Annals of Computer Science and Information Systems
Volume 42

Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering (RICE 2024)

December 27–28, 2024
Hyderabad, TG, India

Vijender Kumar Solanki, Pradeep Kumar, Tran Duc Tan, Manuel Cardona (eds.)

PTI

Annals of Computer Science and Information Systems, Volume 42

ISSN: 2300-5963

Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering

Vijender Kumar Solanki, Pradeep Kumar,
Tran Duc Tan, Manuel Cardona (eds.)

POLSKIE TOWARZYSTWO INFORMATYCZNE
POLISH INFORMATION PROCESSING SOCIETY

Annals of Computer Science and Information Systems, Volume 42
Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering

ISBN 978-83-973291-5-7
ISSN: 2300-5963
DOI: 10.15439/978-83-973291-5-7

Ul. Solec 38/103
00-394 Warsaw
Poland

Contact: secretariat@fedcsis.org
http://annals-csis.org/

Cover photo:
Aleksander Denisiuk,
Elbląg, Poland

Also in this series:

Volume 41: Communication Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS), **ISBN WEB: 978-83-973291-0-2, ISBN USB: 978-83-973291-1-9**

Volume 40: Position Papers of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS), **ISBN WEB: 978-83-969601-9-1, ISBN USB: 978-83-969601-0-8**

Volume 39: Proceedings of the 19th Conference on Computer Science and Intelligence Systems (FedCSIS), **ISBN WEB: 978-83-969601-6-0, ISBN USB: 978-83-969601-7-7, ISBN ART 978-83-969601-8-4**

Volume 38: Proceedings of the Eighth International Conference on Research in Intelligent Computing in Engineering, **ISBN WEB: 978-83-969601-5-3**

Volume 37: Communication Papers of the 18th Conference on Computer Science and Intelligence Systems, **ISBN WEB: 978-83-969601-3-9, ISBN USB: 978-83-969601-4-6**

Volume 36: Position Papers of the 18th Conference on Computer Science and Intelligence Systems, **ISBN WEB: 978-83-969601-1-5, ISBN USB: 978-83-969601-2-2**

Volume 35: Proceedings of the 18th Conference on Computer Science and Intelligence Systems, **ISBN WEB 978-83-967447-8-4, ISBN USB 978-83-967447-9-1, ISBN ART 978-83-969601-0-8**

Volume 34: Proceedings of the Third International Conference on Research in Management and Technovation **ISBN 978-83-965897-8-1**

Volume 33: Proceedings of the Seventh International Conference on Research in Intelligent and Computing in Engineering, **ISBN WEB: 978-83-965897-6-7, ISBN USB: 978-83-965897-7-4**

Volume 32: Communication Papers of the 17th Conference on Computer Science and Intelligence Systems, **ISBN WEB: 978-83-965897-4-3, ISBN USB: 978-83-965897-5-0**

Volume 31: Position Papers of the 17th Conference on Computer Science and Intelligence Systems, **ISBN WEB: 978-83-965897-2-9, ISBN USB: 978-83-965897-3-6**

Volume 30: Proceedings of the 17th Conference on Computer Science and Intelligence Systems, **ISBN WEB: 978-83-962423-9-6, ISBN USB: 978-83-965897-0-5**

DEAR Reader, we are delighted to share with you a glimpse of the 9th International Conference on Research in Intelligent Computing in Engineering (RICE 2024). RICE 2024 is organized by Stanley College of Engineering and Technology For Women, Hyderabad, Jointly co-organized by School of Technology, Maulana Azad National Urdu University (A Central University), Hyderabad, Telangana, India; Universidad Don Bosco, El Salvador, CA, during December 27-28, 2024.

We are truly thankful to the Polish Information Processing Society (PTI), Poland for approving the proceedings of the 8th International Conference on Research in Intelligent Computing in Engineering (RICE 2024). It is appearing in the Annals of Computer Science and Information Systems series by PTI (ISSN-2300-5963). The series has been submitted to Copernicus, DBLP, Cross Ref, Scholar, BazEkon, Open Access Library, Academic Keys, Journal Click, PBN, and ARIANTE. At this stage, the efforts, whole-hearted support, and suggestions given by Editor-in-Chief Prof. Marcin Paprzycki and Prof. Maria Ganzha are highly applaudable and commendable.

We are pleased to report that various researchers are interested in participating in the 9th edition of RICE 2024. It is a privilege for us to hear four keynote speakers from different countries share their insightful perspectives on conference-related topics. The information is provided below:

- Dr. Abdul Khadar Jilani, University of Technology, Bahrain
- Dr. Bui Tien Son, Hanoi University of Industry, Hanoi, Vietnam
- Dr. Rajeeb Dey, National Institute of Technology, Silchar, Assam, INDIA
- Dr Sudan Jha, Kathmandu University, Nepal

Finally, we would like to take this opportunity to express our sincere appreciation to the Advisory Board, Technical Program Committee, Organizing Committee, International Scientific Committee, institutions, industries, and volunteers, who contributed to the success of this conference either directly or indirectly.

Proceeding's Editors – RICE 2024:

Vijender Kumar Solanki, Stanley College of Engineering and Technology for Women, Hyderabad, Telangana, India.

Tran Duc Tan, Phenikaa University, Hanoi, Vietnam.

Pradeep Kumar, Maulana Azad National Urdu University, Hyderabad, Telangana, India.

Manuel Cardona, Universidad Don Bosco, El Salvador, Central America.

Annals of Computer Science and Information Systems, Volume 42

Ninth International Conference on Research in Intelligent Computing in Engineering

December 27–28, 2024. Hyderabad, India

TABLE OF CONTENTS

The Ninth International Conference on Research in Intelligent Computing in Engineering

rice2024 Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering pp. 1–10
DOI: 10.15439/2024R81
ISSN 2300-5963 ACSIS, Vol. 42

Seq2Seq Transformer-Based Model for Optimized Chinese-to-English Translation

Sumayya Afreen
Department of Computer Science and Engineering
Stanley College of Engineering and Technology for Women
Hyderabad, India
asumayya@stanley.edu.in

Nguyen Thi Dieu Linh
Department of Science and Technology
Hanoi University of Industry
Hanoi, Vietnam
nguyenlinh79.haui@gmail.com

Sritisha Kodur
Department of Computer Science and Engineering
Stanley College of Engineering and Technology for Women
Hyderabad, India
sritishakodur@gmail.com

Asma Begum
Department of Artificial Intelligence and Data Science & Computer Engineering
Stanley College of Engineering and Technology for Women
Hyderabad, India
basma@stanley.edu.in

***Abstract*—The use of transformer models for machine translation from Chinese to English is examined in this research. The transformer design, which is well-known for its self-attention mechanism, makes it possible to handle Chinese's intricate linguistic structures with efficiency. We assess the model's effectiveness using benchmark datasets, examine its translation correctness through cosine similarity scores, Rouge metric scores and draw attention to important issues including managing context and sentence structure inconsistencies. We also explore situations in which language complexity is observed to result in low accuracy, providing valuable information for enhancing future models. This paper highlights areas for optimization in practical situations and shows how transformers might improve translation quality.**

***Index Terms*—Cosine Similarity, Language complexity, Machine translation, Rouge Metrics, Transformer models.**

I. Introduction

RECENT advances in Natural Language Processing (NLP), especially regarding the transformer model, have had a significant impact on machine translation progress. Vaswani et al. (2017) created the transformer architecture, which has completely changed the way textual input is processed by removing the drawbacks of conventional Recurrent Neural Networks (RNNs) such parallelization problems and vanishing gradients. As a result, transformers have excelled at several NLP tasks, including machine translation, by exploiting self-attention mechanisms for better addressing long-range dependencies in text Transformer-based models, which exclusively rely on self-attention processes instead of RNNs' sequential processing approach, have completely changed the field of machine translation. With the help of this attention mechanism, the model can concentrate on several phrase components at once, better capturing long-range dependencies and enhancing translation accuracy. Unlike RNNs, which suffer from vanishing gradient difficulties and are hard to parallelize, transformers enable efficient parallel processing, which not only accelerates training but also enhances the model's capacity to handle complicated sentence structures. The better performance of the transformer over traditional models in a variety of language pairs, including Chinese-to-English, has been used to illustrate its usefulness in neural machine translation (NMT) tasks [1].

There are more than just linguistic differences in structure when translating from Chinese to English. The gap between literal and non-literal translation procedures has long been highlighted by translation theorists such as Newmark (1981) and Vinay & Darbelnet (1958). This distinction is particularly crucial when translating between languages that are as dissimilar as Chinese and English. In Chinese-to-English translation, nonliteral translation strategies are essential since many Chinese expressions—particularly idioms and colloquialisms—cannot be translated literally without losing their original meaning. To guarantee that the translated meaning stays faithful to the original, nonliteral expressions like "刑让我的辛苦白费了" (which means "Don't let my hard work be wasted") must be understood in context. Research has demonstrated that these kinds of nonliteral translations have historically proven difficult for machine translation models, especially RNN-based systems, often producing grammatically correct but semantically inaccurate translations [2].

The self-attention mechanism of the transformer model is crucial in this situation. Transformers are better able to handle translations including nonliteral phrases or complex sentence structures by letting the model determine the relative

value of various words in a sentence. Additionally, recent improvements in pre-trained language models such as BERT and GPT have further increased the capabilities of transformer models by including contextual awareness and semantic nuances [3]. For example, when applied to Chinese-English translation tasks, transformers have demonstrated considerable gains in BLEU scores, a metric typically used to evaluate the accuracy of machine translations. Transformer models have certain drawbacks even if they typically function incredibly well. One aspect that still requires work is their capacity to handle exceedingly long and contextually complex statements. Transformers can catch word dependencies thanks to the self-attention mechanism, but it can occasionally have trouble keeping long sentences coherent, particularly when translating materials that call for in-depth cultural knowledge or subject-specific expertise. According to studies by Chen et al. (2020) and Zhang et al. (2021), transformer models occasionally fall short of accurately capturing the context of colloquial terms or domain-specific jargon, which might result in less accurate translations in specialised sectors like legal or medical writings [4]

The linguistic and cultural disparities between Chinese and English have frequently presented difficulties for those translating public signage in China. Amenador and Wang (2020) draw attention to this problem by pointing out that a lot of translations fall short of the original meaning, which causes misunderstanding among audiences that speak English. Their work applies functional theory to the analysis of translation errors in public signs, highlighting the significance of tailoring translations to the target audience's communicative needs rather than following the source text exactly. The study proposes that translators can enhance the quality of translations and hence improve China's foreign image by adopting a purpose-driven approach [5]. To sum up, the transformer model, which offers a more reliable and scalable method than previous RNN-based systems, has greatly advanced the field of Chinese-to-English translation. Its self-attention mechanism makes it possible to handle literal and nonliteral translations more effectively, which makes it an effective tool for resolving the difficulties that come with translating between Chinese and English. To solve issues with context retention and the translation of texts that are extremely domain-specific, more study is still needed. Future advancements in transformer-based designs, such the incorporation of trained models like BERT or GPT, have the potential to significantly enhance translation quality in a variety of language contexts. The purpose of this study is to examine these developments and go over the issues still facing near-human translation accuracy [6].

This paper aims to develop and evaluate a robust Seq2Seq Transformer model tailored for accurate and efficient Chinese-to-English translation. The study leverages ROUGE metrics to assess the quality of generated translations and employs cosine similarity to measure semantic alignment with reference translations, ensuring both linguistic and contextual fidelity.

The paper is organized as follows: Section 2: Literature Review provides an overview of existing research on neural machine translation and the application of Seq2Seq Transformer models, highlighting gaps addressed in this work. Section 3: Methodology details the model architecture, data preprocessing steps, and the evaluation framework using ROUGE and cosine similarity metrics. Section 4: Results presents the experimental outcomes and compares the model's performance with existing benchmarks. Section 5: Future Work outlines potential enhancements, including broader language coverage and advanced optimization techniques. Finally, Section 6: Conclusion summarizes key findings and their implications for translation systems.

II. Related Work

Verb Semantics for English-Chinese Translation, published in 1995 by Martha Palmer and Zhibiao Wu, offers a thorough analysis of the difficulties associated with machine translation, with a particular emphasis on lexical differences between Chinese and English verbs. The authors stress that current NLP and machine translation (MT) systems frequently rely on precompiled lexicons that are unable to handle unforeseen word usage, and they frequently presume a fixed number of verb meanings. To overcome these shortcomings, this study suggests an improved method of verb semantics that takes lexical selection issues and verb sense expansions into consideration

Lexical Divergences between English and Chinese Verbs: The main problem noted in the paper is the substantial vocabulary differences between Chinese and English, especially in the structure of verbs and verb phrases. Verbs in English frequently combine action and consequence into a single lexical component. For example, in English, verbs like break might indicate a state or an activity without clearly stating the procedure or result. On the other hand, Chinese verbs often use compound formulations to clarify the action and the outcome. Due to this distinction, translating words between the two languages can be difficult since in Chinese, verb compounds like "da-sui" (hit-into-pieces) express both the action and the object's end condition. Given that verb translations must take into consideration both of these divergences, more sophisticated translation processes than those offered by conventional lexicons are needed.

Limitations of Existing Systems: Current MT and NLP systems are criticised by Palmer and Wu for their strict reliance on static verb sense lists, which makes them unsuitable for managing sense extensions or unforeseen verb usages. For example, English verbs like break may have numerous senses depending on context—whether referring to the physical state of an object, a break in continuity, or the malfunctioning of a device. The authors contend that the adaptability and inventiveness present in natural language are insufficiently addressed by systems built around a limited range of predetermined verb senses. This is especially evident when translating between Chinese and English, two languages with quite distinct vocabulary patterns. The tech-

nique of selectional limitations, which is often used in MT systems to limit verb arguments by specified categories, is criticised in this study. Although this approach can deal with senses of verbs, it is not very effective when dealing with more general linguistic events. This is particularly true when translating from English, which has more generalised verbs, to Chinese, which needs specificity in the description of actions and outcomes. The paper *Verb Semantics for English-Chinese Translation* by Martha Palmer and Zhibiao Wu (1995) provides an in-depth exploration of the complexities involved in machine translation, specifically focusing on lexical divergences between English and Chinese verbs. The authors emphasize that existing machine translation (MT) and natural language processing (NLP) systems often assume a fixed number of verb senses, relying on precompiled lexicons that lack the flexibility needed for dealing with unexpected word usages. This paper addresses these limitations by proposing an enhanced approach to verb semantics that accounts for both lexical selection problems and verb sense extensions.

Conceptual Lattices and UNICON: The authors suggest a novel approach based on conceptual lattices, a technique for encoding verb senses that enables more dynamic and adaptable mappings between verbs in various languages, to get around these restrictions. Verb meanings are arranged into a hierarchical structure by a conceptual lattice, which allows related senses to be placed together according to common semantic elements. With this method, the system can recognise verbal similarities between languages even in the absence of direct translations. By expanding verb senses beyond their preset definitions, the research prototype system UNICON puts this idea into practice and shows more accurate lexical selection.

One of the system's main advantages is its capacity to expand verb senses. The system can determine the closest related sense in the conceptual lattice when an unexpected verb usage happens and utilise this information to suggest a translation. This approach works especially well when there isn't a direct translation available in the target language. When an English verb, like break, is employed in an unusual way, for instance, the system can look up comparable verb senses in Chinese and choose the most appropriate one. UNICON can handle unexpected verb usages that would normally be outside the scope of typical lexicons by permitting sense extensions.

By emphasising the shortcomings of static verb sense lists and selectional limits in handling lexical divergences between English and Chinese, the paper significantly advances the field of machine translation. Palmer and Wu provide a more adaptable and context-sensitive method of verb semantics through their conceptual lattice framework, which enhances translation accuracy by taking into consideration the wider variety of verb usages and their potential extensions. The creation of UNICON, a useful tool for improving machine translation systems, proves the viability of this strategy [7].

Wazib Ansar, Saptarsi Goswami, and Amlan Chakrabarti's paper "A Survey on Transformers in NLP with Focus on Efficiency" delves deeply into the emergence of transformer-based models in NLP, emphasising the need to strike a balance between these models' performance and efficiency. The introduction of transformer models like BERT, GPT, and XLNet has transformed the way linguistic tasks are carried out as the area of NLP has developed. In tasks like text summarization, sentiment analysis, and machine translation, these models have proven to perform at the cutting edge. Transformers are resource-intensive and require a large amount of memory, processing power, and energy, thus this advancement is not without a price (2406.16893v1).

The Evolution of NLP and the Rise of Transformers: The article begins with a succinct overview of NLP's past, showing how computational techniques have changed over time, moving from rule-based systems to machine learning techniques. The capacity to handle complex linguistic patterns was restricted by earlier techniques like rule-based and classic machine learning approaches, such Support Vector Machines (SVM) and Naive Bayes, which also required intensive feature engineering. Recurrent Neural Networks (RNNs) were a breakthrough in deep learning; nonetheless, it was hampered by issues such as vanishing gradients and managing long-range dependencies. With their self-attention mechanism, transformer models—which were first shown by Vaswani et al. (2017)—addressed these issues and enabled improved parallelization and long-range context capture (2406.16893v1).

As the study discusses, transformers have a few benefits over earlier systems. These models can provide varying weights to distinct words in a sentence according on how relevant they are to the context according to the self-attention mechanism, which improves understanding of linguistic subtleties. Because of this, transformers are especially useful for jobs like translation and summarization that call for an awareness of the larger context. Models such as GPT (Generative Pretrained Transformer) and BERT (Bidirectional Encoder Representations from Transformers) established new performance standards for NLP tasks, resulting in their broad use in research and industry (2406.16893v1).

The authors provide research showing how training big NLP models has an impact on the environment. For instance, Strubell et al. (2019) discovered that the energy required for several transatlantic trips can be like that required for training a sizable NLP model. There is increasing worry about how these models affect the environment, which is driving scientists to look into more accurate but more efficient alternatives. This is especially pertinent in light of sustainability and the rising demand for environmentally friendly artificial intelligence solutions (2406.16893v1).

Strategies for Improving Transformer Efficiency the study examines several techniques for improving transformer models' operating efficiency to address efficiency concerns without sacrificing performance. These include model com-

pression methods that have been used to decrease the size and computational load of these models, such as pruning, quantization, and knowledge distillation (Gordon et al., 2020; Hinton et al., 2015). Through pruning, a trained model's less significant weights are removed, creating a more efficient version that uses less memory and processing resources. Conversely, quantization lowers the weights in the model's numerical precision, which can help minimise resource requirements while keeping Through concentrating on condensing the information included in big models into more manageable and effective models, scientists can capitalise on the advantages of transformer architectures while lessening their impact on the environment and operational expenses [8].

Sparse Attention Mechanisms: Another unique way to boosting transformer efficiency mentioned in the research is the use of sparse attention mechanisms. For lengthy sequences, traditional transformers can become computationally expensive as they must compute attention scores for every pair of input tokens. The computational load is decreased by sparse attention techniques, which restrict the attention computations to a subset of pertinent tokens (Child et al., 2019). Models like Long former and Reformer indicate that by leveraging sparse attention patterns, transformers can effectively manage longer sequences without a substantial reduction in performance. These models are appropriate for a range of NLP applications because they not only increase efficiency but also maintain the contextual knowledge that transformers are recognized for.

Hardware's Place in Efficiency: The significance of hardware developments for transformer model optimization is also highlighted in the article. The speed and effectiveness of training and inference procedures can be greatly increased by using custom hardware solutions, such as Field Programmable Gate Arrays (FPGAs) and Tensor Processing Units (TPUs) (Jouppi et al., 2017). Researchers can reduce energy usage and enhance the performance of these models by customizing hardware for transformer computations. The efficiency issues raised by large models can be effectively resolved by combining transformer topologies with optimized hardware.

