(1), pp. 128-135.
- [19] Hardik Pandit, et al, "Appl of Digital Image Processing & Analysis in Healthcare Based on Medical Palmistry", *Proc. of Int Conf on Intel Sys & Data Processing*, pp. 56-59, 2011.
- [20] S. Padmappriya, et al, "Digital Image Processing Real Time Applications", *Intl Jrnl of Eng Sci Invnt (IJESI)*, ISSN (Online): 2319 – 6734, pp. 46 -51, 2018.
- [21] Ayatullah Faruk Mollah, et al, "Design of an Optical Character Recognition System for Camera based Handheld Devices", *IJCSI Intl Jrnl of Comp Sci Issues*, Volume: 8, July-2011 .
- [22] Judith A. Redi & et al, "Digital image forensics: a booklet for beginners", *Springer, Multi Tools Appl* (2011) 51:133–162, Oct 2010

QUESTION-ANSWERING WITH PERSONALIZED RESPONSE RESTRUCTURING

Srushti Gajbhiye
Data and AI

Advanced Technology Centers in India
(Accenture)
Chennai, India
srushti.gajbhiye@accenture.com

Anshuman Mahapatra
Data and AI

Advanced Technology Centers in India
(Accenture)
Chennai, India
anshuman.a.mahapatra@accenture.com

Selvakuberan Karuppasamy
Data and AI

Advanced Technology Centers in India
(Accenture)
Chennai, India
s.b.karuppasamy@accenture.com

Subhashini Lakshminarayanan
Data and AI

Advanced Technology Centers in India
(Accenture)
Chennai, India
s.j.lakshminarayanan@accenture.com

Abstract

With content on the rise and ever-evolving, accessibility and ease of understanding is key. Internet has solved the accessibility issue to quite an extent, making it possible for a layman to download, store and even upload data. However, understanding the downloaded information is still taxing for a layman. This deems especially true in intricate fields like finance, medicine or psychology. This paper aims to aid in understanding the document and answering questions in everyday language without using complex jargons, in minimal time. The model translates domain-specific vernacular into easy-to-understand language and also provides a summary of the answer, highlighting the important information, without losing its gist. We have tested our algorithm on multiple domains and have got promising results, enabling us to further enhance its functionalities.

Keywords—*Natural Language Generation, Machine Reading Comprehension, Response Restructuring, Summarization.*

I. INTRODUCTION

When we read any article, the aim is to understand it. When asked to explain, we try to use simpler language, elaborate the meaning of few complicated words and try to rephrase the sentences for clarity and get a narrate a summary of it. This becomes difficult if the article is on an unknown field, uses complicated words or language.

We propose an automated pipeline which helps in understanding the domain specific documents and the jargons and answer the question in an easy and efficient way so that it is helpful for a common person to understand the gist of the article without losing the meaning of it in reduced time.

This integrated pipeline first converts the unstructured input text into a meaningful content by extracting the relevant text out of it with the help of machine reading comprehension (MRC) which helps in focus on the relevant content in the whole text. This raw question and answer of the MRC is passed as input to Simple-NLG for better structured sentence. As this structured sentence consist of many domain-specific jargons in it so it is passed through response restructuring model to make the text simpler and easier to understand the specific terms which is not easily understood by common person. In order to understand the whole summary of the extracted text it is passed through summarization model. Now, the output text we have is

having all the relevant content out of the whole long text without losing any meaning out of it with proper understanding of jargons as well.

This integrated pipeline helps in understanding domain specific documents with improved clarity. It saves human effort and gives better understanding on the subject which is economical in terms of time as well.

components, incorporating the applicable criteria that follow.

II. BUSINESS USE CASES

There is a plethora of content out there which is not easily understandable to the common man. So much so, that it has given rise to alternative occupations of “agents” who have domain knowledge and can translate in everyday language a common man speaks. This becomes a hurdle for many who might not have access or resources for such agents or may not trust the agents or agents may not exist for all domains.

A. Insurance

Insurance is one such example. It is commonplace for everyone to have an insurance today – for them, family and vehicle. Choosing which one is best suitable becomes a task due to its lengthy and complicated documentation. Using an insurance agent not only costs additional money but there is always a doubt they are selling the one which gets them the highest commission. Using our solution which is custom trained on insurance domain, one can get the desired answer by simply uploading the set of documents and asking the question. This breaks down the jargons into everyday synonyms making it simpler and straight-forward.

B. Medicine

The Medicine is another such field. It is improbable for a common man to understand this extensive science. Using online search engines is not advised by doctors. For instance, MRI reports are expensive and very difficult to understand owing to the keywords used. While doctors are the best source for any consultation, our solution tries to explain what specific terms could mean in given context. This helps the patients and their families by giving them some confidence and not carry a black box report.

C. Finance and Legal

Financial documents are tough to understand and too risky to not take professional advice. Let those be tax documents, property papers, stock market or legal contracts. The market and laws surrounding it are constantly evolving, which makes it difficult to keep up. Our solution is regularly trained on the finance and legal domain to ensure giving the most recent news, when asked a question with the document.

III. LITERATURE REVIEW

Machine reading comprehension has been subject to extensive attention in the recent years in the NLP field [1-3]. MRC requires the machine to read text and then answer questions about the text, which is a very challenging task [4]. This requirement is as good as replicating human reading, understanding and answering abilities. A recurring boulder in this field of research has been the lack of high-quality MRC datasets [5], leading to its neglect for a while. There still exists a huge gap between existing MRC and real human comprehension [6], our paper attempts to bridge that gap.

The SimpleNLG package [7] can perform simple tasks to generate grammatically correct English sentences. One study uses this library to generate formal and informal sentences [8] using template-based NLG. This realization engine converts the text output from MRC into a better framed sentence, without adding any new words.

Many techniques have been used in the past for generating effective response restructuring. One of the techniques is using Transformer and seq2seq model for sentence restructure generation [9]. Another approach that has been used is generating response restructuring using reduced vocabulary which consists of building a convolution to sequence model (Conv2Seq) [10]. Similarly, Semi-Supervised Learning in NLU which uses Para-SSL is used for response restructuring generation [11]. Another work on response restructuring generation with semantic augmentation which shows the effectiveness of transformers for rephrase generation and further improvements [12]. A good response restructuring should be adequate -conveys the same meaning, should be fluent - it should be grammatically correct, and it should be diverse as well - to what extent restructuring has changed the original sentence. In this work, we focus on generating a response restructuring which is adequate, fluent and diverse, without changing the meaning of original sentence or domain specific jargons.

Recent approaches based on trained neural models address summarization as a paraphrasing task where the sequence of words in the source text is mapped to the sequence of words that make up the summary. This new paradigm uses sequence to sequence methods to bypass the need for an intermediate representation [13]. This study [14] presents a procedure capable of generating a text summary from a causal graph by introducing new words and expressions similar to the ones appearing in the nodes of the graph. In our paper, we have designed summarization in order to provide highlights of the answer, if it deems to be lengthy. This can be an on-demand step, giving user the ability to view the entire answer or simply the summary.

IV. PROPOSED APPROACH

The pipeline, as described in Fig.1, can be broken down into four major parts:

- Machine Reading Comprehension
- SimpleNLG
- Domain-specific response restructuring
- Summarization

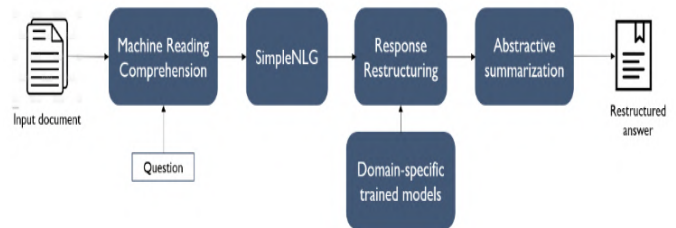


Fig. 1. Custom NLG Pipeline

A. Machine Reading Comprehension

Machine Reading Comprehension (MRC) is a sub-field in NLP which takes a text document as input and identifies the answer spans from the input text article [15], as described in Fig. 2. It is widely used in extraction of specific lines or paragraphs to quote from lengthy text pieces, for example in legal contracts or financial documents. MRC is the first component of the pipeline; where we provide a question, context and the input document and it outputs a paragraph which answers the question. This is the first basic step of QA systems where the answer is extracted from the document. This answer, however, is in a raw form, as it is directly cut from the original document. It needs processing to be in the best presentable format.

B. SimpleNLG

The SimpleNLG is intended to function as a "realization engine" for NLG architectures. Realization [16] is also a subtask of NLG, which involves creating an actual text in a human language from a syntactic representation, using linguistic constituents and features of the sentence.

We give question and raw answer from MRC as two inputs to this engine and get better framed answer which inculcates verbs, subject, object from context and question, as can be noted from Table 1. [7] This is the first step of our approach of transforming the answers into an optimum human consumable format. While SimpleNLG can capitalize the first word, add auxiliary verbs, put whitespace, punctuation and stitch all words in an appropriate grammatical format; it does not have the capability to choose or introduce new words

C. Domain-specific response restructuring

It can be thought of as a restatement or rewording of a paragraph or text, to borrow, clarify, or expand on information without plagiarizing [17].

We take the SimpleNLG output and give it as an input to our domain-specific restructuring model to generate an improved answer which ensures all jargons and keywords are incorporated. For specific domains, restructuring should be

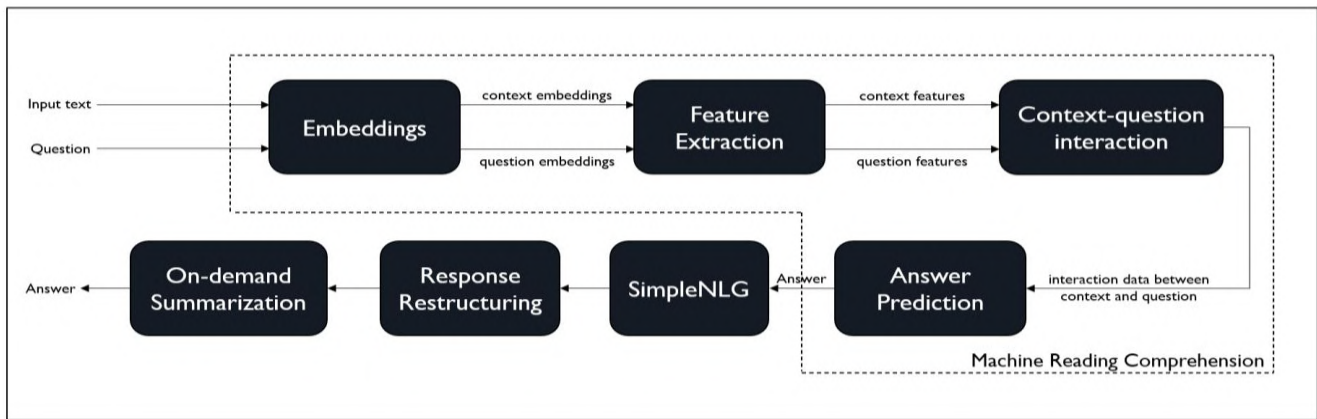


Fig. 2. Architecture Diagram

TABLE 1. FEATURES AND VALUES IN SIMPLENLG

	Feature	Values	Applicable Classes
lexical	AdjPosition	Attrib, PostNominal, Predicative	ADJ
	AdvPosition	Sentential, PostVerbal, Verbal	ADV
	AgrType	Count, Mass, Group.	N
	ComplType	AdjP, AdvP, WhFin, WhInf.	V
	VType	Aux, Main, Modal.	V
phrasal	AdjPosition	Attrib, PostNominal, Predicative	ADJ
	AdvPosition	Sentential, PostVerbal, Verbal	ADV
	AgrType	Count, Mass, Group.	N
	ComplType	AdjP, AdvP, WhFin, WhInf.	V
	VType	Aux, Main, Modal,	V

sensitive to specialized language usage. For example, ‘treat’ in general means ‘address, handle, tend’ and the same word in medical terms means ‘cure, fight, kill’. This will change the meaning of the whole text, if not used properly. Using keywords in the right way while restructuring is essential as the common public is unaware of many industry-based jargons. Domain-specific restructuring comes helps in identifying domain-specific terms and rephrase the text into simpler words without changing its original meaning so that the facts remain intact [18].

For instance, Input: Symptoms of influenza include fever and nasal congestion. Output: A stuffy nose and elevated temperature are signs you may have the flu.

D. Summarization

Summarization is the process of shortening a set of data computationally, to create a subset that represents the most important or relevant information within the original content [19]. Abstractive summarization gives important material as a summary in a new way after interpretation and examination of the text using advanced natural language techniques to generate a new shorter text that conveys the most critical information from the input text. Extractive summarization identifies important sections of the text and

generates them verbatim producing a subset of the sentences from the input text as the summary.

This final block in our pipeline can also be an optional, on-demand step based on the desired length of the final answer, using thresholding logic. We are using abstractive summarization where one more step of advanced language techniques is applied to give out a brief, crisp and precise answer.

V. EXPERIMENTAL RESULTS

NLG has progressed rapidly, but the existing evaluation methods are not up to the mark. There are only two ways to evaluate NLG systems, first is manual human evaluation – which is the most accurate since the aim of NLG is to replicate human language but also the most labor intensive and time consuming. The second method, automatic metrics are quick but often unreliable substitutes for human interpretation and judgement.