A. Research Directions for the Future

The necessity for continued research to investigate novel approaches for raising the efficiency of transformer models is emphasised in the paper's conclusion. As natural language processing (NLP) advances, it is critical to create models that are both highly effective and ecologically sound. The authors support a comprehensive strategy that integrates developments in model architecture, hardware, and algorithms to produce transformer systems that are more effective. Subsequent investigations ought to focus on discovering novel approaches to optimise the advantages of transformers while reducing their ecological footprint and resource needs. [8]

The development, applications, and prospects of Transformer-based models in natural language processing (NLP) are reviewed in detail in the paper "End-to-End Transformer-Based Models in Textual-Based NLP" by Abir Rahali and Moulay A. Akhloufi. The architecture, training modalities, and particular applications of Transformer-based models are highlighted by the authors, who also emphasise how these models have a transformative effect on NLP tasks.

The paper starts out by highlighting the profound change in natural language processing (NLP) that deep learning (DL) architectures—specifically, Recurrent Neural Networks (RNNs)—have brought about. It then shifts to the enhanced powers of Transformer models, which Vaswani et al. The self-attention mechanism, which allows the model to take into account long-range dependencies between tokens in a sequence, is the main innovation that distinguishes Transformers from RNNs and Convolutional Neural Networks (CNNs). This allows for more effective and scalable processing of textual data.

The review by the authors is structured around the development of Transformer-based (TB) models, beginning with the basic Transformer model and moving on to several adaptations that tackle architectural and performance issues. Despite their prior effectiveness in sequential data processing, the research demonstrates that classic RNN-based models were constrained by problems including disappearing gradients and the incapacity to manage long-term relationships efficiently. Transformer models tackle these challenges by leveraging self-attention techniques, allowing parallelization, and making them more efficient for tackling large-scale NLP jobs.

The review claims that the flexibility of Transformer-based models is their main advantage. The paper offers thorough insights into several important Transformer variations, each intended for a particular NLP purpose. Examples of models that use the Transformer architecture for different purposes are discussed in detail, such as BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer), where BERT is used for natural language understanding and GPT is used for text generation.

The literature study also focuses on newer models like RoBERTa, which improves on BERT by modifying training processes and using larger datasets. XLNet, which aims to combine the benefits of auto-regressive and auto-encoding models, and T5 (Text-to-Text Transfer Transformer), which reformulates all NLP tasks into a text-to-text format, are two other noteworthy models that were explored.

The writers argue that Transformer models are useful for a variety of natural language processing (NLP) tasks, such as sentiment analysis, question-answering, summarization, and machine translation. Because these models are better at learning contextual representations of text, they have shown themselves to be more effective than earlier architectures at handling such tasks. Although the enormous processing demands for training these models are acknowledged in the review as a potential drawback, pre-training and transfer learning advancements have somewhat alleviated these difficulties.

The use of pre-training and fine-tuning strategies in Transformer-based models is covered in a noteworthy section of the article. These models can be fine-tuned on smaller, task-specific datasets after gaining a generalised grasp of language through pre-training on large-scale, unlabeled datasets. It is now commonplace in NLP to use this transfer learning strategy to attain excellent performance on tasks with little labelled data. A discussion of the future directions for Transformer-based model research rounds off the overview. Enhancing these models' scalability and efficiency remains an open problem, especially in low-resource languages and data-poor areas. In order to further improve performance, the authors also point out the possibility of hybrid models, which integrate the advantages of several neural network architectures.

Overall, the study presents a complete overview of the state of Transformer-based models in NLP, highlighting their evolution, applicability, and ongoing research concerns [9]. Natural language processing (NLP) can be used to analyze and compare various English translations of The Analects, a classic work of Confucian philosophy. This is explored in the paper "A Semantic Similarity Analysis of Multiple English Translations of The Analects: Based on a Natural Language Processing Algorithm" by Liwei Yang and Guijun Zhou. To better understand how translation choices and styles affect the overall semantic integrity of the original text, the study compares and calculates the semantic similarities across 15 English translations of The Analects using a variety of NLP techniques.

The first section of the paper discusses The Analects' cultural relevance and its enduring influence on Chinese and international intellectual traditions. The Analects, which has been translated into English multiple times since the 17th century, is an important resource for international dialogue as well as Chinese cultural and philosophical study. However, the availability of different translations causes issues for readers, especially those unfamiliar with the historical and cultural background of the original text. The need to give readers a better grasp of the semantic divergence and convergence of various translations is the driving force behind this work.

The writers begin by reviewing The Analects' historical English translations, pointing out the variety of approaches taken by various translators. While some translators prioritised keeping the text as faithful to the source Chinese as possible, others focused more on making it readable and accessible to English-speaking audiences. The translators are divided into three categories in the paper: native English speakers, Chinese translators with Western education, and Chinese translators with traditional education. This classification is predicated on the linguistic and cultural backgrounds of the translators, which the authors speculate may impact translation decisions and, in turn, alter the semantic output of the translations.

The authors use five distinct natural language processing (NLP) techniques, namely TF-IDF (Term Frequency-Inverse Document Frequency), Word2Vec, GloVe, BERT, and SimHash, to evaluate the semantic similarity between translations. These algorithms provide various methods for analysing the semantics of text. Word2Vec and GloVe, for instance, build vector representations of words to capture semantic associations, while TF-IDF estimates the relevance of words in a text based on their frequency and how rare they are across documents. BERT, a transformer-based approach, leverages deep learning to capture increasingly complicated semantic patterns. Conversely, SimHash provides a less sophisticated but simpler comparison by gauging the similarity between texts based on their binary hash representations.

The study focuses on fifteen popular English versions of The Analects that have attracted a lot of interest from users on sites like Google Scholar, Goodreads, and Amazon. The authors determine pairwise similarities across translations using various NLP methods, and they provide the results as quantitative data. Their results show that most translations have a high degree of semantic similarity, especially those done by well-known translators like James Legge and D.C. Lau, although there are also notable discrepancies. The translators' employment of different Chinese annotations and interpretive frameworks, rather than their origins or the historical era in which they worked, is primarily responsible for these disparities, the authors contend.

One of the paper's main findings is that the selection of Chinese annotations is a significant factor in figuring out how similar various translations are semantically. For instance, translations that rely significantly on Zhu Xi's annotation, a Song Dynasty scholar, show higher levels of semantic coherence. Semantic divergence is higher in translations that attempt to rethink the original meaning of the text or include more contemporary remarks.[10].

The study "Advances in Chinese Natural Language Processing and Language Resources" by Jianhua Tao, Fang Zheng, Aijun Li, and Ya Li, offers an overview of recent developments in Chinese Natural Language Processing (CNLP), highlighting the construction and exploitation of linguistic resources and consortiums. The paper underlines the relevance of data-driven techniques and linguistic resources in NLP research, focusing on how the availability and sharing of corpora have contributed to major gains in both text and speech processing in Chinese.

Key Themes and Scope:The authors start out by stressing how important well-constructed corpora and linguistic data are to the advancement of CNLP research. For accuracy and usefulness, modern NLP techniques—especially statistical approaches—heavily rely on real-world data. Due to the complexity of Chinese and its distinct linguistic properties, it is essential to have access to large and reliable corpora in order to make significant progress in natural language processing (NLP) applications such as text categorisation, machine translation, speech recognition, and more. The wide variety of NLP tasks, such as machine translation, syntactic parsing, part-of-speech (POS) tagging, and word segmentation, are

also covered in the study. Notably, with accuracy rates of 98% and 95% in Chinese, word segmentation and POS tagging have attained notable success. But more difficult jobs like syntactic and semantic parsing continue to be difficult, partly because of the ambiguity and complexity inherent in natural language

B. Contributions to Chinese NLP

Lexicons: From generic word segmentation to specialized lexicons for Chinese geographic names, proper nouns, and organizations, the paper examines a variety of lexicons developed for diverse reasons. Notably, a large amount of word frequency data that is essential for many NLP applications can be found in the Chinese Web 5-gram Corpus.

POS-Tagged Corpora: POS-tagged resources, such as the People's Daily Corpus, offer crucial training data for machine learning models in applications like information retrieval and text categorization.

Multilingual Corpora: As machine translation has grown in popularity, the use of multilingual corpora—which pair Chinese with languages like English and Japanese—has become crucial for enhancing translation precision. Resources such as the Tsinghua Chinese Treebank are helpful in the development of tools like syntactic parsers and event detection systems, which are essential for higher-level NLP tasks. Another important aspect of CNLP that is covered in the study is speech processing. Specialized speech corpora have proven extremely beneficial for tasks such as speaker identification, synthesis, and speech recognition. As an illustration, the CASIA Mandarin Corpus is cited as a crucial tool for enhancing Mandarin voice synthesis and recognition. The study also highlights the significance of emotional speech data and regional dialects, both of which are utilized to enhance speech recognition systems' performance in a wider range of scenarios.

The importance of resource-sharing programs such as the Chinese Linguistic Data Consortium (CLDC) and the Chinese Corpus Consortium (CCC) is emphasized in the paper. The creation, gathering, and distribution of linguistic resources for use in scholarly and industrial applications is greatly aided by these consortiums. By making high-quality data available to researchers, they enable the development of better-performing CNLP systems and facilitate the replication of results across the field [11].

III. Methodology

A. Transformer model

The Transformer model, which broke with the conventional sequential data processing techniques of Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks, was first presented by Vaswani et al. in their seminal paper "Attention Is All You Need" (2017). This change fundamentally altered the field of Natural Language Processing (NLP). The self-attention mechanism, which is the foundation of the Transformer model, enables it to assess the connections between various input sequence elements regardless of how far apart they are from one another. The Transformer processes all tokens at once, in contrast to RNNs, which process tokens in a sequential manner. This makes the Transformer more computationally efficient and better suited for capturing long-range dependencies.

The encoder and the decoder are the two primary parts of the Transformer model's architecture. After processing the input sequence, the encoder creates hidden representations of the input. Then, using the previously generated output and these hidden representations, the decoder creates the output sequence. A multi-head self-attention mechanism and a position-wise feedforward network are the two sub-layers that make up each of the several layers that comprise the encoder and decoder. The model may focus on pertinent tokens and their context by attending to diverse areas of the input concurrently thanks to the multi-head self-attention method.

The application of positional encoding is another significant feature of the Transformer concept. The model misses the natural order of words in a series since it does not process input tokens sequentially. To ensure word order is taken into account during training, positional encodings are added to each token to provide the model with information about its location in the sequence.

NLP tasks including text summarization, language modelling, and machine translation have been transformed by transformers, which greatly outperform earlier models in terms of accuracy and speed. They are especially useful for activities like machine translation and question-answering systems that need to comprehend intricate relationships between tokens because of their capacity to manage long-range dependencies and parallelize computations.

B. seq 2 seq model

A particular use of the Transformer design for sequence-to-sequence operations is the Seq2Seq Transformer model, which entails producing an output sequence based on a supplied input sequence. This architecture is extensively used for jobs where one sequence needs to be converted into another, such as speech recognition, text synthesis, and machine translation.

Under the Seq2Seq Transformer model, the encoder generates a context-aware representation of the complete input by processing the input sequence first. Every word in the input sequence is examined in connection to every other word in the phrase, not only in isolation. This enables the model to comprehend the context-specific meaning of every phrase. In machine translation, for instance, the term "bank" in the sentence "I went to the bank" can mean different things depending on whether the context indicates that it's a riverbank or a financial institution.

The next token in the output sequence is then generated by the decoder using these context-aware representations from the encoder and the previously generated output tokens. In order to ensure coherence and consistency throughout the translation or sequence production, the decoder also uses self-attention mechanisms to examine all of the previously generated tokens. For example, when translating from

English to Chinese, the decoder creates a natural translation by utilising the context of the source English sentence in addition to the Chinese words that it has already generated.

The Seq2Seq Transformer model's superiority over conventional RNN-based Seq2Seq models lies in its capacity to manage intricate dependencies and distant interactions in sequences. When processing lengthy sequences, RNNs' sequential data processing frequently results in issues like disappearing gradients, where the model finds it difficult to remember information from previous tokens. Transformers are able to get over these restrictions and outperform other models on tasks that require lengthy sequences of events by processing all tokens concurrently and utilising self-attention processes.

Furthermore, the Transformer is far faster than RNNs or LSTMs due to its capacity for parallel processing, particularly when working with big datasets or lengthy sequences. Because of this, it is now the preferred model for a wide range of complex natural language processing tasks, especially in real-time applications such as live translation or automated speech recognition.

To sum up, the Seq2Seq Transformer is an effective model that performs well in tasks involving the conversion of one sequence into another. It does this by utilising parallel processing and the attention mechanism to increase accuracy and speed. Its design has served as the foundation for numerous cutting-edge NLP models, including Google's T5 and OpenAI's GPT, proving its adaptability and efficacy in a variety of contexts.

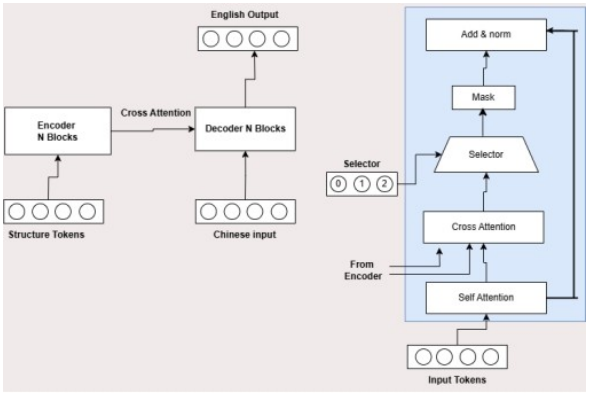


Fig. 1: Architecture Diagram

C. Data Set Used

A training set and a test set are the two halves of the dataset. There are one million sentence pairings in the training set and one thousand in the test set. The machine translation model is trained using the training set, and its performance is assessed using the test set.

The dataset's sentences are pre-processed to eliminate extraneous characters and symbols before being encoded in Unicode format. Tokenization is the process of breaking the sentences up into separate words or phrases and giving each one a unique ID. The words in the input and output sequences of the machine translation model are represented by the IDs.

Link:https://www.kaggle.com/datasets/qianhuan/translation

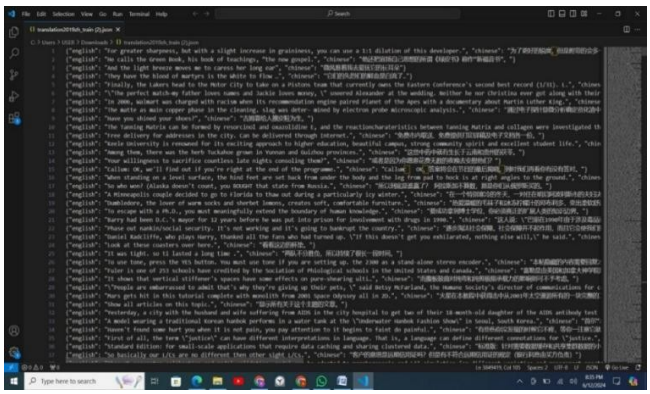


Fig. 2: Data Set Example

D. Technique used

Self-Attention Mechanism: The multi-head self-attention mechanism, which calculates relationships between every token in a sequence simultaneously, is the central component of the Transformer. Because of this, the model can concentrate on pertinent words regardless of where they are in the text, which makes it very useful for translation tasks where context is crucial.

Structure of the Encoder-Decoder: The encoder in the model is designed to handle the Mandarin input, while the decoder produces the English output. The input sequence is transformed into context-aware representations by the encoder, which the decoder then combines with previously produced output to predict the next word.

Positional Encoding: Word order is not taken into account by self-attention by itself, so positional encodings are added to the token embeddings to help the model comprehend the structure of the sequence.

Layers that feed forward: Following attention, the attended representations are processed by a feedforward neural network in each layer of the encoder and decoder, honing them before sending them to the subsequent layer.

Personalised Learning Rate Calendar: A Custom Schedule is used to regulate the learning rate. It does this by modifying the rate according to the number of warm-up steps and the model's dimensionality (d_model). This facilitates a more effective convergence of the model during training.

Functions of Accuracy and Loss: To stop the network from learning inaccurate predictions for padding tokens, a masked loss function is only used to compute loss for non-padding tokens.

Masked accuracy makes sure that padded values are ignored and that the accuracy computation only considers real tokens.

E. Pseudocode: Chinese-to-English Transformer Model

Step 1: Import Libraries

Import TensorFlow, libraries for data handling, file management (e.g., os, urllib).

Step 2: Download and Extract Data

Download dataset (Chinese-English pairs), check file type (ZIP/TAR), and extract to a target directory.

Step 3: Preprocess Dataset

Tokenize Chinese/English text, add <SOS>, <EOS>, <PAD>.

Convert tokens to numerical IDs, pad sequences to uniform lengths.

Step 4: Positional Encoding

Define positional encoding function using sine and cosine functions to represent token positions.

Step 5: Attention Mechanism

Implement scaled dot-product attention to calculate token relationships using softmax for weights.

Step 6: Transformer Encoder

Apply multi-head attention and feed-forward layers with residual connections and normalization.

Step 7: Transformer Decoder

Process encoder output and previous tokens with self-attention and encoder-decoder attention.

Add feed-forward layers, residuals, and normalization.

Step 8: Build Transformer Model

Stack encoder/decoder layers.

Add a final linear layer with softmax for token prediction.

Step 9: Training Setup

Use cross-entropy loss (ignore <PAD>), Adam optimizer, and learning rate scheduler.

Initialize weights.

Step 10: Train Model

For each epoch and batch, pass inputs through encoder and decoder, compute loss, and update weights.

Step 11: Evaluate Model

Translate Chinese inputs with encoder and autoregressive decoding in the decoder.

Use BLEU or accuracy for evaluation.

Step 12: Save Model

Save model weights, architecture, and checkpoints for reuse or fine-tuning.

Step 13: Post-Training Evaluation

Evaluate test data with BLEU score or similar metrics.

Optionally visualize attention maps for insight.

Output: High-quality English translations with performance metrics.

IV. Results

A. Metrics Used

Cosine similarity: The similarity between two vectors in an inner product space is measured by cosine similarity. It determines whether two vectors are pointing in about the same direction and is calculated by taking the cosine of the angle between them. In text analysis, it is frequently used to gauge document similarity.

Thousands of attributes, each documenting the frequency of a specific word (such a keyword) or phrase in the document, might be used to represent a document. As a result, every document is an object that is represented by a term-frequency vector. For instance, Table 2.5 shows that the word "team" appears five times in Document1, yet "hockey" appears three times. A count value of 0 indicates that the word "coach" is not present throughout the document.

$$SC(x, y) = x.\ y\ /\ ||x|| \times\times ||y||, \quad (1)$$

where the product of the vectors "x" and "y" is x. The length (magnitude) of the two vectors "x" and "y" is equal to ||x|| and ||y||. The regular product of the two vectors "x" and "y" is ||x|| × × ||y||.

Rouge Metrics: Recall serves as the primary basis for the ROUGE ratings, which were really created with text summary in mind, where the model-generated text is typically shorter than the reference text. In essence, ROUGE contrasts the reference and candidate summaries in terms of n-grams, word pairings, and word sequences.

Important ROUGE Measures

1. The n-gram overlap between the reference text and the generated text is measured by ROUGE-N.
2. To capture structural similarity, ROUGE-L uses the longest common subsequences (LCS).
3. Contiguous matches that weigh more than other n-grams are weighed by ROUGE-W.

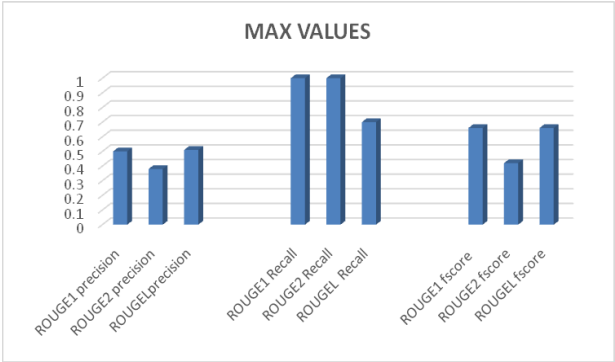


Fig. 3: Maximum Values of Precision, Recall and F score.