We’ve used human evaluation methodology to score our results. We tested our approach on 41 documents (with an average length of six pages) with over 200 questions spread across eight different domains – legal, medicine, finance, sales & marketing, insurance, architecture, information technology and psychology as described in Table 2. The

TABLE 2. DOMAIN-WISE DATASET DETAILS

Domain	No. of documents	Avg. no. of pages	Non-textual content		
			Tables	Images	Formulae
Architecture	3	5	✓	✓	×
Finance	7	6	✓	×	✓
Information Technology	6	5	✓	✓	×
Insurance	5	6	✓	✓	✓
Legal	5	8	×	✓	×
Medicine	5	10	✓	✓	✓
Psychology	4	7	✓	✓	×
Sales & Marketing	6	4	×	✓	✓

evaluation was carried on all stages of the pipeline – MRC, SimpleNLG, response restructuring and summarization output, to optimize the accuracy. We got an overall accuracy of 87.38% with reference to the expert-generated answer.

VI. CONCLUSION AND FUTURE WORK

We present a pipeline for Question Answering systems which presents the answer in its best form. This can be used in various domains, all leading to simplification of information for common man, translating it into knowledge.

In future, we plan to train on more niche domains (for example: laws in each country, cryptocurrency, politics) and making the pipeline multilingual. We are working on a continual training method so that the models are up to date with all advancements. We also plan to have a more detailed evaluation of our results using both human and automatic metrics and have a comparison.

REFERENCES

- Boerma, I.E., Mol, S.E., Jolles, J.: Reading Pictures for Story Comprehension Requires Mental Imagery Skills. *Frontiers in Psychology*, vol. 7. (2016) doi: 10.3389/fpsyg.2016.01630
- Zhang, X., Yang, A., Li, S., & Wang, Y.: Machine Reading Comprehension: A Literature Review. *ArXiv*, abs/1907.01686. (2019)
- Gupta, S., Rawat, B.P.: Conversational Machine Comprehension: A Literature Review. *ArXiv*, abs/2006.00671. (2020)
- Gao, J., Galley, M., Li, L.: Neural Approaches to Conversational AI. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pp. 2–7. (2018)
- Chen, D.: *Neural Reading Comprehension and Beyond*. Ph.D. Thesis, Stanford University. (2018)
- Jia, R., Liang, P.: Adversarial Examples for Evaluating Reading Comprehension Systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2021–2031, Copenhagen, Denmark. (2017).
- Gatt, A., Reiter, E.: SimpleNLG: A realisation engine for practical applications. *Proceedings of the 12th European Workshop on Natural Language Generation*, pp. 90-93. ENLG (2009) doi: 10.3115/1610195.1610208.
- Fadi Abu S., Diana I.: Generation of formal and informal sentences. In *Proceedings of the 13th European Workshop on Natural Language Generation (ENLG '11)*. pp. 187–193. Association for Computational Linguistics, USA. (2011)
- Egonmwan, E., Chali, Y.: Transformer and seq2seq model for Paraphrase Generation. *Proceedings of the 3rd Workshop on Neural Generation and Translation (WNGT 2019)*, pp. 249–255. (2019) doi: 10.18653/v1/D19-5627.
- Nomoto, T.: Generating paraphrases with Lean vocabulary, *Proceedings of the 12th International Conference on Natural Language Generation*, pp. 438–442. (2019). doi: 10.18653/v1/W19-8655
- Cho, E., Xie, H., Campbell, W.M.: Paraphrase Generation for Semi-Supervised Learning in NLU. *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pp. 45–54. (2019).
- Wang S., Gupta R., Chang N., Baldrige J.: A task in a suit and a tie: paraphrase generation with semantic augmentation. *AAAI (2019)*. doi: 10.1609/aaai.v33i01.33017176
- Casamayor G., Mille S., Aleksander Shvates A.: “V4Design”. (2019)
- Puente, C., Olivas, J. A., Garrido, E., Seiseddos, R.: Creating a natural language summary from a compressed causal graph, *Joint IFSA World Congress and NAFIPS Annual Meeting (IFSA/NAFIPS)*, pp. 513-518 (2013). doi: 10.1109/IFSA-NAFIPS.2013.6608453.
- Liu, S., Zhang, X., Zhang, S., Wang, H., Zhang, W.: *Neural Machine Reading Comprehension: Methods and Trends*. *Applied Sciences*, vol. 9 (2019). doi: 10.3390/app9183698
- Wikipedia, [https://en.wikipedia.org/wiki/Realization_\(linguistics\)](https://en.wikipedia.org/wiki/Realization_(linguistics)), last accessed 2022/01/09.
- Literary Terms, <https://literaryterms.net/>, last accessed 2022/01/09.
- Pavlick E., Ganitkevitch J., Chan, T. P., Yao X., Van Durme B., Callison-Burch C.: Domain specific Paraphrase Extraction, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Short Papers)*, pp. 57–62 (2015). doi: 10.3115/v1/P15-2010
- Wikipedia, https://en.wikipedia.org/wiki/Automatic_summarization, last accessed 2022/01/09.

MULTILINGUAL OPEN DOMAIN ENTITY EXTRACTION

Priyank Bhardwaj
Accenture, Advanced Technology
Centers in India, India
Chennai, India
priyank.b.bhardwaj@accenture.com

Venkatesan Paramasivam
Accenture, Advanced Technology
Centers in India, India
Chennai, India
v.d.paramasivam@accenture.com

Balasubramanian Vijayasankar
Accenture, Advanced Technology
Centers in India, India
Chennai, India
b.vijayasankar@accenture.com

Selvakuberan Karuppasamy
Accenture, Advanced Technology
Centers in India, India
Chennai, India
s.b.karuppasamy@accenture.com

Subhashini Lakshminarayanan
Accenture, Advanced Technology
Centers in India, India
Chennai, India
s.j.lakshminarayanan@accenture.com

Abstract—This paper talks about how open domain entity recognition can be used to extract entities irrespective of domain and language. In conventional systems, we would need a domain-specific subject matter expert (SME) to annotate, prepare training data, and extract entities from the given text or document. In this paper, we have proposed an approach that can aid open domain entity extraction without the manual efforts of the SME. Also, we added multiple language support as a feature on top of domain entity extraction. Apart from that, we can extract the entities that are present in the unstructured document – which can be in any language. The proposed approach reports approximately 85 - 90% accuracy on multiple test data and different languages.

Keywords—Named Entity Extraction, Multilingual Named Entity Extraction, Agnostic Named Entity Recognition.

I. INTRODUCTION

A. Background

With the increasing amount of unstructured data in enterprises, it is becoming an uphill task to read, understand and process them to obtain meaningful insights. For example, an enterprise needs to process its user's social media data or google search information, which is unstructured data, to understand what product they can recommend to their users. Similarly, in e-commerce, retail, or manufacturing it is a tedious task for the organization to go through the millions of reviews which is another form of unstructured data they receive for their products or services to better understand the client's like or dislike about a product and the areas of concern. At this point, open-domain entity extraction helps by mapping the keywords in the user-generated data to an entity to which it belongs.

B. Maintaining the Integrity of the Specifications

Currently, enterprises are incurring a huge amount of cost ensuring customers satisfaction. Customers raise their issues in the form of tickets, the ticket management should support multiple languages. It is an enormous manual-intensive task to read the description of all the incoming tickets, categorize and route tickets to relevant teams. If the description is in multiple languages, it becomes very difficult for a customer support engineer to identify the correct team to route the ticket towards.

II. LITERATURE SURVEY

The author J. lee has provided [1] a list of tools available in python to implement the language detection module.

There are four methods stated langdetect, spaCy, langid, and FastText. Among them, langdetect seems to work at better performance across a wide range of languages.

J. Brank et al. [2] has talked about how a Wikipedia-based Wikifier can be used to train any corpus to obtain an entity – mention graph using page rank values. The ways to compute page rank and semantic relatedness are mentioned in this paper. The model can be used across all languages.

Rahul. S et al. [3] has discussed the various processing that is required to be done on Text data like Part of speech (POS) tagging, Word sense disambiguation (WSD), Sentiment analysis, and others. They have also POS tagging and Chunking are important operations to be performed on data.

Broscheit, S et al. [4] have built a special kind of BERT that does per token classification of overall vocabulary. They have found that it works better than over plain BERT based model and other entity linking based models. They have also said to have found that most of the models do not benefit from additional knowledge as much.

Stephanie. S et al. [5] has focused on the research required on Annotated corpora to support Natural language technology in English, Arabic, Chinese and other languages. The research data is obtained from an LDC Corpora which extracts from Chinese, Arabic, and Other News channels, which is later used by ACE, DARP TIDES & DARP EELD programs for enhancing the research work on Linguistic Technology

Avirup. S et al. [6] has discussed the Multilingual Entity Linking and Entity discovery topic and how it works based on Wikipedia Knowledge Base in this journal. It has also discussed the recent development in state-of-the-art Neural EL. And the variety of applications of Cross-Lingual EL like search engines and others are discussed here.

Antoine. L et al. [7] has surveyed various text preprocessing methods for making the text data suitable for analysis. They have spoken about various domains of data and sources from which data is obtained. They have discussed about various methods like Lemmatization, Tokenization, BoW(Bag of Words), Word2Vec, Doc2Vec etc.

Laxmi. R et al. [9] has proposed a method that uses page cumulative weights of Google PageRank to rank pages for users during search operations in best to worst order. They have mentioned that context specific recommendation is

useful for giving better search result and it has given state of the art results.

Jian. N et al. [10] has found that using Wikipedia Multilingual mappings we can obtain high accuracy and high coverage on unseen entities during training. They have used this technique to develop a NER system that can be trained among 6 languages with no human annotation or Language dependent knowledge which had shown higher accuracy on unseen entities even when applied to a new domain or trained with new data.

III. PROPOSED APPROACH

Our proposed approach uses the ticket description from the ticketing system no matter what the language is and responds with the custom entities that will be immensely helpful for the customer support engineer to proceed ahead with the resolution steps.

A. Architecture Diagram

We have segregated the problem of open domain entity extraction into two modules.

1. Language Module
2. Entity Extraction Module

The detailed steps are represented in fig 1.

1) Module 1: Language Model:

1. Language detection
2. Language translation
3. NLP and Text Analytics

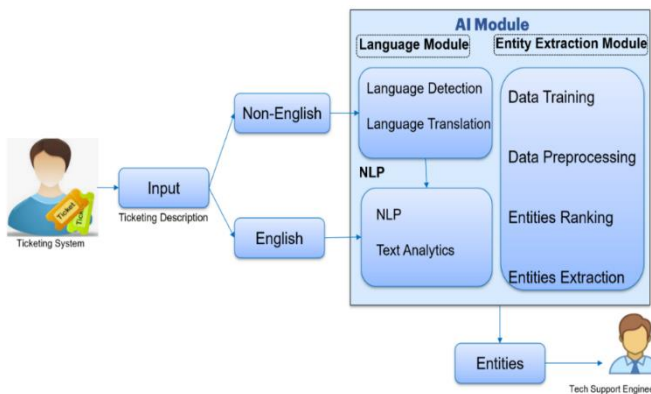


Fig. 1. Block diagram illustrating the architecture of the proposed application

a) Language Detection: The language Detection module predicts the language to which the user input or text documents belongs. Once the description is received from users, the text/document is sent to python's LangDetect [1] package which supports 55 languages like Swahili, Chinese, Hindi, English, etc., which is adapted from Google's LangDetect module built using Java. This is a probabilistic model, not a deterministic model.

b) Language Translation: Once input language is detected from the language detection module then we call the Language Translation module [2]. It will search for the respective pre-trained language translation model and

convert the input text into English. Then the text is passed into NLP and Text Analytics modules to clean the description for our open domain entity extraction model.

c) NLP & Text Analytics: This module implements NLP techniques like [3] tokenizing, part of speech (POS) tagging, lemmatization, and stemming to parse the ticket description. NLP module gets the important keywords from the description to get the context from it. It checks for the syntactic structure of the description and grammatical correction. If the customer description is in English, we must rephrase the text (for example, RE for a reply, FW for a forward) to get the context from the description.

d) Deletion: Delete the author and affiliation lines for the extra authors.

2) Module 2: Entity Extraction Module:

1. Data training
2. Preprocessing
3. Ranking Entities
4. Entities Extraction

a) Data training: To start with the entity extraction operation input descriptions are trained against the Wikipedia data [4-6], where based on the similarity between the data in input description and Wikipedia mentions they are mapped, below are the training steps:

- Model training begins by building an entity graph based on the input document.
- A directed edge between a mention "a" and entity "b" is created.
- And later the edges are clustered into a Concept "c".
- Clusters are semantically related as $c \rightarrow c'$.
- To avoid noise in data, top-200 most frequently used words are avoided in the graph.

b) Preprocessing: Since the data has a lot of special characters, stop word [7] (like and/or/by), and other unwanted data which is not going to add value to the implementation, they are removed using the preprocessing steps mentioned below.

- stop words removal
- removal of punctuations
- removal blank spaces and tabs
- removal of special characters

c) Entities Ranking: The entities are ranked based on their PageRank scores [8-10], which is defined as the relationship between an entity and an input word it will have a high PageRank score e.g., Himalayas and Western ghats can be strongly related to entity Mountains. Entities have been chosen by the highest page rank score. The PageRank is calculated as follows, PageRank is calculated for all the vertices in a text-entity graph using an iterative approach where with each iteration, the PageRank score of one vertex is to its immediately adjacent vertex and so on. Later this PageRank value is used to choose the most appropriate entity for a particular ticket description.

d) *Entities Extraction*: As of now we already have the entities ranked so as the user gives input to the system, based on the keywords identified by the system from the input description the related entity name is displayed and highlighted along with it.

IV. RESULTS AND DISCUSSION

We have tested the model with Wikipedia data as input and the results are shared as below. In stage 1, we have given an English language input from Wikipedia data and the results are posted (see **Error! Reference source not found.**)

Below table 1 shows, the comparison between the ground truth entity value for each word and the model predicted entity value. We have calculated the precision, recall, and f1 score for the results and furnished them in table 2 as below.