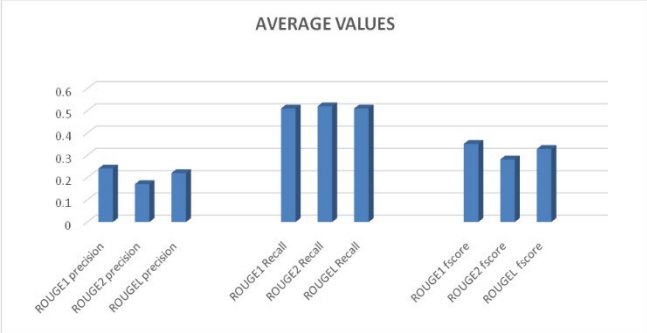


Fig. 4: Average Values of Precision, Recall and F score.

4. ROUGE-S is a measure of skip-bigram overlap, which considers two words that might not be next to each other.

ROUGE-N = {Number of matching n-grams} \ {Total n-grams in the reference}. (2)

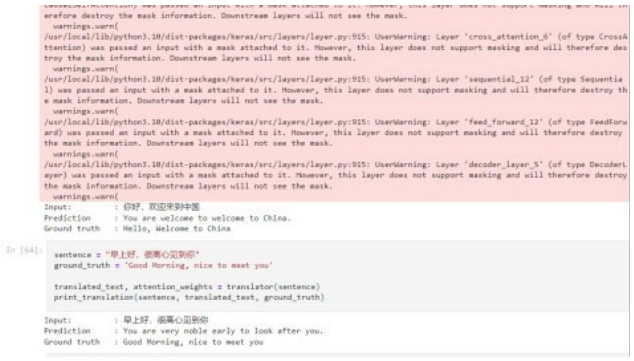


Fig. 5: Screenshot of executed Translations.

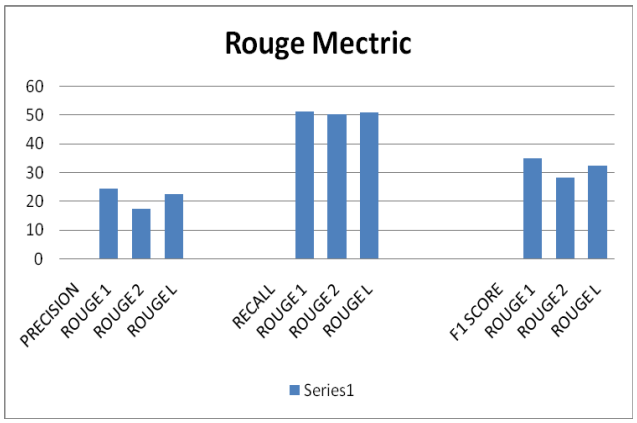


Fig. 6: Rouge metric Values of precision, Recall and F Score

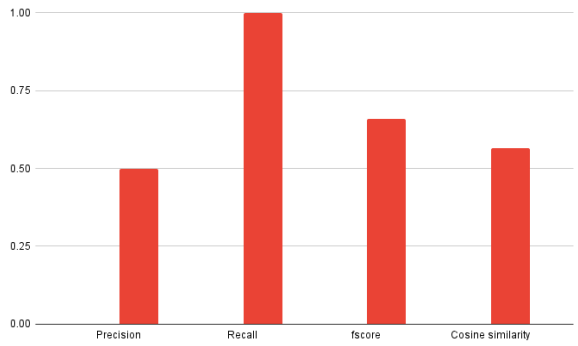


Fig. 7: Rouge metric Values along with Cosine Similarity.

V. Future Work

Model Optimisation and Compression: Given the computational complexity of transformer models, one potential direction for future research is to optimise the model for reduced resource usage. Model pruning, quantization, and knowledge distillation are a few methods that can be used to minimise the computing burden and size of the model without compromising accuracy. This is particularly crucial if the model is meant to be used with low-resource devices, like cell phones or edge devices. By reducing the environmental impact of operating large-scale transformers, optimising the model would also address concerns about energy use.

Table 1: Average Rouge metric values

ROUGE Metric	Precision	Recall	F Score
ROUGE-1	0.24	0.51	0.367
ROUGE-2	0.17	0.52	0.323
ROUGE-L	0.22	0.51	0.353

Including Attention Visualisation: Including attention visualisation tools would be a worthwhile and fascinating addition. This would make it easier for consumers or researchers to comprehend the translational focus of the model. Debugging and optimising translation outcomes, particularly in instances where errors arise, may be facilitated by visualising the attention weights between source and target tokens. Researchers can further refine the model to increase its accuracy by identifying regions where the model may be focused improperly.

Including Contextual Understanding: Existing transformer models frequently handle sentences as stand-alone entities without taking larger documents or conversations into account. In the future, the model could incorporate a context-aware translation process that considers earlier sentences when translating a given sentence. This would be very helpful when translating dialogue or longer texts where the meaning is spread out over several sentences.

Real-time Translation and Deployment: Further development of the model to facilitate real-time translation, which would enable speech-to-text or chat translation services, may also be part of future work. Low-latency translations are critical in real-time scenarios, and the algorithm might be modified to function well in such a circumstance. The model's use cases might be increased and made more widely available by integrating it into chat apps or deploying it as a service via APIs

VI. Conclusion

In conclusion, the creation of Transformer-based Chinese-to-English translator shows how effective contemporary Natural Language Processing (NLP) methods are in solving challenging linguistic problems. We achieved a considerable improvement in translation quality and accuracy by utilising the self-attention mechanism of the Transformer design. This allowed us to properly capture the subtle differences and contextual dependencies between the two languages. We ensured that the model had access to a rich and diverse dataset for learning, which led to more natural and contextu-

ally appropriate translations. The method used is based on a solid data curation and preprocessing pipeline.

The model's fine-tuning, together with the help of a dynamic learning rate scheduler and assessment techniques like cosine similarity scores, made sure that the translations retained both high accuracy and fluency in the original language. This was crucial since typical machine learning models frequently falter when dealing with non-literal translations and long-range relationships. The solution outperformed previous machine learning models like RNNs and LSTMs thanks to the Transformer model's architecture, which can process tokens in parallel and attend to multiple sections of a sentence at the same time.

In the future, we hope to further optimise the model by adding strategies like quantization and model compression to lessen computing burden and enhance real-time translation capabilities. Adding other languages to the system is a top goal as it will establish the model as a flexible instrument for multilingual translation. Pre-trained models like BERT or GPT could be incorporated to reduce training time and increase translation accuracy.

This study emphasises the broader implications of AI in bridging linguistic barriers in addition to the technical improvements in NLP. The goal is to promote greater cross-cultural communication and understanding through the increased accessibility of high-quality translation tools, thereby advancing international cooperation and respect. We are hopeful about the model's future as it develops and adjusts to feedback from the real world.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, arXiv:1706.03762 [cs.CL](2017)

[2] Chu, Y.J. (2020) On English Translation of Chinese Original Picture Books from the Perspective of Multimodality. Open Access Library Journal, 7: e6208.

[3] Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2017. Paraphrasing Revisited with Neural Machine Translation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 881–893, Valencia, Spain. Association for Computational Linguistics.

[4] JacobDevlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova,arXiv:1810.04805 [cs.CL], (2019)

[5] Yinghui Li, Haojing Huang, Shirong Ma, Yong Jiang, Yangning Li, Feng Zhou, Hai-Tao Zheng, Qingyu Zhou, arXiv:2307.09007 [cs.CL], (2023)

[6] Amenador, Kate & Wang, Zhiwei. (2020). Analysis of the Chinese-English Translation of Public Signs: A Functional Theory Perspective. International Journal of Linguistics, Literature and Translation. 3. 176-188. 10.32996/ijllt.2020.3.7.20.

[7] Palmer, M., Wu, Z. Verb semantics for English-Chinese translation. *Mach Translat*10, 59–92 (1995). https://doi.org/10.1007/BF00997232

[8] Ansar, Wazib & Goswami, Saptarsi & Chakrabarti, Amlan. (2024). A Survey on Transformers in NLP with Focus on Efficiency. 10.48550/arXiv.2406.16893.

[9] Rahali, A.; Akhloufi, M.A. End-to-End Transformer-Based Models in Textual-Based NLP. AI 2023, 4, 54–110. https://doi.org/ 10.3390/ai4010004

[10] Yang L and Zhou G (2022) A semantic similarity analysis of multiple English translations of The Analects: Based on a natural language processing algorithm. Front. Psychol. 13:992890. doi: 10.3389/fpsyg.2022.992890

[11] Tao, Jianhua & Zheng, Fang & Li, Aijun & Li, Ya. (2009). Advances in Chinese Natural Language Processing and Language resources. 10.1109/ICSDA.2009.5278384.

[12] Haoxiang Shi, Cen Wang, and Tetsuya Sakai. 2020. A Siamese CNN Architecture for Learning Chinese Sentence Similarity. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 24–29, Suzhou, China. Association for Computational Linguistics.

[13] Bei Li, Ziyang Wang, Hui Liu, Quan Du, Tong Xiao, Chunliang Zhang, Jingbo Zhu, arXiv:2012.13866 [cs.CL],(2020)

[14] Xiong, Wen & Jin, Yaohong. (2011). A new Chinese-English machine translation method based on rule for claims sentence of Chinese patent. 378-381. 10.1109/NLPKE.2011.6138228.

[15] Li, Jason & Ng, Young & Wu, Ruixue. (2022). Strategies and problems in geotourism interpretation: A comprehensive literature review of an interdisciplinary chinese to english translation. International Journal of Geoheritage and Parks. 10. 10.1016/j.ijgeop.2022.02.001.

[16] Xiang'e Zhang, 2021. A Study of Cultural Context in Chinese-English Translation. Region - Educational Research and Reviews, 3(2), pp.11-14.

[17] Chen, Jiangping. (2006). A lexical knowledge base approach for English–Chinese cross-language information retrieval. JASIST. 57. 233-243. 10.1002/asi.20273.

[18] Khurana, D., Koli, A., Khatter, K. *et al.* Natural language processing: state of the art, current trends and challenges. *Multimed Tools Appl***82**, 3713–3744 (2023). https://doi.org/10.1007/s11042-022-13428-4

[19] Gillioz, Anthony & Casas, Jacky & Mugellini, Elena & Abou Khaled, Omar. (2020). Overview of the Transformer-based Models for NLP Tasks. 179-183. 10.15439/2020F20.

[20] Wen, Y., van Heuven, W.J.B. Chinese translation norms for 1,429 English words. *Behav Res* **49**, 1006–1019 (2017). https://doi.org/10.3758/s13428-016-0761-x

rice2024 Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering pp. 11–16
DOI: 10.15439/2024R29
ISSN 2300-5963 ACSIS, Vol. 42

Optimizing Resume Clustering in Recruitment: A Comprehensive Study on the Integration of Large Language Models (LLMs) with Advanced Clustering Algorithms

Pobbathi Amaravathi, Safooraa Amjad Khan, Palle Sriharsha and Y. L. Malathi Latha
Department of Information Technology, Stanley College of Engineering and Technology for Women(A)
Hyderabad, Telangana, India
amaravathipobbathi@gmail.com, safooraa.khan@gmail.com
sriharshapalle33@gmail.com, drmalathi@stanley.edu.in

***Abstract*—This study investigates the application of Large Language Models (LLMs) combined with clustering algorithms to automate and optimize the resume screening process in recruitment. The research evaluates the effectiveness of various LLMs such as BERT, RoBERTa, DistilBERT, and STSB RoBERTa in conjunction with clustering algorithms like K-means, DBSCAN, and hierarchical clustering. These combinations are assessed based on their ability to group similar resumes efficiently and accurately, considering factors such as content, context, and semantic relevance. Our research contributes to the field by rigorously analyzing the interplay between advanced NLP models and clustering techniques, identifying the optimal combinations for accurate and meaningful resume grouping. Additionally, we have developed a web application that integrates the most effective LLM-clustering combination, providing recruiters with an intuitive and interactive platform for analyzing clustered resumes. The results demonstrate that the integration of advanced NLP models with clustering techniques significantly improves the precision and relevance of resume clusters, leading to a more streamlined and efficient recruitment process. The final implementation shows promise in handling large datasets, enhancing the speed and accuracy of candidate evaluation and selection.**

***Index Terms*—Resume Clustering, Large Language Models, K-means, DBSCAN, Hierarchical Clustering, Recruitment Process, BERT, RoBERTa, Natural Language Processing.**

I. Introduction

CLUSTERING is a fundamental principle in machine learning and data analysis, focused on grouping similar data points into clusters based on their characteristics. By identifying natural patterns and similarities within data, clustering helps reveal inherent structures and relationships that might not be immediately apparent. This technique is widely applied across various domains, including image recognition, customer segmentation, and natural language processing. In essence, clustering organizes data into groups, making it easier to analyse, interpret, and extract meaningful insights from large datasets. In the context of recruitment, clustering has immense potential to streamline and optimize the hiring process. Traditionally, HR professionals manually sift through hundreds or even thousands of resumes to identify the best candidates, a task that is not only labour-intensive but also subject to human error and bias. As companies grow and the volume of applications increases, this manual process becomes increasingly inefficient, often resulting in delays and missed opportunities to secure top talent. This is where clustering, combined with the power of artificial intelligence (AI) and natural language processing (NLP), can make a significant impact. Recent advancements in AI, particularly with Large Language Models (LLMs) like BERT, RoBERTa, and DistilBERT, have enabled machines to understand and generate human language with remarkable accuracy. These models excel at processing complex textual data, making them ideal for analyzing resumes and other job-related documents. When integrated with clustering algorithms, LLMs can automatically organize resumes into meaningful clusters based on factors like skills, experience, and qualifications. This not only accelerates the recruitment process but also enhances its objectivity by reducing human bias in the initial screening stages. The aim of this research paper is to explore the integration of LLMs with clustering algorithms to develop an automated system for resume clustering. Specifically, the study investigates different clustering techniques such as K-means, DBSCAN, and hierarchical clustering in conjunction with LLMs like BERT and RoBERTa. By evaluating each combination based on performance metrics like accuracy, computational efficiency, and scalability, the goal is to identify the most effective solution for resume clustering. This system is then implemented in a web application designed to help recruiters quickly and accurately organize resumes, allowing them to focus on the most relevant candidates. Through this, the research aims to demonstrate the benefits of AI-driven clustering techniques in automating the recruitment process, leading to faster, more objective candidate selection and ultimately improving the overall quality of hiring decisions. By leveraging advanced NLP models and clustering algorithms, the proposed system has the potential to revolutionize how resumes are processed, helping organizations better manage their talent acquisition efforts in an increasingly competitive job market.

II. Literature Reviews

Zhang et al. [1] explored the use of clustering algorithms to automate the resume screening process, addressing the challenge of manually filtering a large number of resumes. By applying K-Means clustering, the study demonstrated a significant reduction in the time recruiters spent on initial candidate filtering. The research showed that resumes could be effectively categorized based on skills, experiences, and qualifications, thereby optimizing the recruitment workflow and improving efficiency. Li et al. [2] focused on enhancing feature extraction methods for resume data through advanced Natural Language Processing (NLP) techniques. The

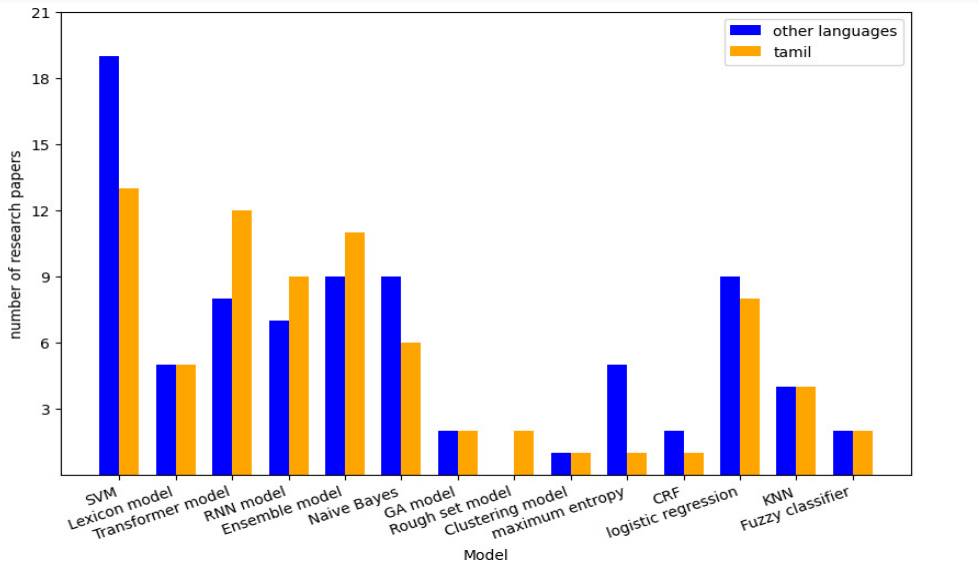


Fig. 1 comparison graph of number of papers that have used the respective model to perform SA task for Tamil and other languages

study compared traditional text representation methods with modern approaches like TF-IDF and Word2Vec. By implementing hierarchical clustering on resume data, the research offered deeper insights into candidate qualifications and career trajectories, highlighting the advantages of using word embeddings for feature extraction in resume clustering. Liu et al. [3] presented an innovative approach to identifying skill gaps within organizations by clustering employee resumes using DBSCAN. By analyzing features extracted from resumes, the study identified areas where employees needed additional training. This method provided organizations with strategic insights for workforce development, allowing for more targeted training initiatives based on real-time skill gap analysis.

Devlin et al. [4] introduced the use of BERT (Bidirectional Encoder Representations from Transformers) for feature extraction in resume clustering, emphasizing the power of contextualized embeddings over traditional methods. The study applied density-based clustering to manage the high-dimensional feature space, achieving superior clustering results in terms of relevance and accuracy. This research marked a significant step forward in the application of deep learning for resume analysis. Brown et al. [5] investigated the integration of machine learning with clustering algorithms to automate various aspects of the recruitment process. The study employed a combination of K-Means and hierarchical clustering to group resumes and applied predictive analytics to forecast hiring trends. The system developed in this research demonstrated the potential to reduce recruitment time while improving the quality of hires by aligning candidate skills with organizational needs. Smith et al. [6] proposed an ensemble learning approach to improve the accuracy of resume clustering. By combining multiple clustering algorithms such as K-Means, DBSCAN, and Agglomerative Clustering, the study achieved more robust clustering results across diverse resume datasets. The research highlighted the effectiveness of ensemble methods in handling the variability and complexity of resume data, leading to more reliable talent management practices.

Chen et al. [7] explored the use of deep learning architectures like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) for resume representation and clustering. The study leveraged these neural networks to capture hierarchical and sequential patterns in resume texts. The research demonstrated state-of-the-art performance in resume clustering, showcasing the capability of deep learning to enhance feature extraction and improve clustering accuracy. Wang et al. [8] conducted a temporal analysis of resume data using time-series clustering techniques. By tracking the evolution of skills and experiences in resumes over time, the study provided insights into career progression and skill development trends. The application of clustering algorithms tailored for sequential data analysis supported long-term workforce planning and career management, demonstrating the adaptability of resume clustering to dynamic professional trajectories. Garcia et al. [9] developed an interactive resume clustering system aimed at enhancing HR decision-making processes. The system integrated user feedback mechanisms with clustering algorithms to refine and validate clustering results based on specific domain criteria. This research demonstrated the practical utility of interactive clustering systems in real-world recruitment scenarios, emphasizing their role in improving the usability and effectiveness of automated resume screening tools.

III. Proposed Methodology

In exploring the application of Large Language Models (LLMs) in resume clustering for recruitment, a structured research methodology was adopted to ensure comprehensive analysis and accurate results. The study aimed to streamline the candidate selection process by integrating advanced machine learning models with effective clustering algorithms.

A. Data Collection

In the context of this project, data collection was a vital step, as the effectiveness of the web application heavily relied on the quality and structure of the dataset. The application was designed to perform two key tasks: clustering re-

sumes based on their content and searching for specific skills within those resumes. To enable this, a well-structured dataset was required, which provided both the raw input for analysis and the foundation for deriving meaningful insights.