The proposed solution gives around 80 – 90% of accuracy for any domain, language detection accuracy of 90 – 95% and it supports.

Below table 3 shows, the comparison between the ground truth entity value for each word and the model predicted entity value. We have calculated the precision, recall, and f1 score for the results and furnished them in table 2 as below.

Elon Musk **human** is a business magnate, industrial designer, and engineer. Elon Musk **human** is the founder, CEO, CTO, and chief designer of SpaceX **aerospace manufacturer**. Elon Musk **human** is also early investor, CEO, and product architect of Tesla, Inc. **automobile manufacturer** Elon Musk **human** is also the founder of The Boring Company **business** and the co-founder of Neuralink **business**. A centibillionaire, Musk became the richest person in the world in January 2021, with an estimated net worth of \$185 billion at the time, surpassing Jeff Bezos **human**. Musk was born to a Canadian mother and South Africa **sovereign state** n father and raised in Pretoria **capital**, South Africa **sovereign state**. Elon Musk **human** briefly attended the University of Pretoria **capital** before moving to Canada aged 17 to attend Queen's University. Elon Musk **human** transferred to the University of Pennsylvania **private university** two years later, where Elon Musk **human** received dual bachelor's degrees in economics and physics. Elon Musk **human** moved to California in 1995 to attend Stanford University **private university**, but decided instead to pursue a business career. Elon Musk **human** went on co-founding a web software company Zip2 **company** with Elon Musk **human** brother Kimbal Musk **human**

Fig. 2. Entity extraction results – English

Table 1. Comparison of actual (ground truth) entity and model predicted entities – English

Input Words	True Label (ground truth)	Predicted Label
Elon Musk	Human	Human
Jeff Bezos	Human	Human
Stanford University	Private University	Private University
University of Pennsylvania	Private University	Private University
Canada	Country	No Prediction
California	State	No Prediction
South Africa	Sovereign state	Sovereign state
Pretoria	Capital	Capital
Neuralink	Business	Business
The Boring Company	Business	Business
SpaceX	Aerospace Manufacturer	Aerospace Manufacturer
Queen's University	Private University	No Prediction
Tesla Inc	Auto Manufacturer	Auto Manufacturer
Kimball Musk	Human	Human

Table 2. Model results for English

Entity	Precision	Recall	F1-score
Human	1.000	1.000	1.000
Location	1.000	0.667	0.800
Education	1.000	0.667	0.800
Business	1.000	0.750	0.850
Avg/total	1.000	0.7725	0.8625

V. CONCLUSION AND FUTURE WORK

Our proposed approach supports entity extraction irrespective of the domain to which the data belongs to e.g., e-commerce, finance, pharma data, etc. apart from this it also provides the multiple language support feature. The multiple languages support, most of the standard languages such as Arabic, French, Mandarin, Hindi, etc. The main benefit of this proposed approach is the elimination of the need for training and annotating data required for entity extraction irrespective of domains. For future work, we have planned to extend our proposed approach to work on scanned PDF documents, and video documents.

REFERENCES

1. Jenny Lee. Benchmarking Language Detection for NLP. (Nov 2020). <https://towardsdatascience.com/benchmarking-language-detection-for-nlp-8250ea8b67c>
2. Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar & Zdeněk Žabokrtský: Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. (Sep 2020). <https://www.nature.com/articles/s41467-020-18073-9>
3. Khyani, Divya & B S, Siddhartha. An Interpretation of Lemmatization and Stemming in Natural Language Processing. (2021). Shanghai Ligong Daxue Xuebao/Journal of University of Shanghai for Science and Technology. 22. 350-357.
4. Broscheit, Samuel. (2019). Investigating Entity Knowledge in BERT with Simple Neural End-To-End Entity Linking. 677-685. 10.18653/v1/K19-1063.
5. Panagiotis Ilyperopoulos, Haoling Qiu, and *Bonan Min*: Concept Wikification for COVID-19. (Oct 2020) <https://openreview.net/forum?id=fJCcajDO8Tx>
6. Avirup Sil, Heng Ji, Dan Roth, and Silviu Cucerzan: Multilingual Entity Discovery and Linking. (Jan 2018). <https://aclanthology.org/P18-5008.pdf>
7. Antoine Ly, Benno Uthayasooryar, and Tingting Wang: A SURVEY ON NATURAL LANGUAGE PROCESSING (NLP) & APPLICATIONS IN INSURANCE. (Oct 2020). <https://arxiv.org/pdf/2010.00462.pdf>
8. Brank, Gregor Leban, Marko Grobelnik. ANNOTATING DOCUMENTS WITH RELEVANT WIKIPEDIA CONCEPTS. (Jan 2021). https://ailab.ijs.si/dunja/SiKDD2017/Papers/Brank_Wikifier.pdf
9. Laxmi Rajani and Urjita Thakar, Webpage Recommendation for Organization Users via Collaborative Page Weight. (Jun 2021). https://www.bhu.ac.in/research_pub/jsr/Volumes/JSR_65_01_2021/29.pdf
10. Sujata Joshi, Shivkumar Goel: Comparative Study of Page Rank and Weighted Page Rank Algorithm. (Sep 2021). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3919359

MEDICAL IMAGE SEGMENTATION AND CLASSIFICATION USING NEURAL NETWORKS

Dr.D.Suganthi,

Assistant Professor,

*Department of Master of Computer Science and Applications,
SRM University, Ramapuram Campus, Chennai
suganthd@srmist.edu.in*

Mr.C.Jeyaganthan,

Senior Software Engineer,

*Navis India Technologies,
Taramani, Chennai
jeyaganthan@gmail.com*

Abstract - Image segmentation plays an important role in the field of image analysis and computer vision which is also regarded as the bottleneck of the development of medical image processing technology applications. Medical Resonance Image (MRI) plays an major role in medical diagnostics and different acquisition modalities were used. A major goal of fMRI data analysis is to recognize activated brain areas and one of the major steps has segmentation. ANN is a computational simulation of a biological neural network, has been classified into many networks. Recurrent neural networks specifically in Echo State Neural Network (ESNN) have implemented fMRI segmentation. The performance of ESNN with CC give 97% accuracy. MATLAB R2011a software was used. The texture features of each class gives a high-efficiency rate. The quantification of the result demonstrates the effectiveness of the proposed method.

Keywords: *Brain tumor, Segmentation, Echo State Neural Network (ESNN), Contextual Clustering (CC), MATLAB.*

I. INTRODUCTION

Medical Resonance Image (MRI) plays an important role in the field of medical diagnostics. In medical imaging for analysing anatomical structures such as bones, muscles, blood vessels, tissue types, pathological regions such as cancer, multiple sclerosis lesions and for dividing an entire image into sub-regions such as the white matter (WM), gray matter (GM) and cerebrospinal fluid (CSF) spaces of the brain automated delineation of different image components are used. MRI can provide plentiful information about human soft tissue anatomy as well as helps diagnose brain tumor.

MR images are used to analyze and study the behavior of the brain. Diseases are shown using MRI signal between neural activity and the local blood flow that results in the BOLD signal. But are looking at how blood oxygen levels change and assuming that this is connected to nerves. Functional magnetic resonance imaging (fMRI) has become

an important method for the investigation of human brain function, both for research and for clinical purposes. Functional areas identified by motor, sensory, and language tasks have been shown to correspond well with intra-operative mapping and also with classically defined anatomical regions responsible for these functions.

Segmentation of an object in an image is performed either by locating all pixels or voxels that form its boundary or by identifying them that belongs to the object. In medical imaging, segmentation is an important analysis function for which lots of algorithms and methods have been built up. Segmentation techniques provide flexibility. The developing platform for the detection is MATLAB. Introduce to acquire high-resolution brain images with ultrahigh field (7T) MR scanner and identify voxels responding to the task using our approach.

II. PREPROCESSING

First, data are analyzed to find the regions with MR signal changes temporally correlate with the experiment paradigm. Second, a threshold is used to discriminate the "inactive" brain regions (i.e., those with signal changes that are more consistent with noise) from the "active" regions. The following pre-processing steps done before segmentation process.

Realignment

Movement-related variance induced by gross head motion in fMRI time-series represents one of the most serious confounds of analysis. Before analysis, head motion detection should be made to evaluate the quality of data. The adjustment may be furthered by correction based on an estimate from a moving average auto-regression model of spin-excitation history effects.

Spatial Normalization

To implement a voxel-based analysis of imaging data, data from different subjects must derive from

homologous parts of the brain. Spatial transformations are therefore applied that move and "wrap" the images that they all conform approximately to some idealized standard brain.

Spatial Smoothing

There are several advantages of spatial smoothing. First, it generally increases SNR. The neuropsychological effects of interest are produced by homodynamic changes that are expressed over spatial scales of several millimetres, whereas noise usually has higher spatial frequencies. In fMRI the noise can be regarded as independent for each voxel and has therefore very high spatial frequency components. Second, it enhances statistical inference.

Image Enhancement and Background Cancellation

To remove the random noise and to maintain the boundary information while producing no additional artifact, the images can be filtered with an anisotropic filter. There are many black background pixels around MRI of the brain, which can be removed before the computation since they are not meaningful for signal calculation or classification. The threshold can be computed experimentally by one-tenth of the maximum pixel value of a MR image.

Segmentation

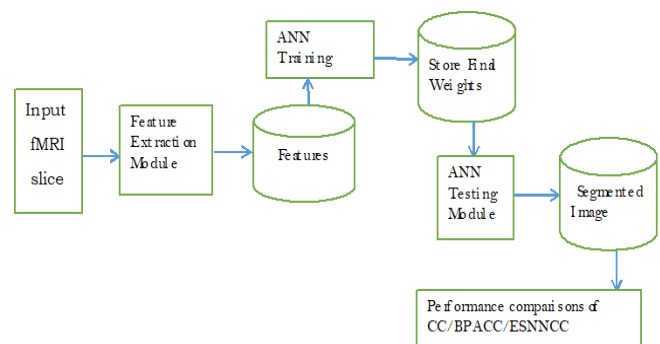
Image segmentation is very useful in separating grey matter, white matter, cerebrospinal fluid, blood vessels, and other brain structures. Segmentation methods usually utilize the differences in intensity distribution of different tissues. An unsupervised approach combines Kohonen self-organizing feature map and fuzzy c-means can classify brain into six different tissues using T1-weighted, T2-weighted and proton density weighted images. In fMRI study, segmenting these structures helps differentiating functional responses in gray matter from large vessels. Thus provides better spatial localization and quantification accuracy.

Statistical inference about specific regional changes requires statistical parametric mapping. Although statistical methods are capable of identifying the functional responses from the MR images, they depend on some prior knowledge or assumption of the physiological response in brain activation. The development of data-driven post-processing methods capable of identifying unknown response pattern

becomes crucial. This approach provides reliable analysis of the known functional responses. Methods belonging to this category include: correlation analysis, t-test, general linear model. The functional activities are performed, detection and identification of tumor in the brain. Pre-processing and Segmentation is an important role in order to distinguish between normal patients and their abnormalities or tumor patients. The proposed method consists of three stages: Feature Extraction module, 2. ANN Training module and 3. ANN Testing module.

III. PROPOSED SYSTEM ARCHITECTURE

Segmentation is an important process that helps to identifying objects in the given image. Existing segmentation methods are not able to correctly segment the complicated profile of the fMRI accurately. Segmentation of every pixel in the fMRI correctly helps in proper location of tumor. The presence of noise and artifacts poses a challenging problem in proper segmentation. This research work proposes a new intelligent segmentation technique for functional Magnetic Resonance Imaging (fMRI). In this segmentation process, the fMRI image can be segmented with contextual clustering method and artificial neural networks.



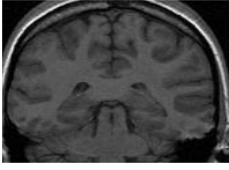
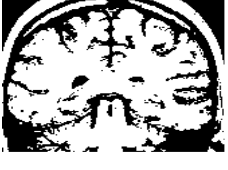
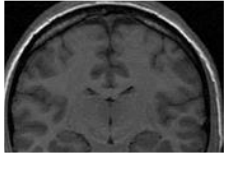
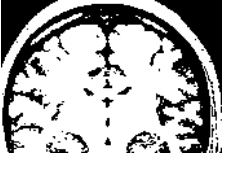
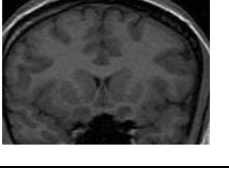

IV. ANN TRAINING MODULE

The ANN training module trains the supervised algorithms namely back propagation algorithm (BPA) and echo state neural network (ESNN) to learn the segmentation of fMRI. The initial weights used for assigning the connection strengths between input-hidden layers, hidden-hidden layer, hidden-output layers.

At the end of the training process in both BPA and ESNN, a set of matrices called final weights is stored. The ANN algorithms used are BPA and ESNN. These two

algorithms undergo two phases before segmenting the fMRI images. Both phases gives mean of the 9 intensity values, summation of 9 intensity values and the V_{cc} obtained from CC algorithm for a moving overlapping window are given as input to the input layer of the ANN topology.

V. SEGMENTATION RESULTS

Segmentation Results for Different images using ESNN		
Image No	Original image	Segmented result
Image 25		
Image 30		
Image 40		

Segmentation Accuracy Evaluation based on Correct Number of Pixels Segmented in the fMRI Slice

The segmentation accuracy based on the correct number of pixels segmented (A_p) is obtained using

$$\text{Equation } A_{\text{pixel}} (\%) = \frac{N_c}{T_p} * 100 \text{ where,}$$

A_{pixel} = segmentation accuracy

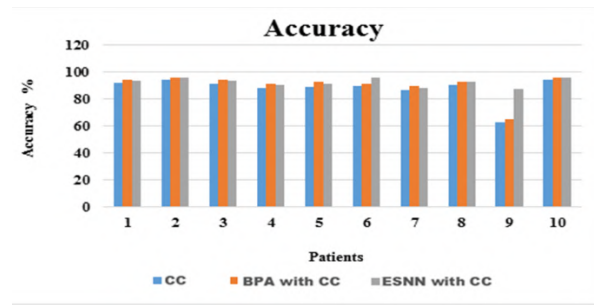
N_c = number of correctly segmented pixels in the fMRI

T_p = total number of pixels corresponding to the fMRI in the unsegmented image (Ground Truth).

This method of evaluating segmentation accuracy helps as follows:

1. The various objects present in the segmented fMRI can be chosen as a whole and the accuracy of segmentation for each object inside the fMRI can be evaluated.
2. As the intensity values are likely to overlap in the adjacent regions of the objects inside the fMRI, the region properties will show how accurately the objects are segmented.

VI. PERFORMANCE IN TERMS OF ACCURACY



The performance analysis of the implemented segmentation method is carried out as follows. For this process, fMRI slices are considered. All the slices are segmented by CC, BPA with CC, and ESNN with CC methods.

VII. CONCLUSION AND FUTURE ENHANCEMENT

The performance analysis of the implemented segmentation method is carried out as follows. For this process, fMRI slices are considered. All the slices are segmented by CC, BPA with CC, and ESNN with CC methods. This segmentation is applied for all the fMRI images. In this work, an intelligent neural network representation model for segmenting image called back propagation algorithm with CC features (BPA with CC) is proposed for effective segmentation of fMRI slices by effectively providing facility for developing a segmentation support system. These features are given as the input layer for proper learning by BPA network and it shows the good experimental results. The experimental results obtained from the proposed BPA with CC segmentation algorithm helps in reducing the false positive rate.

FUTURE ENHANCEMENT

A new combination of features can be generated by considering three to four successive slices in order to segment the volumetric information of fMRI. The number of features can be increased to represent the different properties like density, change in greyscale and change in contrast in the successive slices. For the purpose of enhanced segmentation the Gray Level Co-occurrence Matrix (GLCM) properties can be used combined with genetic algorithm for decreasing the false positive rate.

REFERENCES

- [1]. Afshin S., and Fatemeh J., 2012, Automated technique for medical images using neural network, International journal of Multidisciplinary sciences and Engineering, Vol.3, No.3, pp.38-41.
- [2]. Anamika Ahirwar, 2013, Study of techniques used for medical image segmentation and computation of statistical test for region classification of brain MRI, International Journal of Information Technology and Computer Science, Vol.5, No.5, pp.44-53.
- [3]. Beaulieu J.M. and M. Goldberg, 1989, Hierarchy in picture segmentation: a stepwise optimization approach, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.11, No.2, pp.150-163.
- [4]. Bezdek J.C., Hall L.O., and Clarke L.P., 1993, Review of MR image segmentation techniques using Pattern Recognition, Medical Physics, Vol.20, No.4, pp.1033–1048.
- [5]. Carlos A.P., Khan I., and Robert Kozma, 2003, Automated brain data segmentation and pattern recognition using ANN, International Conference on Computational Intelligence, Robotics and Autonomous systems, pp.27-31.
- [6]. Dzung L.P., Chenyang Xu and Jerry L.P., 2000, Current methods in medical image segmentation, Annual Review of Biomedical Engineering, Vol.2, pp.315-337.
- [7]. Fesharaki M.N, and Hellestrand G.R., 1994, A new edge detection algorithm based on a statistical approach, International symposium on Speech, Image Processing and Neural Networks, Vol.1, pp.21-24.

FUZZY K-MEANS CLUSTERING TECHNIQUE FOR ENERGY EFFICIENT ROUTING IN WIRELESS SENSOR NETWORKS

Dr.S.Lavanya¹, Dr.P.Calista Bebe² Dr.C.Sudha³

¹Head & Associate Professor, Department of Software Applications,
Thiruthangal Nadar College, India

² Associate Professor, Department of Computer Applications,
New Prince Shri Bhavani Arts & Science College, India

³ Assistant Professor, Department of Computer Science,
Sri Sankara Arts & Science College (Autonomous), India

¹lavanya.tnc20@gmail.com ²calista@newprincearts.edu.in ³srisudhasri.kpm@gmail.com

Abstract

Optimization of energy could be a major thought in wireless device network once planning and designing the WSN routing protocols. Cluster utilized in routing has been verified its potency to avoid wasting energy in device network. Formation of applicable clusters is extremely vital in planning of cluster-based routing protocols of WSN. This paper focuses on proposing associate rule economical cluster formation approach by discrimination fuzzy K-means (FKM) cluster rule that improves the energy conservation by making extremely uniform clusters and scale back the overall communication distances. Thus, the projected cluster approach will extend the network period of time. Novelty: The proposed routing protocol is enforced in k-means simulation surroundings and compared to C-Means. Simulation results reveal that the projected approach defeats C-Means protocol in terms of saving energy and prolonging network period of time.

Keywords: *Wireless sensor networks, clustering, fuzzy logic, Euclidian mean value, Residual energy.*

1. Introduction

The improvement of wireless sensor community changed into initiated in army programs which include battlefield surveillance and goal tracking; these days WSNs are used in lots of civilian and business software areas, which includes device fitness tracking, business system tracking and control, healthcare programs, visitors control, domestic automation, climate forecasting, surroundings and habitat tracking. [7][8]Specially, with the growing of the Internet of Things (IOT), WSN received a sustainable improvement. Routing is the system of forwarding the records from supply to destination. It is accomplished with

the aid of using the community layer. Routing protocols can't be constant for the reason that layout necessities for a community have modified with the software [9][10].

Because the sensor nodes of network can be deployed in inaccessible and threatening environments, recharging or changing in their personal strength assets is neither viable nor economical. Therefore, improving strength intake to increase the community lifestyles time is a vital hassle in WSN [11][12]. Clustering has been tested as a powerful routing approach to reduce the strength intake of sensor nodes, stability strength intake among the nodes and lengthen the life of community. In clustering approach, sensor nodes are dividing into corporations named clusters [13][14][15].

2. Literature work

In this work, we present the fluffy rationale method to the appointment of group heads and join with k-Means bunching approach to accomplish the ideal energy productivity in the sensor organization. In this paper, we propose and analyze a Distributed k-Means grouping convention which is called Fuzzy Logic based k-Means Routing Protocol for WSN.

MUHAMMAD RIZWAN ET AL.(2019). [1] ENERGY PROTECTION IS ONE OF THE MAIN EXPLORATION CHALLENGE. FLUFFY LOGIC BASED MULTI-JUMP ENERGY EFFICIENT ROUTING PROTOCOL (FMEEP) FOR HETEROGENEOUS WSN, WHICH UTILIZES FLUFFY RATIONALE SURMISING FRAMEWORK (FIS), NOVEL ROUTING APPROACH FOR ALTOGETHER INCREMENT THE SECURITY TIME FRAME, NETWORK LIFETIME AND THROUGHPUT OF THE SENSOR ORGANIZATION. THE MULTI-JUMP APPROACH IS UTILIZED FOR GROUP HEADS THAT SURPASS THE CORRESPONDENCE EDGE, $D > D_0$, TO BASE STATION. IN THIS MULTIHOP SYSTEM, EACH GROUP HEAD

UTILIZES INSATIABLE FORWARD APPROACH AND FORWARD PARCELS TOWARDS THE BS.

J.S.Pan et al.(2019) [2] In this paper, Best Route determinations were made dependent on interface weight cost (crisp).In prior work Adaptive Energy Saving and Reliable Routing Protocol(AESRRP) joined connection still up in the air dependent on the boundaries hub's leftover energy and transmission boundaries. In this paper fluffy rationale approach is utilized to compute connect cost between two neighboring hubs. K – Mean bunching of m hubs and consequently powerful group head choice by populace age of GA improves the organization lifetime. another technique encodes a mysterious pixel into $m + n(n-1)$.

Sasikumar Periyasamy et.al (2019) [3] An adjusted k-implies (Mk-implies) calculation for grouping was proposed which incorporates three bunch heads (at the same time picked) for each bunch. These group heads (CHs) utilize a heap sharing component to pivot as the dynamic bunch head, which monitors lingering energy of the hubs, in this way expanding network lifetime.

Gaurang Raval et.al (2020) [4] This paper presents the examination of different Wireless Sensor Network (WSN) bunching conventions like LEACH-Centralized, KMeans based grouping, Fuzzy C-Means grouping and Harmony Search Algorithm based bunching. The conventions have been contrasted with deference with network lifetime, energy utilization and adequacy of bunching. HSA based strategy showed prevalent execution when contrasted and different conventions.

Anand Gachhadar et al. (2019) [5] An energy proficient novel grouping plan is planned to give low energy utilization, decreasing over-burden on sensor hubs and increment network lifetime of remote sensor organization. Brought together bunching engineering and KEAC calculation is proposed to give energy proficient and draw out network lifetime in remote sensor organizations.

Pramod Kumar et al. (2019) [6] In this work, to resolve these issues employments of k-implies and fluffy C-implies calculations are explored for bunches arrangement and resulting choice of group heads (CHs). For this load of recently framed groups; choice of bunch head is done dependent on part sensor hubs leftover energy status (RES) trailed by assessment of Euclidean distances.

3. Methodology

This study workouts following problems:-

- K-means clustering based optimum Euclidian distance measures the distance between the nodes to initialize the centroids and also centroids to data points.
- Clusters formation and finding new centroids are done based on Distributed k-means clustering technique.

3.1 Euclidean distance-based cluster classification

K-Means is the least difficult calculation utilized for un-supervised Clustering. This algorithm subsets the data set into k clustering utilizing the Euclidian distance mean, resulting in expanding intra-cluster similitude and decreasing between bunch closeness. K-Means is iterative in nature.

The distance between the information focuses is determined utilizing Euclidean distance characterized by,

$$\text{Dist}(x_1, x_2) = \text{Sqrt of } \sum_{i=1}^n (x_{1i} - x_{2i})^2 \dots\dots\dots 1$$

Utilizing this condition, network climate is isolated into two virtual layers, in regards to separate from base station. It is isolated utilizing Equation (1) least distance among sensors and base hub, greatest distance among sensors and base hub, and mean value is estimated. In this manner network climate is isolated into 2 virtual layers.

```

Administrator: Command Prompt
Microsoft Windows [Version 6.1.7601]
Copyright (c) 2009 Microsoft Corporation. All rights reserved.

C:\Users\Administrator>cd C:\
C:\>cd distance
C:\distance>set path=C:\Program Files (x86)\Java\jdk1.7.0_25\bin
C:\distance>javac Eculidian.java
C:\distance>java Eculidian
enter x1 point
-1
enter y1 point
5
enter x2point
1
enter y2 point
6
distancebetween<-1,5>,<1,6>====>2.23606797749979
C:\distance>

```

Figure.1 Example Output: Euclidian Distance Between two data points

3.2 Distributed k-Means Clustering

In this sort of clustering, each hub achieves all the fundamental data for grouping from any remaining hubs. Since the k-implies calculation depends on the essential guideline of Euclidian distances and leftover energies of the sensor hubs (for picking CH), the data of the area of hubs and their comparing lingering energies is acquired by each hub by exchanging messages among themselves. In the wake of social affair the data pretty much every one of the hubs every hub runs the calculation (k-implies).

The k-implies calculation for grouping and the calculation for picking CH are particularly comparable as the calculations utilized in unified bunching. As each hub runs a similar calculation, each hub realizes its parent bunch and it's CH. So here there is no course of Declaration of Cluster Head as in brought together. Accordingly, the appropriated grouping measure is finished. At first k group places are subjectively picked and every one of the hubs is distributed a highlight the closest focus. Then, at that point, the bunches are refreshed by decision the mean of the different part designs, and similar advances are rehashed until the calculation combines. Given a bunch of perceptions ($x_1, x_2 \dots x_n$), k-implies grouping segments the set into k groups to such an extent that the inside group amount of squares (WCSS) is limited.

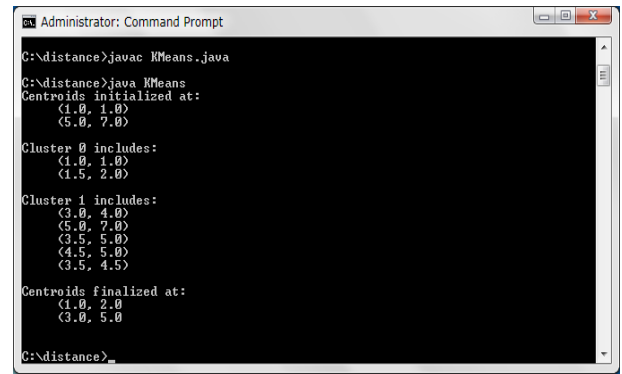
Stage 1: Deliberate the information sources:

- (I) k -number of groups, set of information focuses (hub areas) are $\{x_1, \dots, x_n\}$.
- (ii) Connotation of x_i will be done uniquely to one group.
- (iii) i indicates i th emphasis in the grouping system.

Stage 2: Domicile k centroids in arbitrary spots $\{C_1, \dots, C_k\}$: $k=1, \dots, .K$.

Stage 3: Repeat until assembly: For each point, find the closest centroid and apportion the point x_i to centroid C_j .