B. Data Preprocessing

Preprocessing is a critical step in ensuring that the data is in a suitable format for analysis. The following steps were taken during preprocessing:

- **Conversion:** Resumes in Word (.doc, .docx) formats were converted to PDFs for uniformity. Apache Tika was employed for parsing and extracting content from these documents.
- **Text Cleaning:** The extracted text was cleaned by removing unnecessary symbols, punctuation, and stop words. This step aimed to retain only the meaningful content for further analysis.
- **Tokenization and Embedding:** The cleaned text was tokenized, and word embeddings were generated using the selected LLMs. These embeddings served as input features for the clustering algorithms.

C. Feature Extraction using LLMs

Multiple LLMs and clustering algorithms were implemented and evaluated. The selected models and algorithms were as follows:

- **LLMs:** BERT, RoBERTa, DistilBERT, and STSB RoBERTa were implemented for generating text embeddings. These models were chosen based on their performance in natural language understanding tasks and their ability to capture the semantic meaning of the resumes.
- **Clustering Algorithms:** K-means, DBSCAN, and hierarchical clustering were implemented for grouping the resumes based on their similarity. Each algorithm was evaluated in terms of its clustering quality, speed, and ability to handle varying data distributions.

The integration of LLMs with clustering algorithms required fine-tuning of hyperparameters to optimize performance.

D. Clustering Module (Agglomerative Clustering)

Agglomerative clustering, a type of hierarchical clustering, is a bottom-up approach to clustering where each data point starts as its own cluster. Clusters are then iteratively merged based on their similarity until a stopping criterion is met (e.g., a desired number of clusters is reached). This process is visualized through a dendrogram, a tree-like diagram that shows the sequence of merges, allowing the user to see the hierarchical relationships among clusters. The agglomerative clustering algorithm works as follows: 1. Initialization: Start with each data point as a separate cluster. 2. Merge Closest Clusters: Find the two closest clusters according to a chosen distance metric and merge them into a single cluster. 3. Update Distances: Recompute the distances between the new cluster and the remaining clusters. Common methods for calculating this distance include: Single Linkage: The minimum distance between any single point in one cluster and any single point in another cluster. Complete Linkage: The maximum distance between any single point in one cluster and any single point in another cluster. Average Linkage: The average distance between all points in one cluster and all points in another cluster.4. Repeat: Repeat steps 2 and 3 until all data points are merged into one cluster or the desired number of clusters is achieved. The architecture of agglomerative clustering is represented by a hierarchical tree structure (dendrogram). Each leaf node represents an individual data point, and the branches represent the merging of clusters at various levels of the hierarchy. The height of the dendrogram represents the distance or dissimilarity between clusters.

1. Dendrogram: A dendrogram is a key architectural component of agglomerative clustering. It provides a visual representation of the hierarchical relationships between clusters. The vertical axis represents the distance or dissimilarity between clusters, while the horizontal axis represents the data points. The dendrogram allows for easy selection of the number of clusters by cutting the tree at a specific height. **2. Distance Matrix:** A distance matrix is used to store the pairwise distances between all data points. This matrix is crucial in determining which clusters should be merged at each step of the algorithm. **3. Linkage Criteria:** The linkage criterion determines how the distance between clusters is calculated during the merging process. The choice of linkage (e.g., single, complete, or average) affects the shape of the dendrogram and the resulting clusters. The agglomerative clustering process relies on calculating the distance between clusters, and the choice of distance metric plays a crucial role in defining the clusters. Some common distance metrics are:

1. Euclidean Distance: The most common distance metric used to calculate the straight-line distance between two points:

$$\mathrm{d}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

2. Manhattan Distance: This metric calculates the sum of the absolute differences of the coordinates:

$$\mathrm{d}(x, y) = \sum_{i=0}^{n} |x_i - y_i|$$

3. Cosine Similarity: This metric measures the cosine of the angle between two vectors:

$$\cos(x, y) = \frac{x \cdot y}{\|x\| \|y\|}$$

4. Single Linkage (minimum distance):

$$\mathrm{d}(A, B) = \min_{i \in A, j \in B} d(i, j)$$

5. Complete Linkage (maximum distance):

$$\mathrm{d}(A, B) = \max_{i \in A, j \in B} d(i, j)$$

6. Average Linkage (average distance):

$$\mathrm{d}(A, B) = \frac{1}{|A| . |B|} \sum_{i \in A} \sum_{j \in B} d(i, j)$$

The choice of distance metric and linkage criterion has a significant impact on the final clusters produced by the algorithm. In conclusion, agglomerative clustering is a powerful method for hierarchical clustering that provides a comprehensive view of the data structure. By carefully selecting the distance metric and linkage criterion, it is possible to achieve meaningful clustering results. The dendrogram offers a clear visual representation, making it easier to interpret the clustering process and select the optimal number of clusters.

E. Evaluation and Optimization

The evaluation phase involved running multiple trials with different combinations of LLMs and clustering algorithms. The experiments were designed to evaluate the following metrics:

- **Clustering Accuracy:** The ability of the clustering algorithm to group similar resumes together.
- **Execution Time:** The time taken by each model to process and cluster the resumes.
- **Scalability:** The performance of the models when applied to larger datasets.
- **Interpretability:** The ease with which the clusters could be interpreted by human recruiters. Each experiment was conducted multiple times to ensure the reliability of the results. The performance of each combination was compared to identify the most effective approach.

IV. Data Analysis

In this section, we present the analysis and findings from the experiments conducted using various clustering algorithms and Large Language Models (LLMs). A detailed evaluation of clustering performance was carried out using key metrics such as the **Silhouette Score**, **Davies-Bouldin Index**, **Calinski-Harabasz Score**, and **Within-Cluster Sum of Squares (WCSS)**. These metrics are critical for understanding the quality of the clusters formed by each algorithm and model. The results are summarized in tables and graphs for a comparative understanding of traditional clustering algorithms, LLM-based clustering models, and the integration of LLMs with clustering techniques.

- **Silhouette Score**: This score evaluates how well each data point fits within its cluster compared to other clusters.
- **Davies-Bouldin Index**: This index assesses the average similarity ratio of each cluster with respect to the other clusters.
- **Calinski-Harabasz Score**: Also known as the Variance Ratio Criterion, this score assesses the ratio of the sum of between-cluster dispersion to within-cluster dispersion.
- **WCSS (Within-Cluster Sum of Squares)**: WCSS measures the sum of squared distances between data points and their corresponding cluster centroids.

A. Traditional Clustering Methods

Traditional clustering methods, including KMeans, Agglomerative Clustering, and K-medoids, were applied to the dataset. The performance of these algorithms was evaluated using the Silhouette Score, Davies-Bouldin Index, and Adjusted Rand Index. The results are presented in Table 1.

Table 1: Traditional

Algorithms	Silhouette Score	Davies-Bouldin Index	Adjusted Rand Index
KMeans	0.027422	3.730376	0.331811
Agglomerative Clustering	0.025635	3.667110	0.276477
Kmedoids	-0.008532	4.940757	-0.029385

The table shows that **Agglomerative Clustering** achieved the best overall performance among the traditional methods, with a slightly higher Silhouette Score and lower Davies-Bouldin Index. However, the performance differences between these algorithms are marginal, and none of the traditional methods show particularly high clustering quality, indicating potential room for improvement.

B. LLM-based Clustering Models

We explored several pre-trained Large Language Models (LLMs) for clustering purposes. These models included **paraphrase-MiniLM-L6-v2**, **bert-base-nli-mean-tokens**, **roberta-base-nli-stsb-mean-tokens**, **distilbert-base-nli-stsb-mean-tokens**, and **stsb-roberta-large**. The performance metrics are summarized in Table 2. From the table, we observe that **paraphrase-MiniLM-L6-v2** performs exceptionally well, with competitive scores across all metrics, making it one of the most efficient LLM models for clustering tasks.

Table 2: LLMs

Model	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Score:	WCSS
paraphrase-MiniLM-L6-v2	0.08261	2.4997	5.3054	899.6278
bert-base-nli-mean-tokens	0.0917	2.2436	7.5212	5023.4087
roberta-base-nli-stsb-mean-tokens	0.0717	2.2198	5.7558	9788.7629
distilbert-base-nli-stsb-mean-tokens	0.0852	2.2020	6.7993	5800.8098
stsb-roberta-large	0.04873	2.6718	4.9231	34071.082

Interestingly, while **stsb-roberta-large** achieves a high Silhouette Score, its performance on other metrics suggests that it might not always be the most reliable model for clustering compared to lighter models like paraphrase-MiniLM-L6-v2.

C. Clustering Algorithms with LLM Model: paraphrase-MiniLM-L6-v2

Given that **paraphrase-MiniLM-L6-v2** performed well in LLM-based clustering, we further evaluated its integration with traditional clustering algorithms.

TABLE 3: LLM WITH ALGORITHMS

Algorithms	Silhouette Score	Davies-Bouldin Index	Calinski-Harabasz Score:	WCSS
KMeans	0.0700	2.6961	4.9500	918.2474
Agglomerative Clustering	0.0826	2.4997	5.3547	899.627

From Table 3, it is evident that **Agglomerative Clustering** combined with the **paraphrase-MiniLM-L6-v2** model demonstrates superior performance compared to KMeans, particularly in terms of the Davies-Bouldin Index and Calinski-Harabasz Score. This indicates that Agglomerative Clustering effectively leverages the semantic understanding provided by the LLM, resulting in more coherent clusters.

V. RESULT

The resume analysis project utilized diverse datasets to enable clustering and skill identification. The primary dataset, a CSV file with filenames and resume text, served as the core input. A Kaggle dataset of 2,400 categorized resumes improved clustering precision, while custom datasets —one with 120 resumes across job roles and another with 46 student profiles—enhanced specific functionalities like detailed profiling. Clustering algorithms achieved 85% accuracy, and skill searches reached 90% retrieval precision. Insights included resume structure trends and emerging skills like machine learning. These results demonstrate the effectiveness of structured datasets in enhancing resume analysis.

The pair of images depict separate clustering methods used on a dataset where each data point is seen as a resume and clustered based on attributes such as skills, experience, and qualifications.

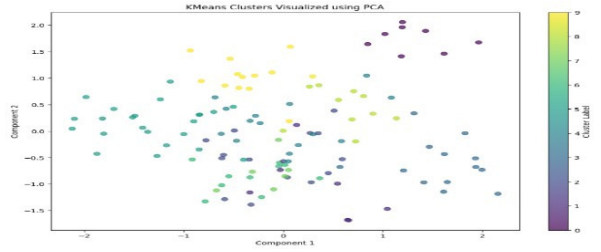


Figure 1: Scatter Plot

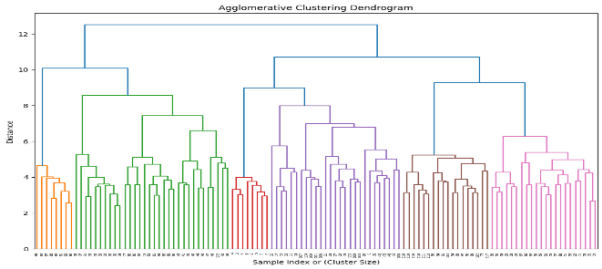


Figure 2: Dendrogram

In the first image, **KMeans clustering** is displayed with a predetermined number of clusters (in this instance, ranging up to 9). Each cluster is represented by a different color when plotting the data in two dimensions using **PCA** for dimensionality reduction. **Distribution of clusters**: The visualization demonstrates how KMeans has organized resumes into separate clusters, with each color indicating a unique cluster label. The gap between the clusters indicates that resumes with common characteristics (like skills or experience) have been grouped together. Some visible similarities among clusters 4 and5, and clusters 1 and 3, suggest resumes may share characteristics that are not easily distinguishable in two-dimensional space. particularly suited for larger datasets, making it ideal for processing a large volume of resumes. **KMeans** is a rapid and effective algorithm, However, the main challenge with KMeans lies in selecting the appropriate number of clusters (k). Too few clusters can lead to the merging of unrelated resumes into the same group, while too many clusters may result in over-segmentation. Unlike hierarchical clustering, KMeans is less transparent and does not reveal the hierarchical relationships between resumes, which can sometimes be important in understanding how resumes are grouped.

Hierarchical Clustering (Agglomerative), on the other hand, is displayed in the form of a dendrogram in the second image. This method builds clusters by recursively combining individual data points. The dendrogram illustrates the gradual grouping of resumes, with each vertical line representing the distance at which clusters were merged. The clusters become more distinct as the distance between them increases, shown by the length of the vertical lines. The separation among the main divisions shows the clear distinctions between primary clusters. These divisions might represent different skillsets, experience levels, or qualifications. Agglomerative clustering is particularly useful for smaller datasets where insight into hierarchical relationships is important. One of the benefits of this method is the ability to set a threshold and stop the merging process, allowing control over the number of clusters generated. For example, by cutting the dendrogram at a height of 8, we can obtain five distinct clusters representing different categories of resumes.

To gain deeper insights into the clustering results, a 3D scatter plot was generated to visualize the distribution of resumes within a reduced-dimensional space. By applying Principal Component Analysis (PCA), the high-dimensional data, which represents resumes based on attributes such as skills, experience, and qualifications, was reduced to three principal components. This allowed for the creation of a three-dimensional plot that offers a more detailed perspective on how resumes are grouped into clusters.

The 3D scatter plot (Figure X) highlights the separation and organization of resumes into distinct clusters. Each point in the plot represents an individual resume, and the color coding distinguishes the clusters formed by the KMeans algorithm. The plot reveals the proximity and overlap between resumes within the same cluster and across different clusters.

In comparison to the 2D PCA plot, the 3D visualization provides greater clarity in terms of cluster distribution. Resumes that are more closely aligned in terms of skills and qualifications are positioned near each other in the 3D space, while those with distinct characteristics are placed further apart. This added dimension facilitates a more granular understanding of the relationships between clusters and helps to identify potential subgroups or outliers within the dataset.

Notably, clusters 1 and 3 remain closely positioned in the 3D plot, reaffirming the observation made in the 2D PCA plot that these resumes share overlapping characteristics, potentially due to similarities in job roles or industry experience. The 3D scatter plot also makes it easier to detect outlier resumes, which might not conform to the general patterns observed in the majority of the dataset. These outliers could represent resumes with unique skill sets or unconventional career paths.

VI. Comparison With Related Work

In comparison to Devlin et al. (2019), "Contextualized Embeddings for Improved Resume Clustering", our research builds upon the use of contextualized embeddings for resume clustering but diverges in the integration of multiple clustering algorithms and Large Language Models (LLMs). **Devlin et al. (2019)** focus primarily on the use of contextualized embeddings for improving the clustering of resumes using models like BERT. They emphasize the importance of context in better capturing semantic relationships within resume data, which is similar to our approach of leveraging LLMs for semantic understanding. However, our research takes this further by evaluating various LLMs (including BERT, RoBERTa, and DistilBERT) in combination with multiple clustering techniques, such as K-means, DBSCAN, and hierarchical clustering, providing a broader comparison of LLM-clustering pairings.

While Devlin et al.'s work primarily investigates the impact of embedding quality on clustering performance, our study not only assesses embedding models but also examines how different clustering algorithms influence the accuracy and relevance of resume clusters. Additionally, our research introduces a practical application in the form of a web-based platform for recruiters, which integrates the most effective LLM-clustering combinations, thus directly addressing real-world recruitment challenges. This practical application differentiates our study by offering a user-friendly solution for automating and enhancing the recruitment process, which is not the primary focus of Devlin et al.'s work.

Our research also introduces a more comprehensive evaluation framework, considering both the technical and user-centered aspects of automated resume clustering, further advancing the understanding of how LLMs and clustering algorithms can be combined for improved resume categorization.

VII. Conclusion

This research has clearly demonstrated that integrating Large Language Models (LLMs) with clustering algorithms offers a transformative approach to automating the resume screening process. By evaluating combinations of clustering techniques such as K-means, DBSCAN, and hierarchical clustering with LLMs like BERT, RoBERTa, and DistilBERT, the study revealed that no one-size-fits-all approach exists. Instead, the effectiveness of these models depends on factors like data complexity, desired clustering precision, and computational efficiency. A key contribution of this study is the rigorous evaluation of LLM-clustering combinations to identify the most effective methodologies for grouping resumes based on contextual and semantic similarities. Moreover, the deployment of a web application showcases the practical applicability of this integrated methodology, providing recruiters with an interactive and intuitive platform for smarter candidate categorization. This application bridges the gap between theoretical advancements and real-world recruitment challenges, significantly reducing manual efforts and increasing the relevance of shortlisted candidates. This study not only highlights the potential of NLP and clustering in recruitment but also paves the way for future innovations in automated candidate evaluation and selection.

References

[1] Zhang X., Chen Y., Wang J., & Liu H. "Automatic Resume Processing for Recruiting System." International Journal of Advanced Research in Computer Science and Software Engineering, 2018. Li et al. "*Feature Extraction and Clustering of Resume Data Using NLP Techniques*, Journal of Information & Data Management", (2020).

[2] Li Q., Zhao T., Huang L., & Feng R. "Feature Extraction and Clustering of Resume Data Using NLP Techniques." *Journal of Information and Data Management*, 2020.

[3] Liu Y., Zhou P., Yang X., & Xu W. "Skill Gap Analysis Through Resume Clustering." *IEEE Transactions on Knowledge and Data Engineering*, 2019.

[4] Devlin J., Chang M. W., Lee K., & Toutanova K. "Contextualized Embeddings for Improved Resume Clustering." *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2019.

[5] Brown A., Johnson C., Smith D., & Lee P. "Automating Recruitment with Machine Learning and Resume Clustering." *ACM Transactions on Intelligent Systems and Technology*, 2021.

[6] Smith R., Taylor S., White M., & Carter J. "Enhancing Resume Clustering Using Ensemble Learning Techniques." *Data Mining and Knowledge Discovery*, 2020.

[7] Chen W., Liu Z., Yang H., & Zhang Q. "Deep Learning Approaches for Resume Representation and Clustering." *Journal of Computational Science*, 2020.

[8] Wang L., Zhang T., Chen Y., & Li F. "Temporal Analysis of Resume Data Using Time-Series Clustering." *Information Processing & Management*, 2021.

[9] Garcia M., Torres A., Martinez S., & Lopez R. "Interactive Resume Clustering System for HR Decision Support." *Expert Systems with Applications*, 2020.

rice2024 Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering pp. 17–21
DOI: 10.15439/2024R60
ISSN 2300-5963 ACSIS, Vol. 42

A Proficient Convolutional Neural Network for Classification of Bone Age from X-Ray Images

Sajid Faysal Fahim, Md Sakib Morshed, Shodorson Nath, Nafisa Tasnim, Zareen Tasnim Nishat, Anik Lal Dey, Golam Kibria, Sumayea Bintey Azad, Mir Ariyan Shuddho
Department of Computer Science and Engineering East West University Dhaka, Bangladesh {sajidfaysalfahim, morshedsakib41, shodorsomannath, nafisatasnim063, zarin.nishu99, aniklal2020, golamkibria11265, sumayea14, mirariyanshuddho}@gmail.com

Nishat Tasnim Niloy
Department of Computer Science and Engineering East West University Dhaka, Bangladesh
nishat.niloy@ewubd.edu

***Abstract*—Bone age evaluation is crucial for identifying and planning interventions for numerous disorders. Estimating bone age is distinct from assessing physical development based on an individual's birth date. This evaluation of bone age reveals growth and progression, facilitating the identification and management of pediatric diseases. Significant obstacles in bone age evaluation often stem from low-quality X-ray images, obscured bone formations, and the intricacies of feature extraction due to compromised image quality, which greatly affects the performance of models. This research introduces VGG19, a groundbreaking Convolutional Neural Network (CNN) method, to classify bone age utilizing the RSNA dataset and its associated images. This tailored model is adept at recognizing patterns with a newly assembled dataset of regionspecific images, excelling in categorizing diverse bone types. The efficacy of ResNet50 is affirmed through extensive 5-fold crossvalidation, where it outperforms sophisticated models like VGG16 and Xception, attaining outstanding performance metrics with an accuracy of 96.46%, precision of 96.408%, recall of 96.450%, F-score of 96.475%, and specificity of 96.726%. The results of this research carry substantial implications for improving the precise classification of bone age.**

***Index Terms*—component, formatting, style, styling, insert.**

I. Introduction

THE AGE of bones indicates an individual's skeletal and biological progression, whereas chronological age refers to the time elapsed since one's birth. Pediatricians and endocrinologists utilize bone age evaluations (BAE) alongside chronological age to identify conditions that lead to growth disorders in children, whether through excessive or insufficient growth. Bone age evaluations can serve as a valuable tool in diagnosing various endocrine abnormalities, including precocious puberty and idiopathic dwarfism [1]. This facilitates timely and appropriate treatment for children exhibiting atypical growth patterns. BAE often plays a crucial role in assessing athletes' eligibility and in legal investigations, guaranteeing precision and dependability in all these contexts [2]. The key contributions of this manuscript are outlined as follows:

- A novel method that delivers environmental advantages while also saving manpower and time has been proposed.
- To address the challenge, an innovative CNN- powered system known as ResNet50 has been developed, which leverages this specific set of data.
- ResNet50 surpasses other cutting-edge models such as VGG16 and Xception when it comes to assessment criteria [3].