Stage 4: Break when the emphasess merge: No hub changes its group in successive cycles.



```

Administrator: Command Prompt
C:\distance>javac KMeans.java
C:\distance>java KMeans
Centroids initialized at:
(1.0, 1.0)
(5.0, 7.0)

Cluster 0 includes:
(1.0, 1.0)
(1.5, 2.0)

Cluster 1 includes:
(3.0, 4.0)
(5.0, 7.0)
(3.5, 5.0)
(4.5, 5.0)
(3.5, 4.5)

Centroids finalized at:
(1.0, 2.0)
(3.0, 5.0)
C:\distance>

```

Fig 2. Sample Output: Finding Centroids

4. Implementation of fuzzy k-means clustering

Fuzzy K-Means (likewise called Fuzzy C-Means) is a delay of K-Means, the predominant straightforward bunching procedure. Fuzzy K-Means is another measurably benevolent strategy and fortunes delicate groups where a positive point can fit to more than one bunch with unmistakable likelihood.

Fuzzy System

Fuzzy k-implies bunching convention customs fluffy in the event that standards to increase the life expectancy of remote sensor organizations. The hypothesis of fluffy rationale was advanced by Dr. Lotfi Zadeh in 1965. Fluffy rationale arrangement is involved of a fuzzier, fluffy surmising machine, defuzzifier and a standard base [1]. All the fluffy rationale based energy ingesting techniques are same as in Fuzzy rationale set up Multi-jump energy productive convention for heterogeneous remote sensor network [1].

Framework model and suspicions

This part gives the framework model of the proposed directing plan. Leave n alone the quantity of sensor hubs arbitrarily coordinated in 100 x100 square meter detecting district. There are two kinds of sensor hubs utilized for example ordinary and advance hubs. Advance hubs are outfitted with extra introductory energy than ordinary hubs. Let E_0 is the underlying energy of the ordinary hubs and $E_0 \times (1 + \alpha)$ be the underlying energy of m division of the development hubs, where α capitals that advance hubs contain α periods additional underlying energy than typical

hubs $(1-m)n$ [1][15]. All out Initial energy of WSN model is given by:

$$E_{total} = N.E_0(1-m) + N.m(1+\alpha)E_0$$

$$E_{total} = N(1+\alpha m)E_0 \dots\dots\dots(2)$$

The radio energy exorbitance model is displayed in the fig 1. Free space and multipath winding down direct are utilized in this energy model. On the off chance that the distance among transmitter and got is less than a limit esteem then unhindered space model is utilized in any case multipath misfortune model is utilized. The measure of energy required to convey L pieces message over a distance d starting with one hub then onto the next hub is accepted by:

$$E_{Tx}(L,d) = \{ L \times E_{elec} + L \times E_{fs} \times d^2 \text{ if } d < d_0, L \times E_{elec} + L \times E_{fs} \times d^4 \text{ if } d > d_0 \} \dots\dots(3)$$

E_{elec} is the unreasonable energy to run the transmitter or beneficiary. The imperatives E_{fs} and E_{mp} is the measure of energy guilty pleasure which relies on the distance d_0 .

In this paper, our sensor network comprises of n sensor hubs which are discretionarily situated in the detecting locale to unremittingly screen the climate. Presently we depict some suspicion for our proposed directing convention:

- ❖ The sensor network comprises of heterogeneous sensor hubs.
- ❖ Sensor hubs are sent arbitrarily in the detecting field.
- ❖ The base station is fixed and situated at the focal point of detecting field.

5. Results and Discussion

For evaluating the performance of the fuzzy k-means, there are some performances metrics area unit accustomed compare and analyze the leads to the Wireless sensing element Network. These metrics are: Energy Consumption and Network life. Energy consumption is that the total quantity of energy consumed by all the sensing element nodes that kind the networks and it represents the distinction average between the initial energy state and therefore the current level of remaining energy every node in each spherical. Network life is measured because the time period taken from the beginning of the network configuration until the death of the last sensing element

node in network. It's painted as variety of rounds created by the nodes.

5.1 Performance measures of Proposed Fuzzy K-Means Algorithm

Like K-Means, Fuzzy K-Means chips away at those substances which can be meant in n -dimensional vector space and a distance apportion is divided. The calculation is like k -implies.

1. Initialize k groups
2. Until united
 - ❖ Calculate the likelihood of a direct fit toward a group for each <point, cluster> pair
 - ❖ Recomputed the group communities utilizing straight above likelihood participation ethics of focuses to bunches

Fuzzy k -implies expressly cuts to manage the issue where focuses are somewhat in the middle of focuses or in any case ill-defined by exchanging distance with likelihood, which obviously could be some capacity of distance, for example, having likelihood virtual to the opposite of the distance. One should get a handle on that k -implies is remarkable instance of fluffy k -implies when the likelihood work utilized is basically 1 if the information point is nearest to a centroid and 0 in any case.

Calculation: Fuzzy K-Means (FKM)

Info: position of hubs

Yield: centroids of group

Start
introduce U_f
rehash

For group j to $1 = C$ do
 $C_j \leftarrow$ register bunch centroid

End for
update U_f
until the calculation meets
return $\{C\}$

End

Model Program for Fuzzy K-Means Clustering

Info : $D = \{d_1, d_2, \dots, d_n\}$ / set of n components d_i , $C = \{c_1, c_2, \dots, c_k\}$ / set of k centroids / $k=3$.

Steps:

1. Compute the distance of every component d_i ($1 \leq i \leq n$) to every one of the centroids c_j ($1 \leq j \leq k$) as $d(d_i, c_j)$;
2. For every component d_i , track down the adjoining centroid c_j and relegate d_i to group j .
3. Set $\text{ClusterId}[i]=j$; j : Id of the close by bunch
4. Set $\text{Nearest_Dist}[i] = d(d_i, c_j)$;
5. For each bunch j ($1 \leq j \leq k$) centroids:
6. Rehash
7. For individual component d_i .
8. Figure its distance is less than or indistinguishable from the contemporary closest group
9. On the off chance that this distance is not exactly ($<$) or equivalent ($=$) to the current closest distance, the component visits in the group: Else
10. For every centroid c_j ($1 \leq j \leq k$) Calculate the distance $d(d_i, c_j)$; End for;
11. Dispense the component d_i to the bunch with the closest centroid c_j
12. Set $\text{Cluster ID}[i]$;
13. Set $\text{Nearest_Dist}[i] = d(d_i, c_j)$;
14. End for;

For each bunch j ($1 \leq j \leq k$), recalculate the centroids

Until the combination models is experienced.

5.2 Simulation and Analysis Results:

The proposed algorithm is obtainable and compared C means Leach protocol in WSN. System simulation is performing under Fuzzy k-means cluster with the parameter listed in table 1.

Table1: Accurate Parameters in Simulation

1	Limitations	Value
2	Network Size	100 M ²
3	Number of hubs	100
4	Base Station Location (BSL)	50,200

5	Usage of Cluster	5
6	Packet Size	4000 bit

200 static sensor hubs are arbitrarily positioned to cover the sensing area. Each hub sends packet size 4000 bit to BSL through cluster during each round.

Table2: Accurate Parameters in Simulation

Run	K-Means routing Protocol	C-Means routing Protocol	Difference
1	1200	1100	100
2	1255	1105	150
3	1212	1110	102
4	1260	1135	125
5	1266	1150	116
6	1280	1141	139
7	1262	1145	117
8	1260	1155	105
Average	1249	1130	119

The use of Fuzzy k-means clustering (Means) to decrease transmission distance since saving energy and increase network lifetime speed. Applied this algorithm have better formation where the mean distance for every hub to cluster is decreased. It is more effective for equalizing network load and distributing nodes between clusters.


```

C:\distance>java -k f_k_means.java
C:\distance>java f_k_means
Enter the number of clusters
10
Enter 10 elements:
2 1 12 14 16 19 20
Enter the number of clusters:
3
At this step
Value of clusters
K1C 1 12 14 16 19 20 >
K2C 2 1 >
K3C 3 10 >
Value of m
m1=2.6 m2=3.0 m3=13.125
At this step
Value of clusters
K1C 1 12 14 16 19 20 >
K2C 2 1 >
K3C 3 10 >
Value of m
m1=2.6 m2=3.5 m3=14.428571428571429
At this step
Value of clusters
K1C 1 12 14 16 19 20 >
K2C 2 1 >
K3C 3 10 >
Value of m
m1=2.6 m2=3.5 m3=14.428571428571429
The Final Clusters By Kmeans are as follows:
K1C 1 12 14 16 19 20 >
K2C 2 1 >
K3C 3 10 >
C:\distance>

```

Fig 3. Output: Fuzzy k-means clustering iterations

The performance results of k-means and FKM calculations are analyzed as far as energy utilization per question and are displayed in Fig 4. The structured presentation in this figure shows that the energy needed to deal with a solitary inquiry by utilizing k-implies calculation is **0.7229 mJ**, while, for a solitary question it's add up to **0.3614 mJ** utilizing FKM calculation. In this manner, to deal with a solitary inquiry on normal premise the FKM calculation requires around half of the energy than that of the k-implies calculation. Along these lines, for guaranteed and restricted energy save, FKM calculation is more effective as it is equipped for handling around twofold measure of questions. In total terms, this scaled inquiry taking care of limit can be aligned as far as spending time in jail and indisputably we can say that the utilization of FKM calculation broadens the organization life time around twofold than that got on utilizing k-mean calculation.

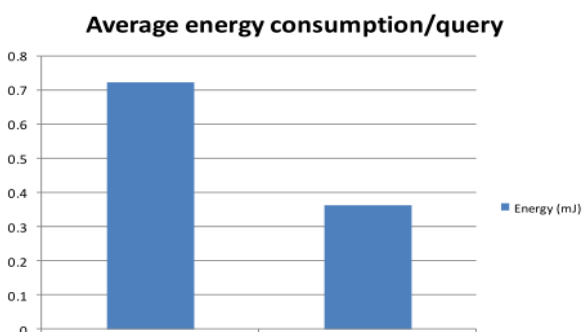


Fig 4. Comparison of energy consumption on per query basis

6. Conclusion

The study designed an efficient k-means routing set of rules based fuzzy clustering to deal with difficulty of energy conservation for WSN. The proposed method makes use of the FKM clustering algorithms for clusters creation to produce a tremendously uniform clustering of 8 nodes by

lowering the spatial distance for the intra-cluster communication of average 1249. The creating equal clusters by creating use of fuzzy k-means purpose decrease in strength consumption of detector nodes that ends up in extending the WSN. The simulation results additionally proved that the purposed routing protocol considerably elevated the steadiness period 5, community lifetime 119 and throughput of the sensor network compared to K-means. Within the future, the projected work ought to be evolved and improved to help applications that need event detection the employment of mobile nodes and versatile BS.

References

- [1] Muhammad Rizwan 1 , Muhammad S. Nisar, 2 Hongbo Jiang 3,," F-MEEP: Fuzzy Logic Based Multihop Energy Efficient Routing Protocol for HWSN" IJCST, ISSN 2277-3061 Volume 15 Number 14.
- [2] J. S. Pan, S. C. Chu, T. K. Dao, and V. C. Do. "Improved Performance of Wireless Sensor Network Based on Fuzzy Logic for Clustering Scheme." In International Conference on Smart Vehicular Technology, Transportation, Communication and Applications, Springer, pp. 104-113, Cham, 2019.
- [3] Sasikumar Periyasamy, Sibaram Khara, and Shankar Thangavelu, "Balanced Cluster Head Selection Based on Modified k-Means in a Distributed Wireless Sensor Network", International Journal of Distributed Sensor Networks Volume 2019, Article ID 5040475, 11 pages.
- [4] Gaurang Raval 1 , Madhuri Bhavsar 2 , Nitin Patel 3,,"Performance Comparison of Various Clustering Techniques in Wireless Sensor Networks", IJCSC VOLUME 5 NUMBER 2 JULY-SEPT 2020 (ISSN).
- [5] Anand Gachhadar, Om Nath Acharya, "K-means Based Energy Aware Clustering Algorithm in Wireless Sensor Network", International Journal of Scientific & Engineering Research, Volume 5, Issue 5, May 2014, ISSN 2229-5518 ,IJSER © 2019.
- [6] Rajesh Patel, Sunil Pariyani, Vijay Ukani, Energy and Throughput Analysis of Hierarchical Routing Protocol (LEACH) for Wireless Sensor Network IJCA(0975 8887) Volume 20 No.4, April 2020.

- [7] Ran, H. Z., "Improving on LEACH Protocol of Wireless Sensor Networks Using Fuzzy Logic," *Journal of Information & Computational Science*, 2020.
- [8] Reebha SA. Fuzzy Logic Based Clustering With Firefly Optimized Routing Protocol For QoS Aware Wireless Sensor Networks. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*. 2021 Apr 28;12(10):2693-714.
<https://turcomat.org/index.php/turkbilmat/article/view/4886>
- [9] Talmale R, Bhat N. Energy Attentive and Pre-Fault Recognize Mechanism for Distributed Wireless Sensor Network Using Fuzzy Logic Approach. DOI: <https://doi.org/10.21203/rs.3.rs-226503/v1>
- [10] Trust Aware Energy Efficient Clustering for Secure Packet Transmission in Wireless Sensor Networks V. Nandalal, M.S. Sumalatha ,V. Anand Kumar and C. Santhosh Kumar. *Indian Journal of Science and Technology*, Vol 12(23), DOI: 10.17485/ijst/2019/v12i23/145375, June 2019
- [11] Energy Efficient and Secured Clustering Algorithm using Fuzzy Logic with K-means Method in MANET A. Y. Prasad and R. Balakrishna. *Indian Journal of Science and Technology*, Vol 12(19), DOI: 10.17485/ijst/2019/v12i19/144195, May 2019
- [12] Jasim, A.A; Idris, M.Y.I.; Razalli Bin Azzuhri, S.; Issa, N.R.; Rahman, M.T.; Khyasudeen, M.F.b. Energy-Efficient Wireless Sensor Network with an Unequal Clustering Protocol Based on a Balanced Energy Method (EEUCB). *Sensors* 2021, 21, 784.
<https://doi.org/10.3390/s21030784>
- [13] Mishra, P.K.; Verma, S.K. FFMCP: Feed-Forward Multi-Clustering Protocol Using Fuzzy Logic for Wireless Sensor Networks (WSNs). *Energies* 2021, 14, 2866. <https://doi.org/10.3390/en14102866> An Energy Efficient Clustering Algorithm in Wireless Sensor Networks for Internet of Things Applications
- [14] Xingjie Ci., Kang Wen, Ying Sun, Weifan Sun and Wei Deng. The 2nd International Conference on Computing and Data Science (CONF-CDS 2021) *Journal of Physics: Conference Series* 1881 (2021) doi:10.1088/1742-6596/1881/4/042035
- [15] Sachithanantham NC, Jaiganesh V . (2021) Enhanced Energy Efficient Routing Protocol (EEE-RP) to forward the Data Packets and to improve QoS in Wireless Sensor Networks by Means of Machine Learning Methods. *Indian Journal of Science and Technology* .14(14):1122-1132. <https://doi.org/10.17485/IJST/v14i14.477>

FAKE NEWS DETECTION USING MACHINE LEARNING

Dr. T. S. Suganya¹ Deepthi Jayadevan², Nethra R³, Praveen Kumar⁴

¹Assistant Professor, Department of Computer Applications, SRMIST, Ramapuram Campus, Tamil Nadu, India

^{2,3,4} UG Students Department of Computer Applications, SRMIST, Ramapuram Campus, Tamil Nadu, India

ABSTRACT

With the rapid growth of information online through the World Wide Web and social media sources like Facebook, Instagram, Twitter, etc., it has become a serious issue and it is really hard to identify whether the information is true or false. People start sharing this information without verifying it. This leads to a major problem of Fake news. Fake news imitates the real news and has a turn of events. This rapid growth of fake news has to be stopped and people should be able to identify whether the news is real or fake before sharing it with others.

Keywords: Fake news, Social media, Internet, Artificial Intelligence, Machine Learning, Passive Aggressive Classifier.

INTRODUCTION

The Internet plays a significant role in our lives today. Social media is one of the major uses of the Internet. We spend most of our time online on social media platforms. It has become an integral part of our lives. We gather a lot of information and news from social media than newspapers and news channels. But the standard of the news available in social media is less when compared to other news resources. Since the quality is low, the news is not trustworthy and this brings in the problem of fake news. Fake news contains false information which should be checked before it is shared.

OVERVIEW

In the previous works, they have used the Support Vector Machine to get a higher accuracy for the used datasets. Logistic regression and Support Vector Machine models combined together give better scores with larger datasets but the performance is doubtful. When Naive Bayes and Decision Tree algorithms were used the margin was not accurate when tested with samples. The advantage of Naive Bayes is that it works well with smaller datasets and has considerable importance. Whereas, Decision tree performs poorly and is not a good fit for fake news classification. While using Neural networks, the accuracy rate was very low consistently. Neural Networks did not work well with simple Machine Learning algorithms. N-Grams and Linguistic Analysis might give better results when combined with Neural Networks. There will be better performance when larger sample sizes are used. When N-Gram encoding and Bag of words were used to extract features with the Support Vector Machine there were better results. When Support Vector Machine, Naive Bayes and Semantic Analysis are used together a better prediction can be obtained.

When Multilayer Perceptron, Convolutional Neural Network(CNN), Recurrent Neural Networks(RNN) and Hybrid-CNN-RNN used CNN provided a robust solution.

BACKGROUND

Machine Learning is a branch of Artificial Intelligence that is used for predicting outcomes with a higher accuracy level with the help of data and algorithms. These Machine Learning Algorithms use historical data as input and are trained to predict new outcomes and possibilities.

The **Passive-Aggressive Classifier** is an algorithm in Machine Learning. This algorithm works best for applications that require and receive data in a continuous stream. The system can be trained by feeding the data individually, sequentially and in small groups. The system remains passive for correct predictions and is aggressive for incorrect predictions.

METHODOLOGY

The first step involved is to create or select datasets from various sources. We then use the **TfidfVectorizer** on our dataset. In the TF-IDF vectorizer, TF is the Term Frequency which counts the frequency of the words in the dataset and the IDF is the Inverse Document Frequency which measures how significant the term is. The vectorizer converts the raw document collector to a vectorized matrix with Tfidf features. Once the data is vectorized, we use the Passive-Aggressive classifier to classify whether the news is fake or real. And at last, the accuracy score tells us how accurate our model is. These are classified under four modules named collection of datasets, pre-processing, feature extraction and classification.

We first have to install the packages and libraries needed using pip. The libraries needed are NumPy, Pandas and sklearn. These can be installed using the following commands pip install numpy, pip install pandas and pip install sklearn. We are then importing these libraries and the model.

Datasets

A large number of news datasets are collected. Today the internet contains large amounts of electronic collections that often contain high-quality information. However, usually, the web provides more information than is required. The user wants to select the best collection of data for particular information needed in the minimum possible time. A data set is a collection of information. The data set lists values for each of

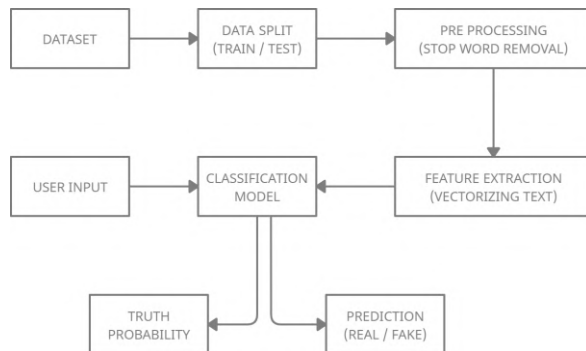
the variables used in the data set and multiple datasets are combined into one entity. The dataset that we are using is named “real_or_fake_news.csv” and has a shape of 6335 * 4. There are 4 columns in the dataset. They are id, title, text and label. The label denotes if the news is true or false. After importing the libraries the dataset is read and is split into training and test sets.

```
[6335 rows x 4 columns]
0    FAKE
1    FAKE
2    REAL
3    FAKE
4    REAL
...
6330   REAL
6331   FAKE
6332   FAKE
6333   REAL
6334   REAL
Name: label, Length: 6335, dtype: object
4741  NAIROBI, Kenya -- President Obama spoke out Sun...
2069  Killing Obama administration rules, dismantlin...
4274  Dean Quindlan, a former attorney, is the hea...
5376  WashingtonBlog VICHN's Jake Tapper hit the ...
6028  Some of the biggest issues facing America this ...
...
5910  From the day we are born into this world, we...
3915  Chosistan , Iraq, Phenomenon of Terrorism By...
1428  Senate Minority Leader Harry M. Reid (D-Nev.) ...
4367  WASHINGTON -- The U.S. government started keepi...
2822  Gary Johnson is the presidential nominee for t...
Name: text, Length: 5908, dtype: object
```

IMPLEMENTATION

System Architecture:

In the **Pre - Processing** process, the given input news text is processed for removing redundancies, inconsistencies and separate words



and documents are prepared for the next step. We use Stop Word Removal for removing such words. Stop words are a group of commonly used words in a language. Some of the stop words in English are “a”, “the”, “is”, “are”, etc.

In the **Feature Extraction** process, we use the Text Vectorization (TF - IDF vectorizer) algorithm to transform the text into a usable vector. This counts the number of words with a higher frequency and measures how significant the term is. Then the TF-IDF vectorizer is initialized for stop word removal with a frequency of 0.7 for English words.

In the **Classification** process, we find the truth accuracy of the news and predict the output for the user. The prediction can either be Fake or True. This is done using the Passive-Aggressive Classifier Algorithm. The Passive-Aggressive Classifier is initialized and fit in the vectorizer. The accuracy score is calculated and the prediction for the test set is done.

```
import pandas as pd
import pandas as np
import itertools
from sklearn.model_selection import train_test_split
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.linear_model import PassiveAggressiveClassifier
from sklearn.metrics import accuracy_score, confusion_matrix
```

CONCLUSION

The proposed work uses the Passive-Aggressive classifier for the prediction. This can work on large datasets since the data can be fed in small terms. The accuracy rate seems to be above 90%. But the datasets have to be updated on a regular basis. This might work better when it is combined with other Machine Learning methods and algorithms.

FUTURE WORKS:

This can further be developed using multiple machine learning algorithms and a larger number of datasets for better prediction and accuracy scores. Other machine learning algorithms can also be combined together for future works. This can also be developed into a desktop or mobile application which is more reliable.

REFERENCES

- [1]. Hassan Ali, Muhammad Suleman Khan, Amer Alghadhban, Meshari Alazmi, Ahmad Alzamil, Khaled Al-Utaibi, Junaid Qadir. 2021. “All Your Fake Detector Are Belong to Us: Evaluating Adversarial Robustness of Fake-News Detectors Under Black-Box Settings”
- [2]. Allcott, H. and Gentzkow, M. (2018). Social Media and Fake News in the 2016 Election. NBER.
- [3]. MykhailoGranik and VolodymyrMesyura. “Fake news detection using naive Bayes classifier.” First Ukraine Conference on Electrical and Computer Engineering (UKRCON). Ukraine: IEEE. 2017.
- [4]. Understanding Support Vector Machine algorithm from examples. Retrieved March 2, 2018.
- [5]. J. A. Nasir, O. S. Khan, and I. Varlamis, “Fake news detection: A hybrid CNN-RNN based deep learning approach,” *Int. J. Inf. Manage. Data Insights*, vol. 1, no. 1, Apr. 2021, Art. no. 100007.
- [6]. Z. Zhou, H. Guan, M. Bhat, and J. Hsu, “Fake news detection via NLP is vulnerable to adversarial attacks,” in *Proc. 11th Int. Conf. Agents Artif. Intell.*, Feb. 2019, pp. 794–800.
- [7]. Parikh, S. B., & Atrey, P. K. (2018, April). Media-Rich Fake News Detection: A Survey. In 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR) (pp. 436-441). IEEE.
- [8]. Stahl, K. (2018). Fake News Detection in Social Media.
- [9]. Zhang, J., Cui, L., Fu, Y., & Gouza, F. B. (2018). Fake news detection with the deep diffusive network model.
- [10]. Helmstetter, S., & Paulheim, H. (2018, August). Weakly supervised learning for fake news detection on Twitter. In 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM) (pp. 274-277). IEEE..

AUTOMATIC NUMBER PLATE RECOGNITION SYSTEM

¹Dr. S.Jayachandran, ²Mrs.V. Devi, ³Mr. J.Sanurag Nair, ⁴Mr Nithish Kumar,

^{1,2}Assistant professor, SRM Institute Of Science and Technology, Ramapuram Campus, Chennai, Tamil Nadu, India

^{3,4} Final Year BCA Students, SRM Institute Of Science and Technology, Ramapuram Campus, Chennai, Tamil Nadu, India

Abstract

Automatic number plate recognition (ANPR) is the extraction of vehicle license plate information from an image. The system model uses already captured images for this recognition process. First the recognition system starts with character identification based on number plate extraction, splitting characters and template matching. ANPR as a real-life application has to quickly and successfully process license plates under different environmental conditions, such as day time. It plays an important role in numerous real-life applications, such as automatic toll collection, traffic law enforcement, parking lot access control, and road traffic monitoring. The system uses different templates for identifying the characters from input image. After character recognition, an identified group of characters will be compared with database number plates for authentication. The proposed model has low complexity and less time consuming in terms of number plate segmentation and character recognition. This can improve the system performance and make the system more efficient by taking relevant samples. at the same time compared their advantages and disadvantages, which provide the basis for license plate recognition.

Keywords: ANPR, Granulometry, Morphological, Grayscale, OCR, Ostu's Classifier and etc.

1. Introduction

In the fields of artificial intelligence, augmented reality, and other advancements, text detection in natural photographs is critical. It aids in the removal of image noise and the recognition of text. However, because of the diversity in imaging conditions, such as lighting, specular reflections, turbulence, obscurity, and the proximity of blocks over the content, as well as the changeability of the content itself, such as scale, introduction, literary style, and style, it is a problematic issue. As a result, great text finding calculations should be robust against such swings, Because it is used in vision-based applications, text detection is essential in everyday life. Backdrop complexity, text orientation, background complexity, scene text diversity, and interference problems are among the challenges it now faces.

When security forces pursue a car or are unable to apprehend a vehicle that has broken traffic laws, it is clear that complications arise. On a busy day, authorities find manually logging vehicle numbers in a parking lot to be extremely time consuming. So, in order to make the entire process autonomous, we may install this system that will automatically recognize the vehicle that is breaking the traffic regulations, take a picture of it, and save the number in a database so that the owner can be fined later. The system can be used in parking lots to

capture pictures of cars and log their license plates in a database (or the cloud, if connected to the internet).