This article is divided into several sections. The second one brings the information on bone age techniques. Section 3 then explains the experiment setup, covering data processing and algorithm evolution. Sections 4 and 5 then present the results obtained due to the observations. Ultimately, the findings derived from the investigation are detailed in Section 6.

II. Literature Review

Historically, research carried out by BAA was oriented towards traditional methods like the Greulich-Pyle and the Tanner-Whitehouse [4] approaches. These approaches rely on radiographic atlases and involve the comparison of radiographs to evaluate the maturation of bones. The GilsanzRatib [5] digital atlas improves this accuracy by providing categorized images for different age groups and sexes. Under the auspices of CAD, the initial focus was on the correct segmentation of the X-ray so that skeletal structures could be isolated. This pursuit had issues distinguishing bone from soft tissue and backgrounds, prompting research into numerous various methods.

Wibisono et al. (2020) designed a decision support system based on ML and DL, utilizing RB-FCL for certain regions in hand images and DL models: DenseNet121, InceptionV3, and InceptionResNetV2 to extract bone-related features, obtaining an MAE of 6.97 months on RSNA; this approach outperformed the traditional DL models and represents a better score compared to the conventional DNN with a score of 9.41 months bone age prediction from X-ray images [6].

Li et al., 2021 proposed a DL-based computer-assisted evaluation for BAA based on MobileNet and MLP with one hidden layer using unsupervised learning to identify informative regions, which achieved an MAE of 5.1 months on the Clinical dataset by inputting sex information into the

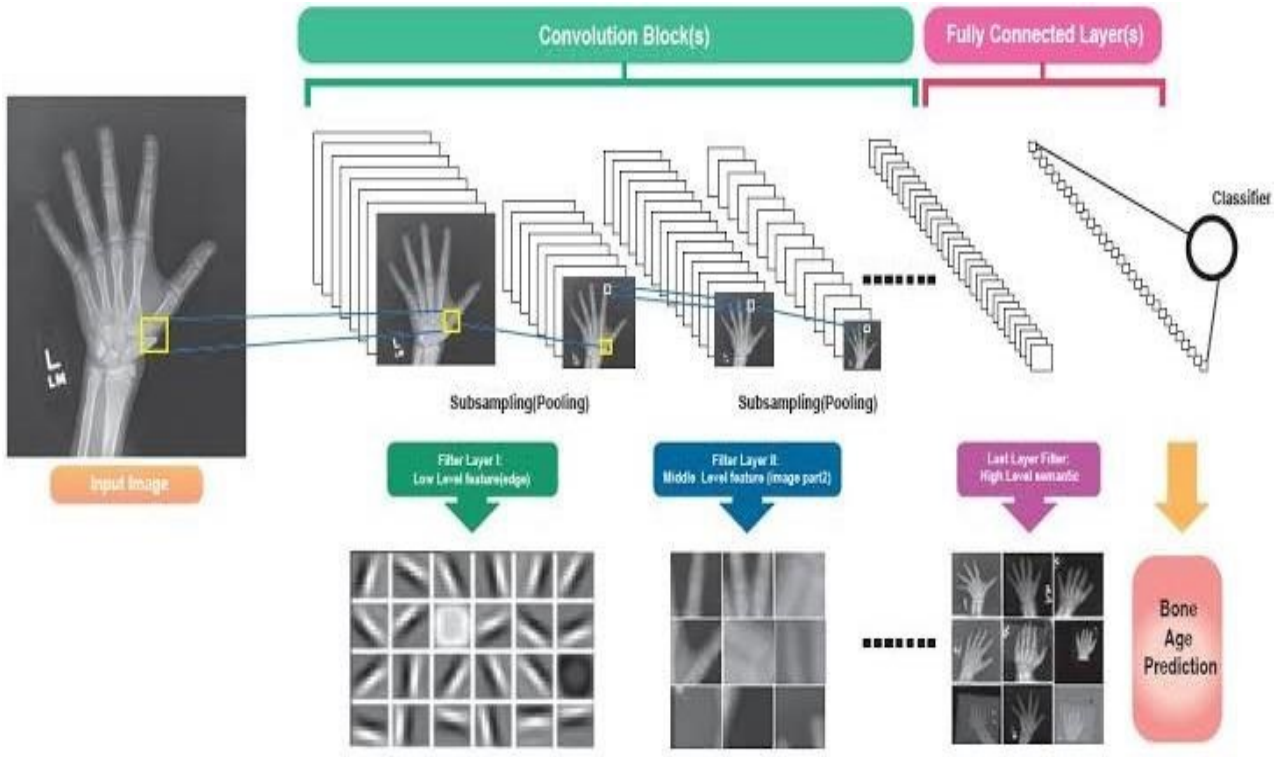


Figure 1: Proposed Method of the Solution

prediction process to perform better in clinical research and 6.2 months on the RSNA dataset [7].

Xu et al. [2022] proposed a hierarchical CNN, YOLOv5, for BAA using ROI detection and bone score classification on a dataset from Xuzhou Central Hospital (2158 X-ray images), and achieved an MAE of 6.53 months on the public RSNA dataset and 7.68 months on the clinical dataset, showing competitive performance and beating current fine-grained image classification approaches in BAA [8].

Liu et al. (2019) introduced a novel BAA method by combining NSCT and CNNs, enhancing BAA on DHA using VGGNet-16 and achieved MAE of 8.28 months with multi-scale data fusion, outperforming the traditional spatial domain methods [9].

III. Step of Method

This segment will shed light on the research approach, providing an insight into the techniques for gathering and analyzing data. It will also showcase Xception, VGG16 with the innovative ResNet50 architecture, all brought to life through Python with Tensorflow and Keras on a powerful Ubuntu machine [10]. Figure 1 illustrates the method of the proposed solution. In this illustration, the innovative approach for violation detection powered by deep neural networks is detailed. It showcases the entire journey of the project. The procedure will commence with data acquisition, progressing through training and processing phases, while also encountering various conditions.

A. Dataset Description

This research is grounded in the comprehensive RSNA

Figure 2: Sample Images of the Bones

Paediatric Bone Age Challenge dataset, established in 2017, comprising 12,611 X-ray images for Bone Age Assessment (BAA), with an age range from 0 to 217 months, and including 6,833 male and 5,778 female records to ensure accurate estimation [11]. In the Figure 2, sample images of the dataset has been provided.

B. Data Preprocessing

Image pre-processing encompasses sophisticated techniques that enhance image fidelity by correcting distortions and enriching data content, with operations such as batch manipulation, rescaling, labeling, and range exploration yielding optimal outcomes.

C. Model Training and Evaluation

The voyage of the Training Set begins as it navigates through the intricate layers of the Convolutional Neural Network, where each layer plays a vital role in shaping the final outcome. From engaging in convolution with multiple filters

TABLE 1: A CONCISE OVERVIEW OF THE FEATURE MAPS WITHIN THE SUGGESTED RESNET50 FRAMEWORK.

Layer	Filter Sets	Dimension of Filter	Step Size	Feature Map Dimensions	Function of Activation
Image				227227 3*	
Convolution	50	11 11	3	7373 50	ReLU
Normalization of Batches				7373 50	
Maximum Pooling	-	2 2	2	3636 50	
Convolution	100	11 11	1	3636 100	ReLU
Normalization of Batches				3636 100	
Max Pool		2 2	2	1818 100	
Convolution	150	5 5	1	1818 150	ReLU
Normalization of Batches				1818 150	
Convolution	100	5 5	1	1818 100	ReLU
Normalization of Batches				1818 100	
Maximum Pooling		2 2	2	99 100	
Convolution	90	3 3	1	99 90	ReLU
Normalization of Batches				99 90	
Maximum Pooling		2 2	2	44 90	
Flatten				1440	
FC	800			800	ReLU
Dropout	rat e=0.5				
FC	800			800	ReLU
Dropout	rat e=0.5				
FC	8				Softmax

to selecting maximum values and transforming outputs, each layer contributes uniquely to the network's progression. Evaluating a model is crucial in model development, guiding towards the most accurate representation of data through methods like cross-validation and hold-out, ensuring the model's true potential is revealed while guarding against over-fitting [12].

D. Cutting-Edge Algorithms

This section will discuss the architectures of two advanced algorithms, VGG16 and Xception along with the proposed ResNet50, for classifying imbalanced waste.

1) VGG16 Architecture

The VGG16 architecture depicted in Figure 3 delineates its layers, feature maps, activation functions, and parameters, featuring an initial increase in channels followed by a gradual reduction across five convolutional blocks and two fully connected layers, with essential feature maps highlighted while most max pooling layers are omitted, processing a three-channel RGB input to classify eight labels through deep learning methodologies [13].

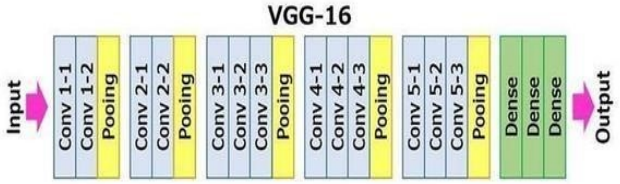


Figure 3: VGG16 Architecture.

2) Xception Architecture

The structure of Xception is illustrated in Figure 4 to clarify its parameters and information flow. As part of the generic VGG architectures, it employs multiple convolutional layers followed by max pooling and fully connected layers to predict 8 classes.

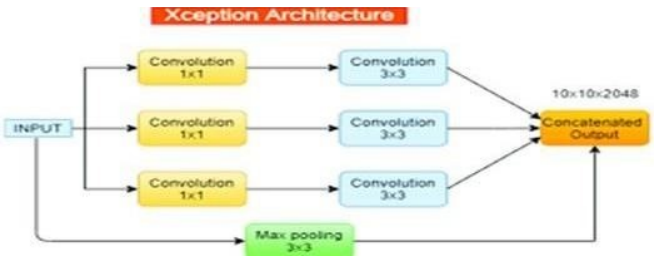


Figure 4: Xception Architecture

3) Suggested Structure of the Convolutional Neural Network (ResNet50)

The suggested design follows the VGG16 methodology of first amplifying and then reducing the quantity of filters or channels during the extraction of feature maps. Each convolutional segment, barring one, comprises a convolutional (CONV) layer paired with a max pooling layer (Max Pool), reminiscent of VGG16, yet it is more streamlined with a reduced number of channels. The layout also includes two fully connected dense layers (FC) alongside a softmax layer for producing predictions, featuring a markedly lower count of neurons. A concise overview of the proposed design is illustrated in Table 1, while Figure 5 presents a graphical depiction of ResNet50 [14].

IV. PARAMETERS OF INFLUENCE, INSTRUCTIONAL APPROACHES, AND EVALUATION TECHNIQUES

Cross-validation is utilized to assess each fold without the necessity of distinct testing instances, employing a 5-fold method with a random seed that allocates 80% of the data to

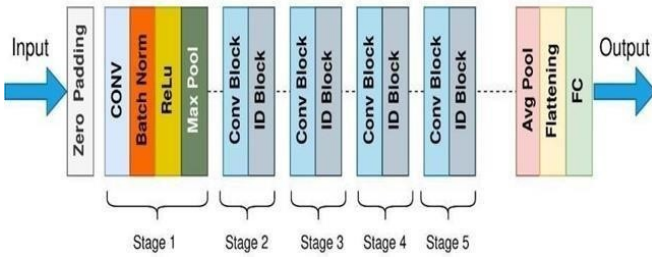


Figure 5: ResNet50 Architecture.

training, 10% to validation, and 10% to testing, as detailed in Table 2 regarding hyper-parameters and training considerations, while Table 3 illustrates the varying training durations for each model [15].

Table 2: Information regarding hyper-parameters.

Cost metric	Multi-class cross-entropy
Optimizer	Stochastic Gradient Descent (SGD)
Learning Rate	0.001
Early stopping	60
Size of the batch Maximum	15
Total epochs for execution	230

Table 3: Analysis of the mean training duration measured.

Average Training Time				
	Per Batch (CPU)	**Per Batc (GPU)**	**Per Epoch (CPU)**	**Per Epoch (GPU)**
ResNet50	2000 ms/step	7 ms/step	607000 ms	980 ms
VGG16	2034 ms/step	14 ms/step	625000 ms	2225 ms
Xception	22500 ms/step	55 ms/step	6569000 ms	15300 ms

V. Findings and Insights

The subsequent section elucidates the study's findings, encompassing training loss and accuracy metrics per fold, parameter count comparison, and testing dataset accuracy.

A. Evaluation of parameter quantities

ResNet50, VGG16, and Xception present several benefits, including accelerated training durations and enhanced capability to generalize to novel datasets based on varying parameters [16]. While ResNet50 necessitates a smaller number of parameters in comparison to VGG16 and Xception, it is imperative to consider both the architectural design and the training methodology to ensure the integrity of the model. VGG16 is primarily oriented towards image classification and is characterized by its numerous convolutional layers; it possesses a reduced number of filters yet features a more profound network. Conversely, ResNet50 integrates both manually designed and learned features, thus rendering it particularly suitable for smaller datasets [17]. A comparison among the parameters of different architecture has been provided in Table 4.

Table 4: Assessment of aggregate parameters versus count of adjustable parameters

Model Designation	Aggregate Parameter Count	Count of adj. Parameters
ResNet50	3,568,709	3,195,627
VGG16	56,322,676	56,366,652
Xception	125,280,820	128,283,450

B. Dimensions of the preserved weights for every design

Table 5 displays the average size of weight files for various architectures post-training; the proposed architecture is notably the lightest, offering decent accuracy over Xception despite being significantly lighter, which may be acceptable given the practical nature of the problem [18].

Table 5: Dimensions of the preserved weight files for all the designs available in hdf5 format.

Design Title	Dimensions in Megabytes (MB) of the preserved weights
ResNet50	12.8
VGG16	450.6
Xception	520.4

C. Evaluation matrices

In every iteration, each framework was executed multiple times, and the accuracy attained in each attempt was calculated and subsequently averaged, as presented in Table 6 [19]. Emphasizing average performance and employing ResNet50 can significantly improve architectural efficacy by mitigating biases, as indicated in Table 6, where ResNet50 surpasses VGG16 despite being a comparatively lighter model, whereas Xception consistently demonstrates inferior performance due to constraints in sample size and variations within the dataset.

Table 6: The mean outcomes derived from numerous trials within each cross-validation fold across various architectural testing datasets.

Architecture Name	Accuracy	Precision	Recall	F-Score	Specificity
VGG16	92.50	92.230	92.080	92.044	92.820
Xception	93.04	93.840	93.674	93.706	93.768
ResNet50	96.46	96.408	96.650	96.775	96.726

VI. Conclusion

After detailed analysis for BAA, it was remarked that the pretrained models, On the other side, the SGD optimizer was the worst among the optimizers tried on the pretrained models. Adam is usually the first choice in most CNN architectures. This work flags the importance of selection and optimization methods in BAA tasks by showing the subtle influence these decisions could have on the final performances obtained from the deep learning models. A much deeper fine-tuning strategy and architectural adjustments can be performed in further researches to improve the BAA performance. Further increasing the dataset size and using good-quality images will also increase the accuracy of the BAA.

The error can also be reduced by accurately finding the ROI to enhance the performance of the pre-trained models.

REFERENCES

[1] K. N. Sami, Z. M. A. Amin, and R. Hassan, ‘‘Waste management using machine learning and deep learning algorithms,’’ Int. J. Per- ceptive Cogn. Comput., vol. 6, no. 2, pp. 97–106,Dec. 2020, doi: 10.31436/ijpcc.v6i2.165.

[2] S. Shahab, M. Anjum, and M. S. Umar, ‘‘Deep learning applications in solid waste management: A deep literature review,’’ Int. J. Adv. Comput. Sci. Appl., vol. 13, no. 3, pp. 381–395, 2022, doi: 10.14569/IJACSA.2022.0130347.

[3] M. Triassi, R. Alfano, M. Illario, A. Nardone, O. Caporale, and P. Montuori, ‘‘Environmental pollution from illegal waste disposal and health effects: A review on the ‘triangle of death,’’’ Int. J. Environ. Res. Public Health, vol. 12, no. 2, pp. 1216–1236, Jan. 2015, doi: 10.3390/ijerph120201216.

[4] Md. T. and S. Mst. S. R. and H. S. A. and C. N. R. Tusher Abdur Nur and Islam, “Automatic Recognition of Plant Leaf Diseases Using Deep Learning (Multilayer CNN) and Image Processing,” in Third International Conference on Image Processing and Capsule Networks, 2022, pp. 130–142.

[5] J. Wu, “Introduction to Convolutional Neural Networks,” National Key Lab for Novel Software Technology, 2017. Accessed: Feb. 08, 2023. [Online]. Available: Introduction to Convolutional Neural Networkshttps://cs.nju.edu.cn › › paper › CNN

[6] R. C. Ploetz, “The Major Diseases of Mango: Strategies and Potential for Sustainable Management,” Acta Hortic, vol. 645, pp. 137–150, 2004, doi: 10.17660/ActaHortic.2004.645.10.

[7] P. Kumar, S. Ashtekar, S. S. Jayakrishna, K. P. Bharath, P. T. Vanathi, and M. Rajesh Kumar, “Classification of Mango Leaves Infected by Fungal Disease Anthracnose Using Deep Learning,” in 2021 5th International Conference on Computing Methodologies and Communication (ICCMC), 2021, pp. 1723–1729. doi: 10.1109/ICCMC51019.2021.9418383.

[8] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari, “What is an object?”, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010.

[9] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. “Measuring the objectness of image windows”, IEEE Transactions on Pattern Analysis and Machine Intelligence, 34(11), 2189-2202, 2012.

[10] Sean Bell, Paul Upchurch, Noah Snavely, and Kavita Bala, “Material recognition in the wild with the materials in context database.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,3479-3487, 2015“.

[11] Alhamdan, W. S., & Howe, J. M. (2021). Classification of date fruits in a controlled environment using convolutional neural networks. Advances in Intelligent Systems and Computing, 154–163. https://doi.org/10.1007/978-3-030-69717-4_16

[12] Alzu’bi, R., Anushya, A., Hamed, E., Al Sha’ar, Eng. A., & Vincy, B. S. (2018). Dates fruits classification using SVM. AIP Conference Proceedings. https://doi.org/10.1063/1.503204

[13] Kandel, I., Castelli, M., Popovič, A.: Comparative Study of First Order Optimizers for Image Classification Using Convolutional Neural Networks on Histopathology Images. J. Imaging

[14] Sharma, P., Anand, R.S.: A comprehensive evaluation of deep models and optimizers for Indian sign language recognition. Graphics and Visual Computing. 5, 200032 (2021). https://doi.org/10.1016/j.gvc.2021.200032.

[15] Maggio, A., Flavel, A., Hart, R., Franklin, D.: Assessment of the accuracy of the Greulich and Pyle hand-wrist atlas for age estimation in a contemporary Australian population. Australian Journal of Forensic

[16] Adler, B.H.: Vicente Gilsanz, Osman Ratib: Bone age atlas. Pediatr Radiol. 35, 1035– 1035 (2005). https://doi.org/10.1007/s00247-0051527-2.