1.2 Web Technology

Web Technology has become a highly important aspect in today's globe because to advanced terminologies. It describes both design and code methodologies, therefore we'll employ both basic and advanced web abilities in this project. APIs will become increasingly crucial as all technologies move toward an API-centric approach. Nowadays, RESTful Web service frameworks are now available for every major programming language. The REST architecture style is a networked hypermedia application architecture. We're utilising the Bootstrap framework for the front end and a SQL database for the back end. The project is dynamic because it is run using the Java programming language. Here rating facility is also available i.e. every patient authorized patient can give rating to the doctor in terms of stars so it will become easy to choose best doctor for patient.

2. System Analysis

2.1 Existing System

On real photos, the proposed system correctly detects and recognizes the car number plate. This system can also be utilized for traffic management and security. The Past The term "digital image processing" refers to the use of a computer to process digital images. It can also be described as the application of computer techniques to improve an image or extract relevant data. When photos were first transferred by submarine cable between London and New York, one of the early applications of digital image was in the newspaper industry. In the early 1920s, the Bart lane cable image transmission system lowered the time it took to send a picture across the Atlantic from more than a week to less than three hours.

2.2 Proposed System

When security forces pursue a car or are unable to apprehend a vehicle that has broken traffic laws, it is clear that complications arise. On a busy day, authorities find manually logging vehicle numbers in a parking lot to be extremely time consuming. So, in order to make the entire process autonomous, we may install this system that will automatically recognize the vehicle that is breaking the traffic regulations, take a picture of it, and save the number in a database so that the owner can be fined later. The system can be used in parking lots to capture pictures of cars and log their license plates in a database (or the cloud, if connected to the internet).

Instead of using sensors or RFID, ANPR cameras are utilized to capture the license number plates of automobiles. Vehicle parking management systems that are currently in use. ANPR stands for Automatic Number Plate Recognition. Cameras have a distinct advantage over other technologies.

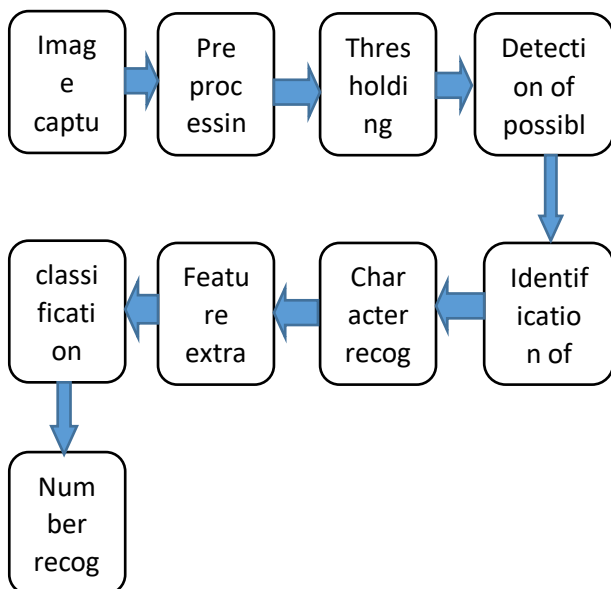
3. Proposed System Architecture

3.1 Objective

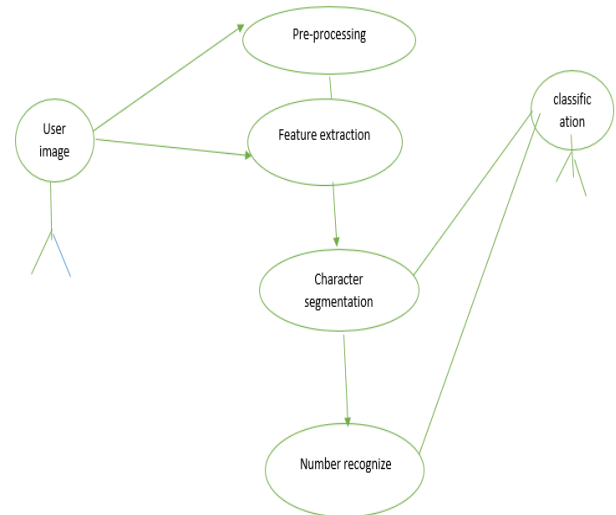
Automatic Number Plate Recognition (ANPR) is a computer vision system that quickly and accurately recognizes car license plates from photos without the need for human interaction. It has grown in importance in recent years as a result of three primary factors: the increasing number of cars on the road, the rapid development of image processing techniques, and the large variety of real-world applications that this technology provides [1]. Traffic enforcement, automatic toll collection, and parking lot access management are some of the most common uses of ANPR systems. However, this technology is widely employed.

The development of ANPR systems, on the other hand, is no easy process, as it must contend with several problems posed by environmental and number plate changes. In the case of the former, changes in illumination or backdrop patterns have a significant impact on number plate identification. In effect, shifting illumination can diminish the clarity of the car image, and backdrop patterns complicate the operation of locating the number plate. The placement, quantity, size, font, color, or inclination of number plates, in particular, are extremely problematic aspects in the development of a consistent ANPR system.

3.2 Architecture Diagram



3.3 Data Flow Diagram:



4. System Implementation

Pre-Processing: - The process of preparing data for future analysis, and it includes the procedures listed below.

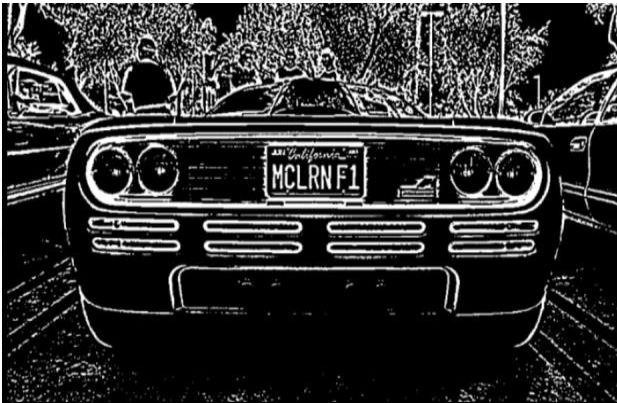
1. Resizing
2. Color transformation
3. Filtering is number three.

To remove unnecessary data and filter our input image, we employ a pre-processing technique. It will compress our input image, which will speed up the rest of the process. If it was a picture, the image is translated into frames, which are then utilized to identify the car number. After that, the image or frame is transformed from RGB to grayscale. To remove various types of sounds from the resultant grey scale image, a median filter is utilized. The median filter also emphasizes on the image's high frequency portions, which aids with edge identification.

The training requires software and tools such as classifiers, which feed vast volumes of data, analyze them, and extract key features. The categorization process' goal is to sort all of the pixels in a digital image into one of several categories. The classification is usually done using multi-spectral data.



Thresholding:- It's a method of turning a grayscale or full-color image into a binary image. This is frequently done to aid image processing by distinguishing between "object" or foreground pixels and background pixels. Thresholding is an image segmentation technique in which the pixels of an image are changed to make the image easier to analyze. Thresholding is the process of converting a color or grayscale image into a binary image, which is just black and white. The term "thresholding" refers to the process of determining whether or not.



Classification: -The purpose of image classification is to recognize and describe the features in an image in terms of what they truly represent on the ground as a distinct grey level (or color). Image categorization is the most important component of digital image analysis. Because object categorization is difficult, picture classification has become an important task in computer vision. The act of categorizing and classifying images is known as image classification.

Feature-extraction:- Feature extraction is a step in the dimensionality reduction process, which divides and reduces a large set of raw data into smaller groupings. As a result, processing will be simpler. The fact that these enormous data sets have a large number of variables is the most crucial feature. To process these variables, a large amount of computational power is required so.

OCR:- OCR stands for optical character recognition, and it is used to tackle the problem of identifying a large number of different characters. Both handwritten and printed characters can be identified and converted into a machine-readable digital data format. Consider any serial number or code made up of digits and letters that needs to be converted to a digital format.

OCR can be used to transform these codes to digital output. The technology employs a variety of techniques. To put it another way, the image is processed, the characters are recovered, and the image is finally identified. The automated conversion of an image-based PDF, TIFF, or JPG into a text-based machine-readable file is a popular application of OCR technology. Digital files that have been OCR-processed, such as receipts, contracts, invoices, and financial statements.

6. Conclusion

In this project, the proposed Automatic Number Plate Recognition System has been implemented successfully and passively in the above perspectives and the proposed system can be used for technological and social advancement in the present scenario for the future improvements. This system proposed works with higher accuracy and enhances the capabilities.

7. Future Enhancement

Vehicle owner identification, vehicle model identification, traffic management, vehicle speed regulation, and vehicle position monitoring can all be done with ANPR. It may also be used as a multilingual ANPR to automatically detect the language of characters based on the training data. It can give a variety of benefits, including traffic safety enforcement, security-in the event of suspicious vehicle behavior, ease of use, immediate information availability-in comparison to manually checking car owner registration records, and cost effectiveness for any country. Some image improvement methods, such as super resolution [30], [31], should be focused on low resolution photos. The majority of ANPR systems are designed to process a single vehicle number plate, but in real-time, many vehicle numbers plates may be present while the images are being taken.

References

- [1] Hong Yamin. The two generation License Plate Segmentation and Recognition photo identification [J]. Journal of Lanzhou industry college, 2013,06:18-22.
- [2] Zhao Xingwang, Li Tianyang, Wang Liang, Zhou Jing. The two generation License Plate Segmentation and Recognition number recognition system of [J]. computer and modernization based on digital equipment, 2014,06:132-136.
- [3] Fu Ronghui. The research and design of vehicle license plate recognition system in traffic management system [J]. International Journal of Signal Processing, Image Processing and Pattern Recognition, v 9, n 3, p 445-456, 2016.
- [4] Yingyon Zou, Jian Zhai, Yongde Zhang, Xinyan Cao, Guangbin Yu, Juhui, Chen. Research on algorithm for automatic license plate recognition system[J]. International Journal of Multimedia and Ubiquitous Engineering, v 10, n 1, p 101-108, 2015.
- [5] Tao Hong, Gopalakrishnam A.K. License plate extraction and recognition of a Thai vehicle based on MSER and BPNN[J]. Proceedings of the 2015 7th International Conference on Knowledge and Smart Technology (KST), p 48-53, 2015.

PIXEL HIGH DENSITY NOISE FILTER METHOD FOR DENOISING IMAGES USING IMAGE PROCESSING TECHNIQUES

Suriya Priyadharsini M¹, Dr. J. G. R. Sathiaseelan²

Research scholar, Department of Computer Science, Bishop Heber College (Affiliated to Bharathidasan University), Trichy, Tamilnadu, India

Julimca.sigc@gmail.com

Associate Professor, Department of computer Science, Bishop Heber College (Affiliated to Bharathidasan University), Trichy, Tamilnadu, India

jgrsathiaseelan@gmail.com

ABSTRACT:

Noise is a serious issue. While sending images via electronic communication, Impulse noise, which is created by unsteady voltage, is one of the most common noises in digital communication. During the acquisition process, pictures were collected. It is possible to obtain accurate diagnosis images by removing these noises without affecting the edges and tiny features. In this paper the comparison of denoising is discussed and a new decision-based Pixel median filter used to remove impluse noise. Using Mean Square Error (MSE), Peak Signal to Noise Ratio (PSNR), and Structure Similarity Index Method (SSIM) metrics, the paper examines the performance of Gaussian Filter (GF), Adaptive Median Filter (AMF), Median Filter and PHDNF. The picture quality metrics between the original clean photographs and denoised images affected by various amounts of pretend salt and paper noise, as well as speckle noise, are calculated and provided as experimental results. According to quality metrics, the PHDNF Method produces a superior result than the existing filter method.

Key Term: Denoising, Median filter, Adaptive Median Filter (AMF), Median filter, Gaussian Filter, pixel High Density Noise Filter (PHDNF)

I. INTRODUCTION

In the acquisition and transmission of an image, impulse noise is frequently affected; there are two forms of impulse noise: fixed-valued impulse noise and random-valued impulse noise.[1] Fixed-valued impulse noise, often known as salt and pepper noise, is one of the most common types of image noise; it has a significant impact on image processing and analysis, segmentation, and other tasks. As a result, good impulse noise reduction is necessary. The mean filter [2] and median filter [3] were first developed for removing fixed-valued impulsive noise. The mean filter, on the other hand, was shown to be unable to preserve the image's structure and edge information, but the median filter is preferred due to its ease of use and high performance. However, the standard median filter was shown to be unsuitable of simultaneously removing noise and preserving structural information, particularly for high density noise, because it processes all pixels independently whether they are noisy or not, destroying the noise-free pixels.

Jianhua Pang [2] proposed a new median filter based on the proportion of variance. The algorithm contains two steps to restore the corrupted picture. The first step is to identify the random valued noise and the second is to process the noisy points. But the algorithm is adept in removing the random valued impulse noise and restoring contaminated images. Ashwani Kumar Yadav [3] presented an improved median filtering algorithm which utilizes the previously processed neighboring pixel value to get the better image quality. The proposed improved median filtering is slightly better than SMF, and it will also loss the details of an image while removing the noise. Saroj K.Meher and Brijraj Singhawat[4] proposed an improved recursive and adaptive median filter for high density impulse noise. Adaptive operation of the filter is justified with the variation in size of working window which centered at noisy pixels. The noisy pixels are filtered through the replacement of their values using both noise free pixels of current working window and previously processed noisy pixels of that window. It is one of the improved adaptive median filters which suit for high density noise.

Low rank based Weighted Nuclear Norm Minimization (WNNM) is done on detail coefficients to disclose the sparsity property of DWT, as proposed by Saurabh Khar[8]. To approximate the low-rank denoised version of the subbands, WNNM is applied to the group matrix of non-local comparable patches of DWT detail subbands. Furthermore, the DWT approximation coefficients contain less edge and structure information. Non-local Means (NLM) filter with Square-Chord distance is also employed to denoise speckle noise from DWT approximation coefficients in the suggested method.

SidheswarRoutray [9] present an innovative framework for image denoising based on non-subsampled shearlet transforms and bilateral filtering (NSST). They begin

by decomposing a noisy input image into high and low frequency coefficients using the NSST. The noise from the low incidence constants is detached using the weighted bilateral filter (WBF), while noise from the high frequency coefficients is detached by means of thresholding. The outputs from both processes are merged to create the final image.

II. EXISTING SYSTEM

Median filter

The median filter is a simple non-linear filter that can be used to remove noise. The targeted noisy pixels are replaced by the median value of their neighbours in this. The number of neighbours is determined by the size of the filtering window then using intermediate value of the sort sequence to replace the certain point of the window. The mathematical expression of median filter is as follows:

$$I(i, j) = \text{Median}(n(k))$$

Among the formula, K is the number of pixels in the window, n is a sort of gray value sequence, i is the pixel horizontal coordinate, and j is the pixel vertical coordinate.

Adaptive median filter

The adaptive median filter can handle much more spatially dense impulse noise, and also performs some smoothing for non-impulse noise. In AMF the filter size changes depending on the characteristic of the image. Adaptive impulse detection uses centre weighted median filters. AMF has a lot higher computing efficiency, and one of its advantages is that it keeps non-impulse pixels.

Weighted median filter

A weighted median filter controlled by evidence fusion is proposed for removing noise from images with contrast. It has a great potential for being used in rank order filtering and image processing. The weights of the filter are set based on intensity value of the pixels in the image. Here we used four weights such as 0, 0.1, 0.2 and 0.3. if the intensity value of the pixel is 0 then consider the weight of the pixel is 0. Else if the range of pixel intensity between 1-100 then the weight is 0.1, else if the range of pixel intensity between 101-200 then the weight is 0.2, otherwise the weight of the pixel is 0.3. the above weights are multiplied with pixel intensity. after that the median filter is applied for calculate weighted median filter.

Gaussian Filter

Gaussian is a linear filter that uses Gaussian function to pick weights [8]. Gaussian smoothing filters are kind of efficient low-pass filters in the space field or the incidence domain, in particular to suppress noises that have been distributed normally. It has a broad prospect of using images. The following can be expressed for one-dimensional Gaussian zero mean function:

$$g(x) = e^{-\frac{x^2}{2\sigma^2}} \quad (1)$$

The Gaussian distribution parameter describes Gaussian functions in their width. And, in the processing of images, here also use the 2-d discrete Gaussian zero-medium feature as a smooth filter [9].

$$g(i, j) = e^{-\frac{i^2+j^2}{2\sigma^2}} \quad (2)$$

III. PROPOSED SYSTEM

Let's have a look at how PHDNF works. We determine a window size for each pixel in a noisy image by gradually extending the window from 3x3 to the maximum size. If the grey value of the window's centre pixel is equal to the maximum or minimum value, it is considered noisy; otherwise, it is a noise-free pixel. if the center of pixel is noisy then the maximum number of pixels in the window that are repeated then compute mean value and assigned as it's a new grey value. For an 8-bit greyscale image min=0 and ma=255.

Let P: = $[p(i; j)]$ consisting of pixels $p(i; j)$, where i and j are in the range of 1 to n, respectively, is considered noisy.

IV. THE METHOD OF APPROACH

Step 1. For all i and j,

Step 1.1. If $p(i; j)$ is noisy, and at least one of the pixels in the window with a 3x3 size that accepts this as its center is not noisy, then for all of the pixels I and j in the window if there is at least one $p(i; j)$ such that $245 < p(i; j) < 255$ or $0 < p(i; j) < 10$, then

- Find the maximum number of pixels in the window that are repeated except min and max
- The mean of the values should be calculated.
- Replace this value with $p(i, j)$

Step 1.2. If $p(i; j)$ is noisy, and at least one of the pixels in the window with a 5x5 size that accepts this as its center is not noisy, then for all of the pixels I and j in the window if there is at least one $p(i; j)$ such that $245 < p(i; j) < 255$ or $0 < p(i; j) < 10$, then

- Find the maximum number of pixels in the window that are repeated except min and max
- The mean of the values should be calculated.
- Replace this value with $p(i, j)$

Step 1.3 If $p(i; j)$ is noisy, and at least one of the pixels in the window with a $(2k + 1) \times (2k + 1)$ size that accepts this as its center is not noisy, then for all of the pixels I and j in the

window if there is at least one $p(i; j)$ such that $245 < p(i; j) < 255$ or $0 < p(i; j) < 10$, then

- Find the maximum number of pixels in the window that are repeated except min and max
- The mean of the values should be calculated.
- Replace this value with $p(i,j)$ where $0 < k < \min[m,n]$

Step 2. Otherwise, keep the value of $x(i; j)$.

Exercis 1.1 Assume P is a 512x512 image with a lot of noise. Accept $p(314,350)$ as the using the 3x3 window accepts center is noisy. Therefore, the window satisfies the conditions given in Step 1.1. In that case, the maximum repetitive pixel values in the window are found as 238 and 246, and the mean value of these values is evaluated as 242. Therefore, the value of 242 is set to the noisy pixel, and the window becomes as in Figure 1

238	225	246
251	0	255
246	287	238

(a)

238	225	246
251	242	255
0	246	238

(b)

$$(238 + 246)/2 = 242$$

Fig1.Illustration of Example 1.1

Exercis 2.2 Assume P is a 512x512 image with a lot of noise. Accept $p(314,350)$ as the centre pixel using the 3x3 window shown in Figure 2a. The pixels in this window are all noisy. As a result, the window fails to meet the requirements established in Step 1. 1. In Figure 2b, accept the centre pixel p in the 5 x5 sized windows (314,350). Since $p(312,350) = 6$, at least one of the pixels in this window looks to be noise-free; As a result, the window's maximum repeating pixel values are discovered to be 6 and 10, and the median value is calculated as 8. Figure 2c the result of setting value 8 to the noisy pixel.

255	255	255
0	255	255
0	0	255

(a)

255	10	6	0	10
255	255	255	255	0
0	6	255	255	255
255	0	10	255	6
6	0	0	6	255

(b)

255	10	6	0	10
255	255	255	255	0
0	6	8	255	255
255	0	10	255	6
6	0	0	6	255

(c)

$$(6 + 10) = 8$$

Fig1.Illustration of Example 2.1

Statistical measurement

a) Peak Signal to Noise Ratio (PSNR) and Mean Square Error (MSE)

Evaluated to these current approaches, the presentation of the developed methodology is both qualitative and quantitative. Quantitative measurement is conducted using the parameters Peak noise signal ratio (PSNR) and Mean Square Error (MSE). PSNR is listed

$$PSNR = 10 \log_{10} \frac{255^2}{MSE}$$

Where MSE is the Mean Square Error between the original seed image (input seed image) and the resolution image after noise removal (output seed image). MSE is defined as followed:

$$MSE = \frac{1}{NM} \sum_{x=l}^M \sum_{y=l}^N [g(x,y) - f(x,y)]^2$$

Where: M= Number of image rows; N: Number of image columns, g: Input image (Normal image), f: output image (Filtered image). The Less value than MSN is the best result in image.

b)Structural Similarity Index Metric (SSIM)

The structural similarity index metrical tests visual consistency on the basis of the assumption that the HVS has a structure closely associated with the original to collect and development structural information from original images, and a high quality image [10]. The SSIM calculation is performed on a local window by breaking the entire image into N x N-size frames. The three SSIM functions combined to quantify the image quality[11] include light measurement, contrast measurement and structure measurement.

The luminance comparison of image x and y is defined as

$$l(x,y) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1}$$

4.5 Analysis

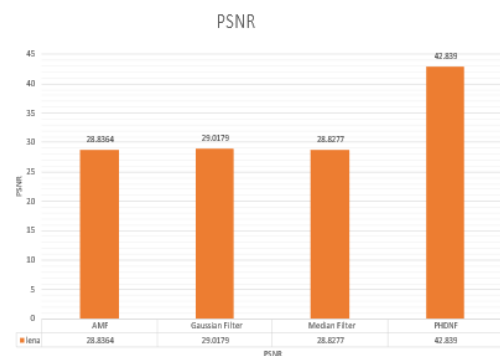
In Analysis here we compute the Peak Signal to Noise Ratio (PSNR) and Mean Square Error (MSE) and Structural Similarity Index Metric (SSIM)

V. Results



Fig 1 Image Filter on Lena Image

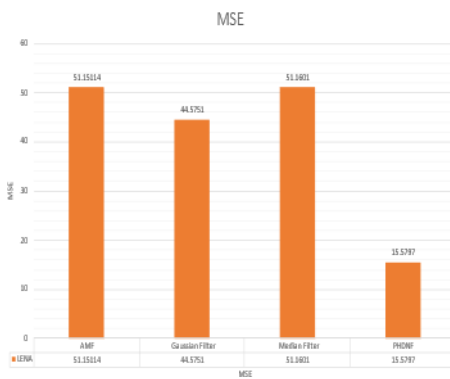
Graphical represent of PSNR values



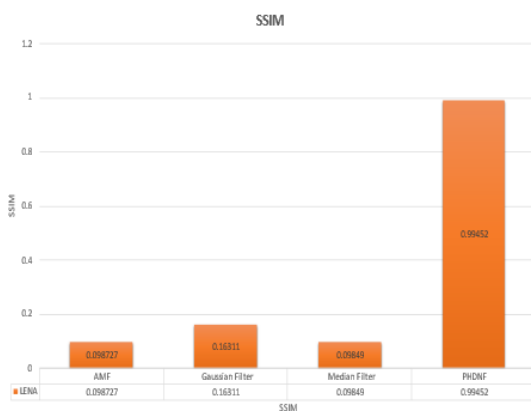
S.NO	AMF	Gaussian Filter	Median Filter	PHDNF
PSNR	28.8364	29.0179	28.8277	42.839
SSIM	0.098727	0.16311	0.09849	0.99452
MSE	51.15114	44.5751	51.1601	15.5797

TABLE I. Average PSNR/SSIM/MSE value of the denoising results of images of the Lena image

Graphical represent of MSE values



Graphical represent of SSIM values



VI. Conclusion

Our proposed methods perform operations on pixel, several image filtering algorithms can be efficiently implemented. In this analysis compared various noise-removal image filtering algorithms. The work concludes, that the proposed filter shows better noise reduction than another algorithm, since the edges for a specific fixed window size can be maintained. A comparative study is conducted by performance analysis of these filters based on quality parameters PSNR, MSE, and SSIM. Experimental results reveal that PHDNF performs well on noise removal in images. The filtering techniques suggested showed the lowest MSE values and the highest PSNR and SSIM values.

REFERENCES

- [1] K. H. Jin and J. C. Ye, "Sparse and low-rank decomposition of a Hankel structured matrix for impulse noise removal," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1448–1461, Mar. 2018
- [2] Jianghua Zhang, Sheng Zhang, and Shaojun Zhang. A Median Filter Based on the Proportion of the Image Variance. In *Technology, Networking, Electronic and Automation Control*, 2016 International conference on . IEEE, 2016 :123-127
- [3] Ashwani Kumar Yadav, Ratnadeep Roy, Raj Kumar and et al. Algorithm for Denoising of Color Image Based on Median Filter. In *Image Information Processing, 2015 3th International conference on.IEEE, 2015 pp.428-432*
- [4] Saroj K.Meher , Brijraj Singhawat. An improved recursive and adaptive median filter for high density impulse noise. *Int. J. Electron. Commun.(AEU)* 2014(68) :1173-1179
- [5] Shen Dehai 卞Liu Dacheng 卞Xing Tao, "Adaptive-selection filtering strategy algorithm o noise density detection,"

Application Research of

Computers 2012(2) : 761-763

[6]Dong, Guanfang, Yingnan Ma, and Anup Basu. "Feature-Guided CNN for Denoising Images from Portable Ultrasound Devices." *IEEE Access* 9 (2021): 28272-28281.

[7]Singh, Himanshu, Sethu Venkata Raghavendra Kommuri, Anil Kumar, and Varun Bajaj. "A new technique for guided filter based image denoising using modified cuckoo search optimization." *Expert Systems with Applications* 176 (2021): 114884.

[8]Khare, Saurabh, and Praveen Kaushik. "Speckle filtering of ultrasonic images using weighted nuclear norm minimization in wavelet domain." *Biomedical Signal Processing and Control* 70 (2021): 102997.

[9] Routray, Sidheswar, Prince Priya Malla, Sunil Kumar Sharma, Sampad Kumar Panda, and G. Palai. "A new image denoising framework using bilateral filtering based non-subsampled shearlet transform." *Optik* 216 (2020): 164903.

[10]Khetkeeree, Suphongsa, and Parawata Thanakitivirul. "Hybrid Filtering for Image Sharpening and Smoothing Simultaneously." In *2020 35th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, pp. 367-371. IEEE, 2020.



SRM INSTITUTE OF SCIENCE AND TECHNOLOGY

Ramapuram Campus

COLLEGE OF SCIENCE AND HUMANITIES

(A Place for Transformation)

DEPARTMENT OF COMPUTER APPLICATIONS (BCA)

1ST INTERNATIONAL CONFERENCE ON
ARTIFICIAL INTELLIGENCE AND
DATA SCIENCE (ICAIDS-2022)

OUR ASSOCIATES



Author

Dr. Agusthiyar R

Professor and Head, Department of Computer Applications,
SRMIST, Ramapuram Campus.

Conference Secretary

Mrs. S. Suriya, Asst Professor

Mrs. J. Shyamala Devi, Asst Professor



978-93-5620-075-3