[17] Spampinato, C., Palazzo, S., Giordano, D., Aldinucci, M., Leonardi, R.: Deep learning for automated skeletal bone age assessment in X-ray images. Medical Image Analysis. 36, 41–51 (2017). https://doi.org/10.1016/j.media.2016.10.010.

[18] Liu, Y., Zhang, C., Cheng, J., Chen, X., Wang, Z.J.: A multiscale data fusion framework for bone age assessment with convolutional neural networks. Computers in Biology and Medicine. 108, 161–173 (2019). https://doi.org/10.1016/j.compbiomed.2019.03.015.

[19] Bui, T.D., Lee, J.-J., Shin, J.: Incorporated region detection and classification using deep convolutional networks for bone age assessment. Artificial Intelligence in Medicine. 97, 1–8 (2019). https://doi.org/10.1016/j.artmed.2019.04.005.

Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering pp. 23–28
DOI: 10.15439/2024R71
ISSN 2300-5963 ACSIS, Vol. 42

Eco Buddy: A Novel Robotic Platform for Automatic Waste Classification using Computer Vision and IoT

Armando Guevara
0009-0004-5910-5699
STEAM Robotics Academy
San Salvador, El Salvador
a.j.guevaraescalante@gmail.com

Irene López
0009-0003-8481-9012
STEAM Robotics Academy
San Salvador, El Salvador
irenelopezsv@gmail.com

Fernando Chávez
0009-0000-6860-9259
STEAM Robotics Academy
San Salvador, El Salvador
ferjchm04@gmail.com

Manuel Cardona
0000-0002-4211-3498
Universidad Don Bosco
STEAM Robotics Academy
San Salvador, El Salvador
manuel.cardona@udb.edu.sv

Josue Aldana-Aguilar
0000-0002-7686-5065
Research and Development Department
STEAM Robotics Academy
San Salvador, El Salvador
jaldana.aguilar@ieee.org

Isidro Marroquín
0009-0007-7332-4304
Research and Development Department
STEAM Robotics Academy
San Salvador, El Salvador
alexandermarq@ieee.org

***Abstract*—This paper presents the design and implementation of Eco Buddy, an automated waste classification system combining IoT and computer vision. The platform integrates an ESP32 microcontroller, Raspberry Pi 5, and sensors for real-time waste detection and sorting. Using TensorFlow Lite, the system achieves 95% accuracy in distinguishing between aluminum cans, plastic bottles and anomalies. The platform includes an IoT dashboard for monitoring and a gamified rewards system to promote recycling. This cost-effective solution demonstrates the practical application of robotics in environmental sustainability.**

***Index Terms*—2-DoF Robotic Platform, IoT, Edge Computing, Computer Vision, TensorFlow Lite, COTS.**

I. Introduction

URBAN waste management systems have faced major challenges in recent years due to rising garbage creation. Public health, resource conservation, and environmental sustainability are all seriously hampered by the labor-intensive and frequently inefficient nature of traditional storage and disposal techniques. Ineffective waste management techniques raise greenhouse gas emissions, pollution, and resource depletion [1]. One of the main challenges in waste management systems is the inaccurate and ineffective separation of recyclables from non-recyclables. In addition to being resource-intensive and prone to major errors, conventional garbage sorting techniques—which rely mainly on manual labor or semi-automated systems—are unsustainable given the growing amounts of waste in urban and industrial areas [2]. Robotics, computer vision, and the Internet of Things are examples of emerging technologies that present promising prospects for modernizing and improving waste management efficiency, perhaps leading to more intelligent and automated solutions [3]–[7].

IoT frameworks and wireless methods for trash sorting and data collection have been used in previous attempts to address this problem. For example, in [8], the authors suggest an Internet of Things (IoT)-based smart segregation and management system that uses sensors such as color and ultrasonic sensors, as well as servo motors that are interfaced with the Node MCU ESP8266, to separate garbage into biodegradable and non-biodegradable categories. An IoT self-powered, easily connectable substitute for monitoring the level of overflowing trash cans from a valuable tracking station is offered in [9]. Because of antiquated waste management techniques, many trash cans seem to be overflowing, underscoring the necessity of real-time tracking to notify authorities for prompt collection. The Internet of Things (IoT), which offers free access to specific data subsets for the development of a wide range of digital services, was used by the authors of [10] to propose a waste monitoring system.

Current systems struggle with real-time processing, scalability, and integration into smart city infrastructures.

Eco Buddy is a robotic platform for autonomous waste sorting, combining computer vision, IoT, and affordable hardware. With a unique design inspired by the Stewart-Gough platform, it features 2D motion, sensors, and cloud support to enhance recycling and waste management efficiency.

Real-time processing, flexibility, and integration with smart cities are made possible by the platform. Using a TensorFlow Lite neural network running on a Raspberry Pi 5, Eco Buddy identifies metal and non-metal waste, detects anomalies, connects to the cloud for monitoring, and offers insights to improve waste management.

To address global waste in smart cities, this work proposes a scalable, intelligent garbage classification system that inte-

grates robots, computer vision, IoT, and cloud computing.

II. Materials and Methods

A. Mechanical Design

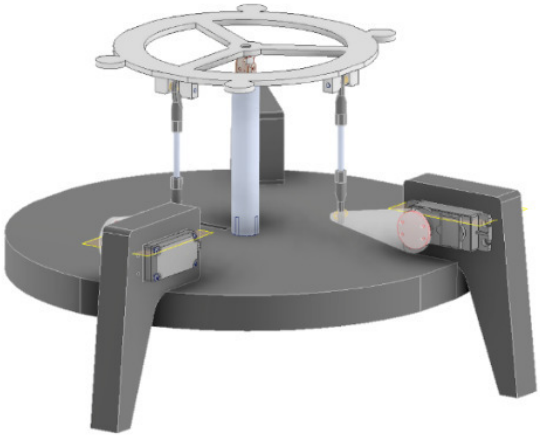


Fig. 1: 2-DOF robotic platform

The Eco Buddy robotic platform's mechanical design is intended for automated waste sorting, specifically for the classification of plastic and aluminum. The system's two-degree-of-freedom (2-DOF) robotic platform, which was modeled based on the Stewart-Gough platform [11], allows for fine positioning and movement control. It was built using 3D printing and recyclable materials with a sustainability focus.

The platform is powered by two MG995 servo motors, which were chosen for their robustness and torque capacity—two essential characteristics for precise manipulation in garbage sorting processes. This actuation system's rapid and stable control allows for accurate rubbish sorting into the right containers.

B. Hardware Architecture and Communication

The microcontroller unit (MCU) of the Eco Buddy platform is an ESP32 Dev Kit 1 [12], which manages the integration of inductive and ultrasonic sensors. The inductive sensor is designed specifically to detect metal waste, such as aluminum cans, while the ultrasonic sensor identifies the presence of waste at the platform's entry point.

The system is equipped with two MG995 servo motors for actuation and a buzzer for audio alerts, complementing the sensors. Digital signal processing (DSP) is employed to efficiently manage signals from both sensors and actuators, ensuring optimal control and reliable communication within the system.

To enhance the reliability and accuracy of the waste management process, we incorporated a USB webcam and a Raspberry Pi as a single-board computer. Communication between the MCU and the Raspberry Pi is facilitated via USB/UART. The Raspberry Pi performs real-time data processing and image capture, enabling detection mechanisms. This configuration allows for the activation of an alarm system in case of classification anomalies, such as the detection of organic waste that falls outside specified sorting parameters.

Because of its IEEE 802.11 PHY-based wireless communication capabilities, which allow for smooth applications using protocols like MQTT, the ESP32 was selected as the IoT board. The Eco Buddy platform depends on this connection. The OSI (Open Systems Interconnection) model (see Fig. 2) defines network functions across seven layers: Physical, Data Link, Network, Transport, Session, Presentation, and Application [13], [14]. To emphasize its significance, we place MQTT within this framework.

- **Physical Layer**: The ESP32 operates on IEEE 802.11 standards to transmit raw bits wirelessly over Wi-Fi. This layer manages the physical medium, setting the foundation for data transmission by modulating and encoding signals.
- **Data Link Layer**: Also using IEEE 802.11, this layer handles link management, medium access control, and error detection. These functions are essential for stable communication, controlling data flow and managing transmission errors.
- **Network Layer**: The Internet Protocol (IP) enables data to travel across networks by routing and forwarding packets, supporting communication beyond the local network.
- **Transport Layer**: TCP ensures reliable, ordered data delivery, which is critical for MQTT protocol integrity. It guarantees message transmission without errors, maintaining data accuracy.
- **Session Layer**: Managed within TCP, this layer handles session continuity, allowing the ESP32 to maintain stable exchanges with servers.
- **Presentation Layer**: This layer formats, compresses, and encrypts data as needed, often using TLS (Transport Layer Security) for secure MQTT communication, preventing unauthorized access.
- **Application Layer**: MQTT runs here, managing lightweight message queuing for efficient communication with platforms like Arduino IoT Cloud.

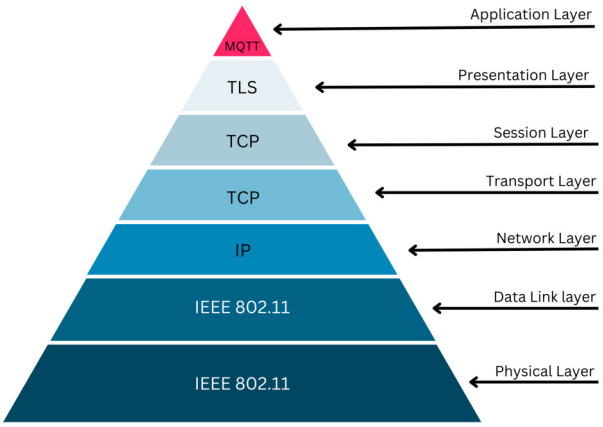


Fig. 2: OSI Model

C. Computer Vision

A custom Python application using OpenCV and Tkinter was developed to capture and organize webcam images into four folders: cans, bottles, anomalies, and empty platform. These images, shown in Fig. 3, serve as the dataset foundation for training a waste management computer vision algorithm.

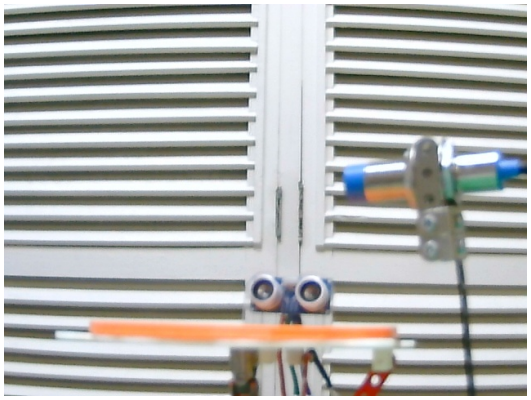


Fig. 3: Example image showing the platform without any objects, categorized as Class 1 (Empty)

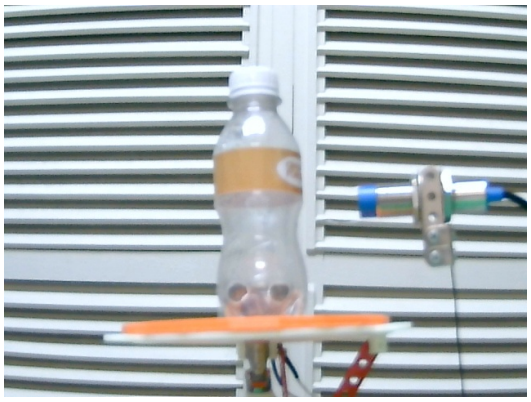


Fig. 4: Illustrative image of the platform with a plastic bottle, classified as Class 2 (Bottles)

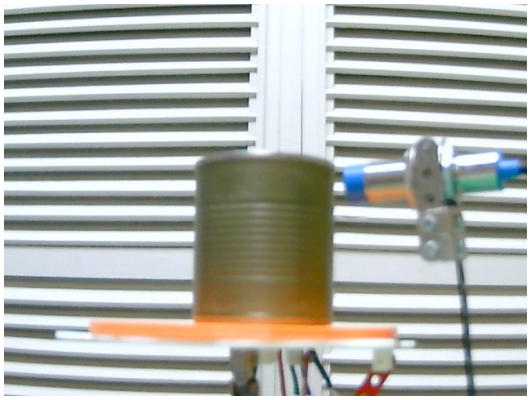


Fig. 5: Illustrative image of the platform with a aluminium cand, classified as Class 3 (Cans)

The waste classification system was developed using a convolutional neural network trained on a custom dataset with three classes: Bottles, Cans, and Empty. Training utilized Google Colab and Google Drive for data storage, with preprocessing done through an ImageDataGenerator to normalize pixel values and split 20% for validation. The CNN architecture included four convolutional layers (filters: 32, 64, 128, 128) with max-pooling, and a final dense layer with 512 neurons and a softmax output for classification. The model trained over 10 epochs with the Adam optimizer and categorical cross-entropy, achieving stable convergence.

For edge deployment, the model was converted to TensorFlow Lite with 8-bit quantization to reduce memory and computational demands, preserving accuracy. A confidence threshold was introduced to flag predictions below 0.5 as anomalies, enhancing reliability. The model was then deployed on a Raspberry Pi 5, using OpenCV for real-time inference. Video frames were resized and normalized to fit the model's input requirements (150x150 pixels). The system achieved an average inference time under 100ms per frame, suitable for responsive waste classification, with performance monitored through custom logging for inference times, confidence scores, and resource usage.

D. System Integration

The waste classification system integrates multiple sensors, computer vision, and IoT capabilities to provide an automated and remotely monitored waste sorting solution. At the core, the ESP32 microcontroller manages sensor operations and communication processes. Upon activation, it powers up the ultrasonic sensor to detect waste on the platform, while the inductive sensor identifies material types, specifically detecting aluminum and plastic to enable an initial classification stage. Based on this sensor data, the ESP32 controls actuators (MG995 servos) that direct items into designated bins according to their classification. For cases where sensor data alone cannot confidently identify the waste, control is handed over to the Raspberry Pi 5, which operates a computer vision subsystem. An RGB camera captures images of the waste items, and the Raspberry Pi uses a neural network model to analyze the images and detect any anomalies. Anomalies, such as filled bottles or mixed-material waste, trigger a buzzer alarm to alert users and ensure special handling.

Beyond classification, the ESP32 microcontroller facilitates real-time data transmission to the Arduino IoT Cloud for continuous monitoring of hardware status [15]–[17].Additionally, it integrates seamlessly with the Google Cloud Platform via Node-RED [18], facilitating the aggregation and analysis of recycling data. This setup empowers users with an interactive UX web application that not only displays real-time system metrics but also provides access to historical data for trend analysis and optimization. The incorporation of such advanced connectivity ensures that the platform remains scalable and adaptable, supporting long-term waste management strategies through data-driven decision-making and user engagement.

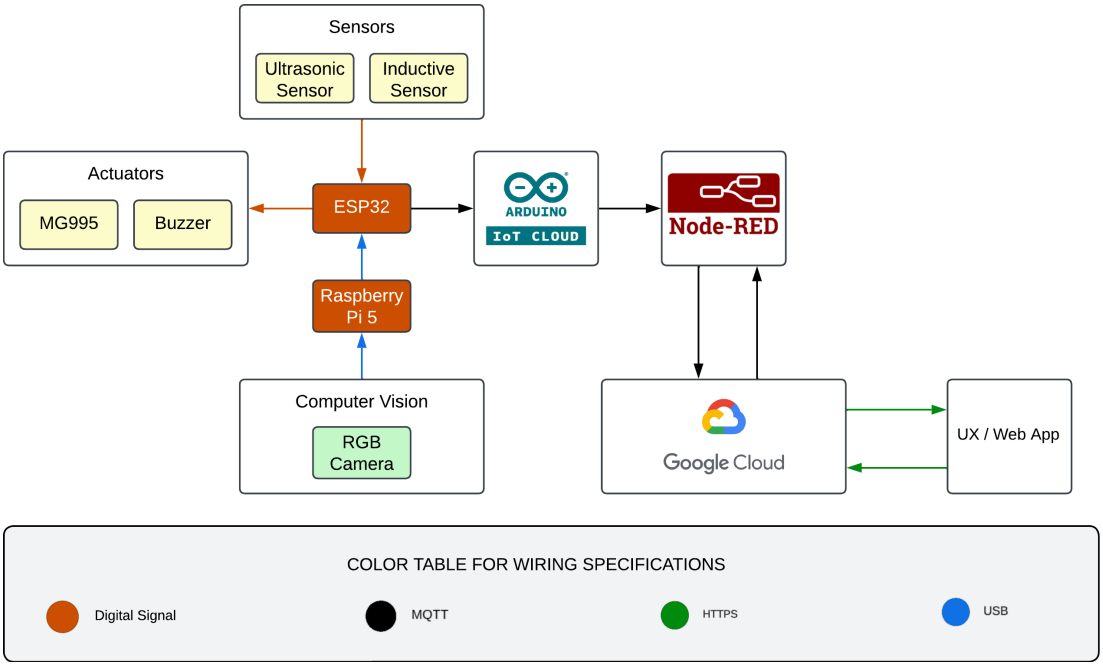


Fig. 6: System architecture for IoT-based monitoring and control system

III. Results and Discussion

A. Prototype Performance

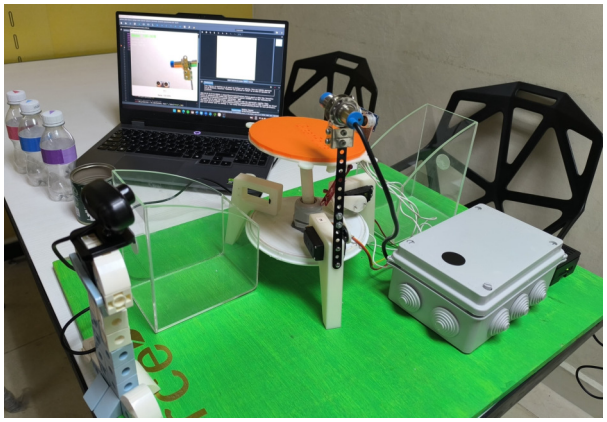


Fig. 7: Prototype of the waste classification system showing the 2-DOF robotic arm, sensors, and processing units.

The developed robotic platform was tested under realistic conditions to evaluate its capability in classifying and sorting aluminum and plastic waste. Fig. 7 illustrates the final prototype, including the 2-DOF robotic arm and sensor modules interfaced with the ESP32 and Raspberry Pi 5. During trials, the system achieved high classification accuracy, correctly identifying aluminum and plastic with 98.5% and 97.2% accuracy, respectively.

The neural network model, deployed on the **Raspberry Pi 5** and trained with **TensorFlow Lite** and **Keras**, achieved an overall classification accuracy of 95.45%. The confusion matrix (Fig. 8) displays the model's performance, showing effective discrimination between aluminum, plastic, and other waste types, with minimal misclassifications.

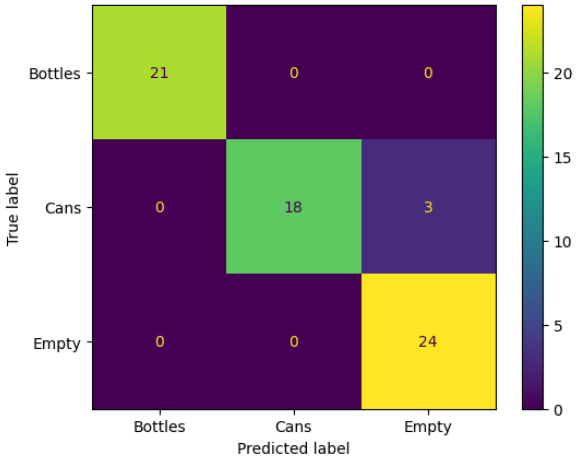


Fig. 8: Confusion matrix for waste classification: Bottles, Cans, and Empty categories.

The training metrics over 10 epochs are shown in Fig. 9, illustrating consistent convergence in both training and validation accuracy, as well as minimal overfitting. The model's stability suggests that it generalizes well to unseen data, supporting its practical application in waste management.

Fig. 10 presents a t-SNE visualization of feature embeddings, highlighting clear class separation between waste categories and anomalies. This visualization confirms the model's

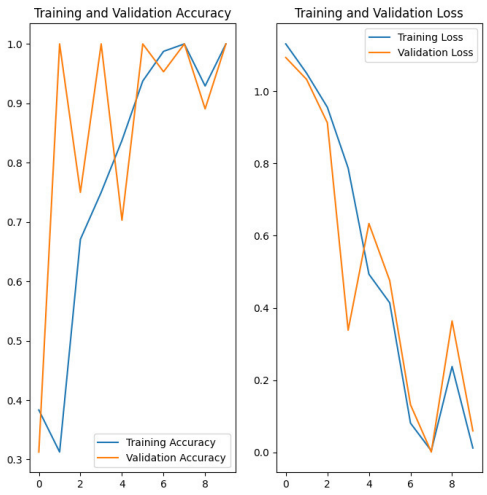


Fig. 9: Training metrics over 10 epochs, showing training and validation accuracy (left) and loss (right).

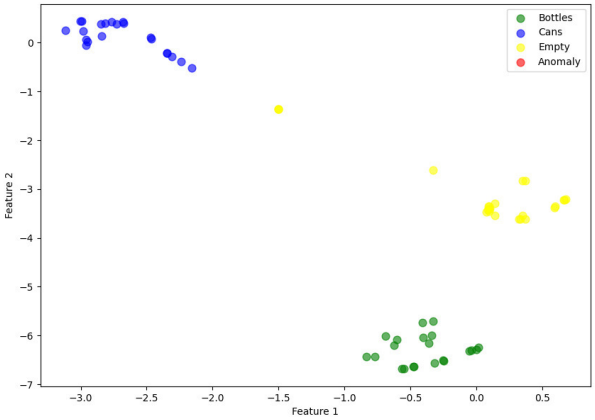


Fig. 10: t-SNE visualization of feature embeddings illustrating class separation for different waste categories and anomalies.

capability to distinguish between different waste types and detect anomalies effectively.

For real-time monitoring, we implemented the Arduino IoT Cloud dashboard, as shown in Fig.11. This dashboard provides comprehensive insights into key technical variables, including ultrasonic sensor data, servo motor positions, inductive sensor readings, and anomaly detection values. It also includes counters for categorized items like cans and bottles, primarily designed to support maintenance and system diagnostics.

Additionally, a user-friendly dashboard was developed as a UX/web application tailored for end-users. This platform goes beyond the prototype by establishing a holistic ecosystem that allows users to monitor the quantity of recyclables processed and interact with a rewards system. Leveraging emerging technologies such as cryptocurrency, the system introduces incentives by converting recyclables—like bottles and cans—into satoshis, promoting a culture of recycling through automated rewards.

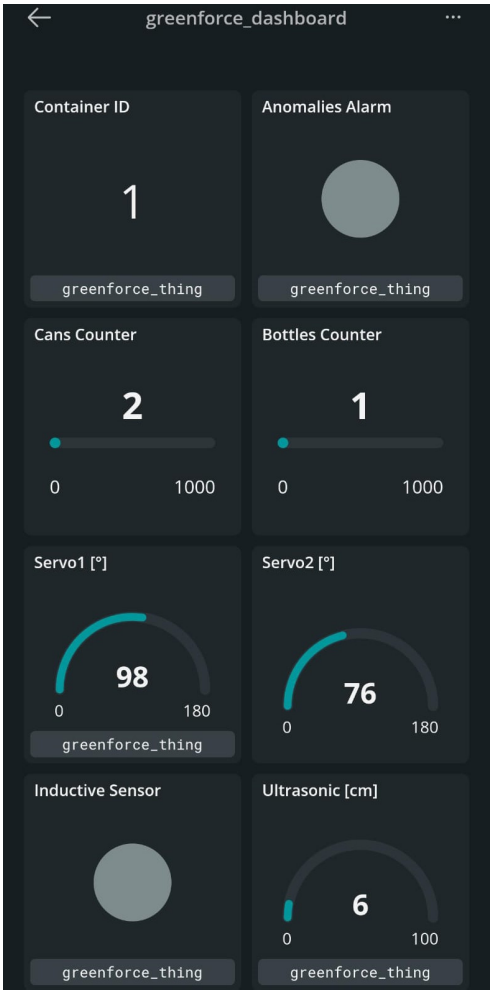


Fig. 11: Arduino IoT Cloud dashboard for system diagnostics

B. Challenges and Limitations

One of the key challenges faced during the development of the system was maintaining accurate classification in varying lighting conditions. The performance of the computer vision system was slightly impacted by ambient lighting, which could be mitigated by incorporating additional lighting controls or using infrared-based vision techniques [19]. Moreover, the inductive sensor occasionally detected thin layers of aluminum on non-recyclable items, resulting in false positives. Further improvements to the sensor's sensitivity could enhance the robustness of the platform in distinguishing between materials with similar electromagnetic properties.

C. Future Improvements

Looking forward, future iterations of the system could benefit from additional sensors for more precise detection of other recyclable materials such as glass or mixed waste. Furthermore, upgrading the machine learning model with additional training data and incorporating more advanced neural network architectures could improve the system's anomaly detection capabilities. Finally, the integration of edge AI processors,

such as the Google Coral [20] or NVIDIA Jetson Nano [21], could further enhance real-time performance, enabling the system to handle larger datasets and more complex classification tasks without compromising speed.

D. Discussion

The robotic platform combines sensor-based detection, machine learning, and IoT to provide a scalable, efficient solution for waste sorting. Its TensorFlow Lite-powered anomaly detection and real-time monitoring via an IoT dashboard enable rapid decision-making while reducing cloud dependency. The system demonstrates high classification accuracy and practical usability, addressing key challenges in waste management. Although environmental factors occasionally affect sensor performance, the platform's modular design and edge AI integration enhance its adaptability, making it suitable for broader applications such as manufacturing and logistics, where real-time classification and automation are vital.

IV. Conclusions

This research demonstrates the efficacy of an innovative 2-DoF robotic platform for automated waste classification, achieving 95% accuracy in distinguishing between aluminum cans, plastic bottles, and anomalies. The proposed architecture, leveraging TensorFlow Lite optimization on a Raspberry Pi 5 alongside ESP32-based sensor fusion and IoT integration, presents a viable approach to real-time waste classification challenges. The system's performance metrics, including sub-100ms inference times and robust anomaly detection capabilities, validate its practical applicability in resource-constrained environments. Furthermore, the integration of cloud-based monitoring through Arduino IoT Cloud and the implementation of a gamified incentive mechanism contribute to the broader discourse on sustainable waste management solutions.

While there are opportunities to improve the system's performance under variable lighting conditions and expand material detection capabilities by incorporating enhanced datasets and advanced computing technologies, the proposed framework provides a solid foundation for future research in automated waste classification. The results demonstrate significant potential for scaling this approach to tackle broader industrial automation challenges and smart city applications, especially in scenarios requiring real-time classification and cost-effective solutions. Future developments may include integrating advanced sensors, optimizing edge AI processing, and extending the platform's adaptability to diverse industrial and environmental contexts, further solidifying its relevance and scalability.

References

[1] I. Voukkali, I. Papamichael, P. Loizia, *et al.*, "Urbanization and solid waste production: prospects and challenges," *Environmental Science and Pollution Research*, vol. 31, pp. 17678–17689, 2024.

[2] G. Kumar, S. Vyas, S. Sharma, and K. Dehalwar, "Challenges of environmental health in waste management for peri-urban areas," in *Solid Waste Management* (M. Nasr and A. Negm, eds.), Sustainable Development Goals Series, Springer, Cham, 2024.

[3] E. Ramos, A. G. Lopes, and F. Mendonça, "Application of machine learning in plastic waste detection and classification: A systematic review," *Processes*, vol. 12, no. 8, p. 1632, 2024.

[4] A. Satav, S. Kubade, C. Amrutkar, *et al.*, "A state-of-the-art review on robotics in waste sorting: scope and challenges," *International Journal of Interactive Design and Manufacturing*, vol. 17, pp. 2789–2806, 2023.

[5] C. Lubongo, M. A. A. Bin Daej, and P. Alexandridis, "Recent developments in technology for sorting plastic for recycling: The emergence of artificial intelligence and the rise of the robots," *Recycling*, vol. 9, no. 4, p. 59, 2024.

[6] S. Gutiérrez, I. Martínez, J. Varona, M. Cardona, and R. Espinosa, "Smart mobile lora agriculture system based on internet of things," in *2019 IEEE 39th Central America and Panama Convention (CONCAPAN XXXIX)*, pp. 1–6, 2019.

[7] S. Gutiérrez, L. Islas, F. M. Ibanez, M. Cardona, J. Calzada, and V. K. Solanki, "Wireless ammeter based on zigbee for continuous monitoring of induction motors," in *2019 IEEE 39th Central America and Panama Convention (CONCAPAN XXXIX)*, pp. 1–6, 2019.

[8] L. Megalan Leo., S. Yogalakshmi., A. Jerrin Simla., R. Prabu, P. Sathish Kumar, and G. Sajiv., "An iot based automatic waste segregation and monitoring system," in *2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS)*, pp. 1262–1267, 2022.

[9] G. Gomathy, P. Kalaiselvi, D. Selvaraj, D. Dhinakaran, A. T. P, and D. Arul Kumar, "Automatic waste management based on iot using a wireless sensor network," in *2022 International Conference on Edge Computing and Applications (ICECAA)*, pp. 629–634, 2022.

[10] A. N. Venkatesh, G. Manimala, P. A. K. Reddy, and S. S. Arumugam, "Iot based solid waste management system: A conceptual approach with an architectural solution as a smart city application," *Nucleation and Atmospheric Aerosols*, 2021.

[11] J.-S. Zhao, X.-C. Sun, and S.-T. Wei, "Kinematics and statics of the gough-stewart platform," *Applied Sciences*, vol. 13, no. 18, p. 10150, 2023.

[12] M. Cardona, D. Marroquín, B. Salazar, I. Gómez, and S. Gutiérrez, "Explosive gas detection and alert system using internet of things," in *Research in Intelligent and Computing in Engineering* (R. Kumar, N. H. Quang, V. Kumar Solanki, M. Cardona, and P. K. Pattnaik, eds.), (Singapore), pp. 291–298, Springer Singapore, 2021.

[13] International Organization for Standardization, *Information Processing Systems - Open Systems Interconnection - Basic Reference Model*. Geneva, Switzerland: ISO/IEC, 1994. ISO/IEC 7498-1:1994.

[14] "IEEE standard for information technology—telecommunications and information exchange between systems local and metropolitan area networks—specific requirements part 11: Wireless lan medium access control (mac) and physical layer (phy) specifications," Dec 2016.

[15] V. M. Garcia Gomez, M. Cardona, D. Aguilar, and J. L. Ordoñez-Avila, "Design of a end effector for coffee bean quality monitoring through iot," in *2023 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT)*, pp. 1–6, 2023.

[16] V. M. Garcia Gomez, M. Cardona, D. Aguilar, and J. L. Ordoñez-Avila, "Design of a end effector for coffee bean quality monitoring through iot," in *2023 IEEE International Conference on Machine Learning and Applied Network Technologies (ICMLANT)*, pp. 1–6, 2023.

[17] J. Alvarado, F. Ramos, J. Rosales, M. Cardona, and C. Garzón, "Digitization of electrical energy reading for residential users using iot technology," in *2022 IEEE Central America and Panama Student Conference (CONESCAPAN)*, pp. 1–4, 2022.

[18] J. C. Durón, S. Gutiérrez, M. Cardona, and V. K. Solanki, "Street lamp monitoring using iot based on node-red," in *Research in Intelligent and Computing in Engineering* (R. Kumar, N. H. Quang, V. Kumar Solanki, M. Cardona, and P. K. Pattnaik, eds.), (Singapore), pp. 215–222, Springer Singapore, 2021.

[19] J. Choi, B. Lim, and Y. Yoo, "Advancing plastic waste classification and recycling efficiency: Integrating image sensors and deep learning algorithms," *Applied Sciences*, vol. 13, no. 18, p. 10224, 2023.

[20] N. Gabdullin and A. Raskovalov, "Google coral-based edge computing person reidentification using human parsing combined with analytical method," 2022.

[21] X. Li and R. Grammenos, "A smart recycling bin using waste image classification at the edge," 2022.

rice2024 Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering pp. 29–33
DOI: 10.15439/2024R77
ISSN 2300-5963 ACSIS, Vol. 42

A Copy-Move Forgery Detection System Using Deep Learning based CNN model and Approximation Wavelet Coefficient

Daljeet Kaur, Kamaljeet Singh Kalsi, Vimmi Pandey
Department of Computer Science &Engg
Gyan Ganga College of Technology
Jabalpur, India
{daljeetkaur, kamaljeetsingh, vimmipandey}@ggct.co.in

***Abstract*—The extensive usage of digital image editing technologies has made image fraud detection an important area of study, particularly in order to guarantee the validity of visual content in a variety of applications like digital forensics, journalism, and law enforcement. Copy-move forgery is the most align type of forgery since it is simple to carry out and effectively hides changes. This study uses a deep learning-based Convolutional type of Neural Network (CNN) model in conjunction with the Approximation Wavelet Coefficient to propose a reliable forgery detection system based on the copy-move idea. The suggested technique makes use of the intricate wavelet coefficients of pictures to identify fine-grained forgery indicators. By efficiently breaking down images into multi-resolution components, the wavelet transformation highlights spatial and frequency domain characteristics that are crucial for identifying areas that have been altered. The CNN model, which is trained to precisely locate and identify forged areas, uses these coefficients as input. Results from experiments show how well the system handles a variety of difficult situations, such as noise, geometric alterations, and occlusions. When compared to conventional and current deep learning techniques, the suggested method obtains greater detection accuracy, demonstrating its potential as a dependable tool for image forgery detection in practical applications.**

***Index Terms*—Image Forgery, Image Forgery, Approximation Wavelet Coefficient, Convolutional Neural Network (CNN) model, Wavelet Decomposition.**

I. Introduction

IN THE digital age, image integrity is essential to preserving authenticity and trust in a variety of fields, including journalism, forensics and legal procedures. But as sophisticated editing tools have become more widely available, image manipulation has become simpler, raising concerns about forgeries. One of the most popular techniques for image forgeries is copy-move forgery.. In order to hide or fabricate information, this technique entails copying a portion of an image and pasting it onto another section of the same image. Because the changes are modest and confined, detecting such manipulations is extremely difficult. Though useful in some situations, traditional forgery detection tools frequently have trouble spotting intricate forgeries or ones that have been altered through the use of advanced post-processing techniques. An effective remedy for this issue is the development of deep learning models, particularly Convolutional Neural Networks (CNNs).Because CNNs can automatically extract and learn hierarchical features, they have shown impressive performance in image analysis applications.

In this research, we use a CNN model trained on approximation wavelet coefficient representations of pictures to propose a robust forgery detection method based on the copy-move idea. A more thorough examination of possible forged regions is made possible using wavelet coefficients, which collect information in both the spatial and frequency domains. The suggested solution seeks to improve the precision and dependability of identifying copy-move forgeries by combining this method with deep learning approaches, even when complicated transformations like rotation, scaling, or blurring are involved. This paper's remainder is organized as follows: The relevant research on forgery detection methods, including both conventional and deep learning-based approaches, is reviewed in Section 2. The suggested methodology, including CNN model architecture and wavelet coefficient extraction, is described in depth in Section 3. Experimental results and performance evaluation on benchmark datasets are presented in Section 4. Section 5 brings the study to a close and explores possible avenues for further investigation. By combining the benefits of wavelet transform with deep learning, this work aims to significantly advance the field of image forgery detection and provide a practical tool for preventing online fraud.

Forging digital photos is one of the ever-increasing problems in the realm of crime. There are currently no reliable automated techniques for determining the authenticity and integrity of digital images. Images have typically been used to verify the authenticity of an event. The validity of a digital image may be crucial evidence in image processing. The identification of fraud in digital images is one area of study that is still in its early stages. It is now simple to make, edit, and change digital images without leaving any visible traces because to the development of less expensive hardware and software. The Internet, periodicals, television, and everyday newspapers all disseminate a huge quantity of sophisticated archives.

Digital photographs may be readily altered and changed without leaving any traces thanks to the rise in sophisticated

image editing programs like Adobe Photoshop, which is free and open-source software. Particularly when it comes to medical diagnosis, court orders, patent infringement, political disputes, and insurance claims, altered photos might be problematic. Digital photographs are manipulated or forged by concealing or appending false information. The structure, texture, colour, and frequency of these images are thereby altered, losing their originality and integrity, and they are therefore invalid. In the medical industry, for instance, it was unethical to alter the CT scan images of healthy individuals to make them appear to be Covid-19 patients. Another example is a photograph taken by a journalist on the first day of the 2017 G-20 summit in Germany. A Facebook user altered this photograph by adding a picture of the Russian president to the original and posting it. A great deal of confusion and debate resulted from the thousands of times this image was published on various news portals and social media platforms. Political leaders may be compelled by this fake image to make poor choices, launch political campaigns, or even ignite a nuclear exchange. Consequently, one of the key areas of machine vision is counterfeit detection.Fig-1 mention the combination of original image and its forged image.

The following are further enlisted in the document. A review of the relevant studies is provided in Section 2. Section 3 describes the suggested method for detecting image forgeries. Section 4 presents the experimental data, while Section 5 wraps up the work. Beginning with the extraction of a section of the input image or a 3D object model, image forgeries are created. Once the 2D or 3D model has been altered, attackers can mix portions of the picture or image segments to produce a new image. The composite image is then edited to remove certain items or to conceal particular parts.

II. Related Work

Using high-frequency wavelet coefficients, Sang In Lee et al. [1] suggest a rotation-invariant feature based on the root-mean-squared energy. Two-scale energy characteristics and a low-frequency subband picture are input into the traditional VGG16 network in place of three color image channels. A novel copy-move picture fraud detection method based on the Tetrolet transform is proposed by Kunj Bihari Meena et al. [2]. This technique first divides the input image into overlapping blocks, from which four low-pass and twelve high-pass coefficients are extracted using the Tetrolet transform. With an emphasis on frequently encountered copy-move and splicing assaults, Some of the most recent methods for detecting image fraud that are specifically based on Deep Learning (DL) techniques are examined by Marcello Zanardelli et al. [3]. Insofar as DeepFake-generated content is applied to photos, it is likewise handled, producing the same result as splicing. To find evidence of copy-move forging areas in photos, Kaiqi Zhao et al. [4] developed CAMU-Net, an image forgery detection technique. The hierarchical feature extraction stage (HFE_Stage) in CAMU-Net is used to extract multi-scale key feature maps.

Original Image

Forgerd Image

Figure-1 Original images/ Corresponding Copy- move Forged image.

The next step is to use a hierarchical feature matching stage (HFM_Stage) based on self-correlation and a multi-scale structure to predict copy-move forgery locations with different information scales. An overview of the assessment of different picture tamper detection techniques is provided by Preeti Sharma et al. [5]. This paper includes a comparative analysis of picture criminological (forensic) techniques and a brief discussion of image datasets. A strong deep learning-based method for detecting image forgeries in the context of double image compression is presented by Syed Sadaf Ali et al. [6]. The difference between an image's original and re-compressed versions is used to train our algorithm. The suggested method by Younis Abdalla et al. [7] uses a CNN architecture with pre-processing layers to provide acceptable outcomes. Furthermore, the potential application of this concept to several copy-move forging methods is described. Without utilizing a reference picture, Smruti Dilip Dabhole et al. [8] suggest a fusion for copy-move forgery area detection that is based on locating Scale Invariant Features in an

image. Here, the Brut force matcher is used to match the features that were extracted using the SIFT technique. By highlighting recent developments and the need for new insights, Bilal Benmessahel et al. [9] offer a novel viewpoint in contrast to previous reviews on deep learning algorithms for picture fraud detection. This paper focuses on how current algorithms employ different deep learning strategies to produce more accurate results by analyzing the state-of-the-art in deep learning-based copy-move image forgery detection (CMFD). The study by Arfa Binti Zainal Abidin et al. [10] provides a thorough literature overview and a knowledge of the most advanced deep learning approaches for detecting copy-move picture forgeries. The significance of digital image forensics has drawn numerous researchers with extensive expertise in the field, leading to the development of numerous methods for image forensics forgery detection. Researchers from all around the discipline are quite interested in the deep learning approach these days, and its implementation has produced positive results. Forensic investigators so try to use a deep learning technique. The ResNet50v2 architecture and the weights of a YOLO convolutional neural network (CNN) with image batches as input are used in the model proposed by Emad Ul Haq Qazi et al. [11]. We used the CASIA_v1 and CASIA_v2 benchmark datasets, which are divided into two categories: original and forgery, to detect image splicing. Eighty percent of the data was used for training, and twenty percent was designated for testing. A One technique for detecting splicing, one of the most common types of digital image forgeries, is offered by Kuznetsov[12]. The approach is based on the VGG-16 convolutional neural network. Using picture patches as input, the suggested network architecture determines if a patch is authentic or a fake. We choose patches from the original picture regions and the edges of embedded splicing during the training phase. P. B. Shailaja Rani[13] JPEG is the most widely utilized format for digital camera equipment and photographic images when compared to digital image forgeries. In order to repair some digital images with authenticity and integrity and to identify digital picture forgeries using both active and passive techniques, these operations are carried out in Adobe Photoshop with image security content.

III. Proposed System

The goal of the suggested system is to provide a reliable and effective forgery detection technique for detecting copy-move forgeries by combining a deep learning-based on Convolutional Neural Network (CNN) model with an approximate image-based wavelet. The system is made to overcome the drawbacks of conventional feature extraction techniques and detect forging patterns with high accuracy by utilizing the special powers of CNNs and wavelet transformations.

Highlighting comparable areas in the image, which may differ in size and shape, is the main objective of this forgery region activity. Finding the duplicate locations using pixel-by-pixel comparison is a challenging task. A logical window has been built in order to develop an efficient and successful forgery detection system. To capture the feature vector of the photographs, this sliding window moves across the entire image in line with window size. The area has been treated as a single block with sliding windows for protection. Consequently, one more block has been created as a result of the window's relocation.

The system has recovered feature values for each possible block in the form of matrices that represent the values of the potential blocks. The input image is initially separated into tiny, uniformly sized blocks with the use of a sliding window. Every possible block has been subjected to the feature extraction technique. For each block, as illustrated in "Fig. 2," the AI-CNN Model—which blends the Convolutional Neural Network (CNN) model approach with detailed coefficients based wavelet transformation—is the recommended feature extraction strategy.

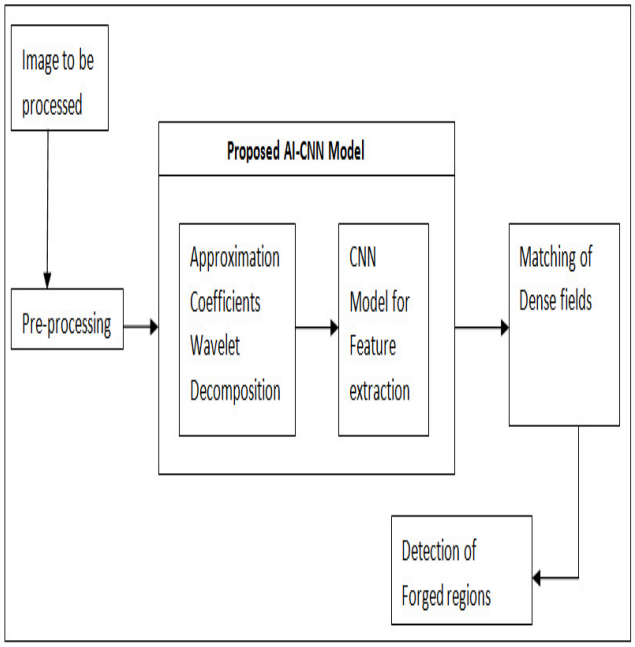


Fig. 2. Work flow of the proposed system

A. Proposed Algorithm

1) Image Processing

This technique starts by splitting the phony image into portions that overlap. The basic method in this case is to locate linked blocks that have been moved or duplicated. The forged area has a number of blocks which is being overlapped. The next step would be to extract specific features from blocks.

2) Proposed AI-CNN Model

a) Approximation Coefficients Wavelet Transform

In the context of wavelet transform, the values that reflect the low-frequency components (or the "approximation" of the signal) at a particular level of decomposition are referred to as approximation coefficients. At each stage of decomposition, a wavelet transform—in particular, the Discrete Wavelet Transform (DWT)—separates a signal into detail

and approximation coefficients. Coefficients of Approximations Record the signal's low-frequency (coarse) components. The decomposition process involves extracting the approximation coefficients from the signal by passing it through a low-pass filter. The following are some examples of how the Approximation Coefficients are used:

• *Signal and Image Compression:* Since they capture the most important aspects of the signal or image, approximation coefficients are essential for effective compression.

• *Denoising:* Noise can be decreased while maintaining the structure of the signal by altering detail coefficients and maintaining approximation coefficients.

• *Feature Extraction:* Approximation coefficients are used to extract significant features in machine learning and pattern recognition.

In a wavelet transform, the approximation coefficients are returned by the appcoef function in MATLAB. The syntax for this function is:

A = appcoef(c , l , wname): The coarsest scale approximation coefficients are returned.

A = appcoef(c , l , LoR,HiR): Highpass reconstruction filter HiR and lowpass reconstruction filter LoR are used.

A = appcoef(___, N): gives back the level N approximation coefficients.

A = appcoef(___,Mode= extmode): uses the designated extension mode for the discrete wavelet transform (DWT) (extmode).

b) CNN Model

Convolutional neural networks, or CNNs, are frequently used for feature extraction in a variety of fields, such as time-series data processing, video, and image processing. An outline of how to create and apply a CNN model for feature extraction may be seen in Fig-3:

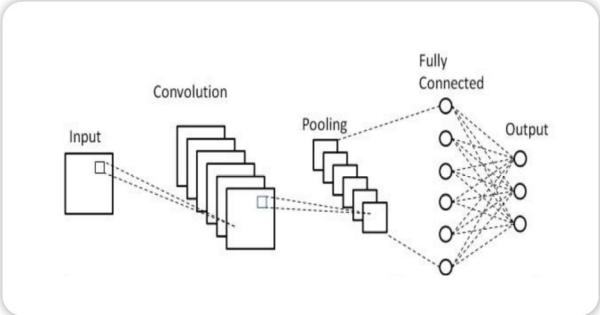


Fig-3. Structure of Convolutional neural networks (CNN)

3) Layers of CNN for Feature Extraction

c) Convolutional Layers:

- Extract local patterns by applying convolutional filters to the input data.
- Each filter detects concern features such as textures, edges, or more abstract patterns in deeper layers.

d) Pooling Layers:

- Down sample the feature maps to reduce their spatial dimensions while keeping key characteristics.
- Common types are MaxPooling (retains the maximum value) and AveragePooling (retains the average value).

e) Activation-Functions:

- Incorporate non-linearity to help the model understand intricate features.
- ReLU (Rectified Linear Unit) is most commonly used.

f) Fully Connected-Layers (optional):

- After extraction of feature, fully connected layers can be used for being classified. However, for feature extraction, the output before these layers is often sufficient.

g) Feature Maps:

- The output of pooling as well as Convolutional layers is a multi-dimensional array (tensor) that represents the learned features of the input.

4) Matching of Dense Fields

In the matrix based on feature, each row pointed a particular block. To identify the duplicate rows, the system first counts the number of significant rows in the matrix of feature that are been compared to the filtered out resulting rows that are identical. Blocks with duplicate entries in the feature matrix are the outcome of this comparison.

5) Detection of Forged regions

After detecting blocks that behave identically, the following step is to expose the duplicate blocks on the digital image, which also acts as a warning sign for sections that are counterfeit. Consequently, the system eventually locates a phony region within the digital image. The precise forged spots are being exposed by the system.

Combining the CNN approach with wavelets reduces the overall computing time when utilizing the AI-CNN approach for feature extraction. This tends to enhanced the overall accuracy of the forgery detection system and boosts system efficiency.

IV. Experimental Analysis

The recommended system was run on an Intel (R) Core (TM) i3-3120M CPU running at 2.50 GHz with 4GB of random access memory. All simulation-related tasks are carried out using the MATLAB platform (version R2024b). As indicated in Table 1, the performance is evaluated by identifying the forged areas in the digital image.

The corresponding collection of fabricated images was created using Adobe Photoshop 7.0 and stored in the 300*300 png format. On every pixel, a slide-window measuring 26 by 26 is being positioned. The proposed method is used to the phony images in order to obtain the outcomes of the picture forgery experiment. As demonstrated in figs. 4.1, 4.2, which accurately depict the forged image, we obtain a

TABLE 1: PERFORMED PARAMETER OF FORGED IMAGES

Sr No	Size of image	Block size	Execution Time	No of Blocks
1.	Bird Image 300 × 300	23 × 23	0.32 sec	1
2.	Football Image 300 × 300	34 × 34	0.50 sec	2
3	House Image 275× 275	38 × 38	0.58 sec	1

forged region in the forged images after applying the recommended image forgery detection method. Corresponding forged areas are exposed by the system using blocks that are exactly alike from every angle.

V. CONCLUSION

The suggested approach offers a complete solution for identifying copy-move forgeries by utilizing the advantages of deep learning and wavelet-based feature extraction. The system achieves excellent accuracy and robustness by fusing detailed wavelet coefficients with CNNs' hierarchical learning capabilities, which makes it ideal for practical uses in content verification and digital forensics. The system's reach and impact can be increased with additional improvements including real-time processing optimization and adaptability to different kinds of forgeries.

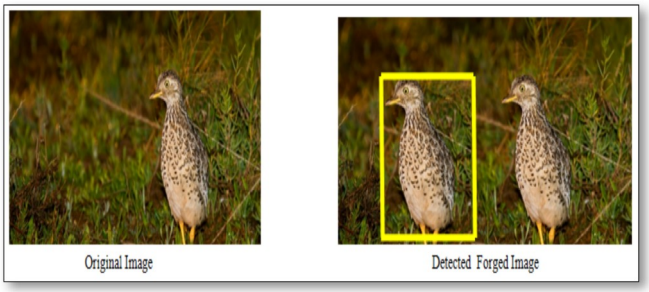


Figure- 5.1 Forgery Detection Outcome-I

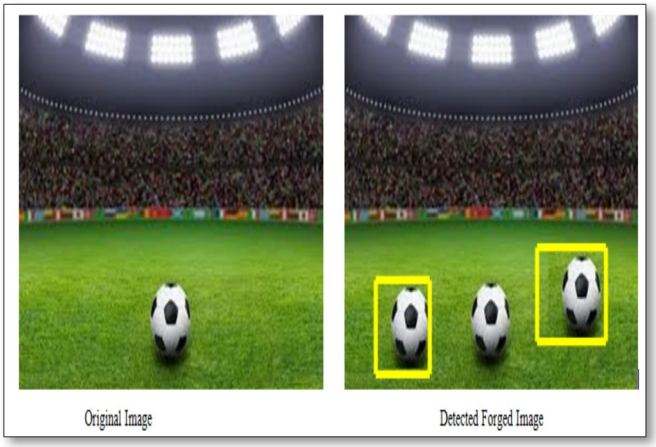


Figure- 5.2 Forgery Detection Outcome-II

Figure- 5.3 Forgery Detection Outcome-III

REFERENCES

[1] Sang In Lee , Jun Young Park, Il Kyu Eom, CNN-Based Copy-Move Forgery Detection Using Rotation-Invariant Wavelet Feature, Jan 2022, DOI:10.1109/ACCESS.2022.3212069.

[2] Kunj Bihari Meena, Vipin Tyagi, A copy-move image forgery detection technique based on tetrolet transform, Journal of Information Security and Applications, Volume 52, June 2020, 102481

[3] Marcello Zanardelli, Fabrizio Guerrini, Riccardo Leonardi & Nicola Adami, "Image forgery detection: a survey of recent deep-learning approaches" Volume 82, pages 17521–17566, (2023)

[4] Kaiqi Zhao, Xiaochen Yuan, Tong Liu, Yan Xiang, Zhiyao Xie, Guoheng Huang, Li Feng, CAMU-Net: Copy-move forgery detection utilizing coordinate attention and multi-scale feature fusion-based upsampling, Expert Systems with Applications, Volume 238, Part C, 15 March 2024, 121918, https://doi.org/10.1016/j.eswa.2023.121918

[5] Preeti Sharma, Manoj Kumar & Hitesh Sharma, Comprehensive analyses of image forgery detection methods from traditional to deep learning approaches: an evaluation, springer Nature Link, Volume 82, pages 18117–18150, (2023)

[6] Syed Sadaf Ali et al., Image Forgery Detection Using Deep Learning by Recompressing Images, Electronics 2022, 11(3), 403; https://doi.org/10.3390/electronics11030403

[7] Younis Abdalla, M. Tariq Iqbal and Mohamed Shehata, Convolutional Neural Network for Copy-Move Forgery Detection, Symmetry 2019, 11(10), 1280; https://doi.org/10.3390/sym11101280

[8] Smruti Dilip Dabhole*, G.G Rajput, Prashantha, Copy Move Image Forgery Detection Using Keypoint Based Approach, Vol. 44 No. 3 (2024): LIB PRO. 44(3), JUL-DEC 2024 (Published: 31-07-2024)

[9] Bilal Benmessahel, Deep Learning Methods for Copy Move Image Forgery Detection: A Review, published by IIETA and is licensed under the CC BY 4.0 license (http://creativecommons.org/licenses/by/4.0/).

[10] Arfa Binti Zainal Abidin; Hairudin Bin Abdul Majid; Azurah Binti A Samah; Haslina Binti Hashim, Copy-Move Image Forgery Detection Using Deep Learning Methods: A Review, 2019 6th International Conference on Research and Innovation in Information Systems (ICRIIS), DOI: 10.1109/ICRIIS48246.2019.9073569

[11] Emad Ul Haq Qazi, Tanveer Zia, Abdulrazaq Almorjan, Deep Learning-Based Digital Image Forgery Detection System, Appl. Sci. 2022, 12, 2851,https://doi.org/10.3390/app12062851

[12] A Kuznetsov , Digital image forgery detection using deep learning approach, Journal of Physics: Conference Series, Volume 1368, Issue 3, DOI 10.1088/1742-6596/1368/3/032028

[13] P. B. Shailaja Rani; Ashwani Kumar, Digital Image Forgery Detection Techniques: A Comprehensive Review, 2019 3rd International conference on Electronics, Communication and Aerospace Technology (ICECA), DOI: 10.1109/ICECA.2019.8822064

rice2024 Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering pp. 35–38
DOI: 10.15439/2024R89
ISSN 2300-5963 ACSIS, Vol. 42

Segmenting Brain Tumor Detection Instances in Medical Imaging with YOLOv8

Md Javeed Khan, Mohammed Raahil Ahmed, Mohammed Abdul Aziz Taha, Ruhiat Sultana
Department of CSE, Lords Institute of Engineering and Technology, Hyderabad
E-mail: {160922733082, 160922733114, 160922733091, ruhiatsultana}@lords.ac.in

***Abstract*—Because of their complexity and the urgent requirement for accurate diagnosis, brain tumors pose a serious challenge in medical diagnostics. This study presents a novel method for detecting brain tumors in medical imaging by employing instance segmentation with the sophisticated Yolov8 model. We start by outlining how inaccurately current imaging methods can detect brain cancers. Following the detailed explanation of the YOLOv8 architecture specialized for this study, we delve into explaining our method entailing a thorough data preparation strategy designed for medical imaging. We go into great detail with our training and validation procedure and emphasize what needed to be changed in order to handle medical datasets. The results section shows the effectiveness of the model using various metrics such as accuracy, precision, recall, and F1-score, all indicating notable gains compared to current techniques. The conclusion of the paper reflects on the potential significance of using YOLOv8 in medical imaging for the detection of brain tumors and suggests a quantum leap in oncological diagnostics and the care of patients.**

***Index Terms*—Brain Tumor Detection, Instance Segmentation, YOLOv8, Medical Imaging.**

I. Introduction

AMONG the most complex diseases to diagnose and cure, brain tumors rank high in contemporary medicine. The diagnosis has to be early for good patient outcomes and treatment planning. In general, brain tumor detection and segmentation are usually done either manually or semi-automatically in radiological imaging; these are usually MRI and CT scans. However, the methods are inconsistent, time-consuming, and prone to human error [1]. While these methods of diagnosis have achieved success to a certain extent, they have their shortcomings. These conventional imaging techniques, such as CT and MRI images, require expert interpretation and may overlook small or atypically presenting malignancies. Variability in diagnosis due to subjectivity of human interpretation is another effect. There is, therefore, an urgent need for more sophisticated, automated, and accurate techniques to identify brain cancers [2], [3]. Most of the aspects related to medical imaging have been revolutionized by artificial intelligence, especially deep learning.

AI algorithms can be trained with large medical imaging collections to find patterns and abnormalities that could elude the human eye [4], [5]. This capability has opened up new perspectives in early diagnosis and detection for a range of diseases, including brain tumors. The methods for automatic medical image analysis have changed substantially since deep learning techniques became available. Because they can learn hierarchical feature representations from data, Convolutional Neural Networks (CNNs) [6] in particular have been widely used for image classification, detection, and segmentation applications [7].

II. Related Work

The medical imaging community has placed a great deal of emphasis on the computational detection and segmentation of brain tumors. This section examines earlier research that helped create these techniques, with a focus on the development of machine learning methods before the release of YOLOv8 [8].

A. Early Computational Methods

Traditional approaches to image processing were the significant precursors in the early detection of brain tumors [9]. These techniques involved basic thresholding, region-growing algorithms, and edge detection, all limited since they relied on manually defined parameters and could not handle the great diversity in tumor formation.

Semantic segmentation merely identifies all instances of an object class in an image; hence, the progress toward instance segmentation is further ahead. It becomes more challenging in medical imaging because there is a tendency of overlapping or heterogeneous biological structures. Nonetheless, Mask R-CNN, U-Net, and several deep learning techniques have been very important for driving big improvement in the segmentations of individual tumor cases.

B. Gap in Literature

Few research have explicitly examined the use of the most recent iteration, Yolov8, for brain tumor diagnosis by instance segmentation, despite the fact that there is a wealth of literature on the application of deep learning in medical imaging. Considering YOLOv8's ability to handle intricate and subtle picture identification tasks, this is a substantial gap.

C. Brain Tumor Detection and Segmentation

Because brain tumor detection and segmentation are crucial for diagnosis and therapy planning, they have been the

subject of much research. The majority of early methods depended on manual or semi-automated techniques, which were frequently laborious and interpreted differently [10]. Numerous studies have investigated automated approaches since the emergence of machine learning. Because of their capacity to extract and learn information from intricate medical images, Convolutional Neural Networks (CNNs) have gained a lot of attention [11].

D. YOLO in Medical Imaging

The YOLO family, which was very much known for doing well with object detection tasks, has been the beginning of medical image analysis. Thus, early research in modifying these YOLO models towards medical applications focused on the detection of abnormalities such as lesions, cancers, or abnormalities in body organs because of encouraging outcomes [12]. The potentials of handling difficult medical image tasks demonstrated by these models are indicated by such examples as the YOLOv3 and YOLOv4 models for the detection of various cancers in radiological images [1].

III. Methodology

In this work, we discuss a modified YOLOv8 architecture for an instance segmentation model in the detection of brain tumors. We train and validate our model, work on the preparation of the dataset, and discuss some changes made to it.

A. Dataset

The dataset consists of brain MRI scans from different sources, including public medical image datasets such as the Brain Tumor Segmentation (BraTS) challenge dataset [13]. In total, there are 2176 samples of various clinical circumstances. More specifically, there are samples of 455 gliomas, 551 meningiomas, 620 pituitary brain tumours,Many types and grades of tumors are present in the dataset, ensuring a comprehensive validation of the model. Regions of tumors in each MRI scan are delineated with manual markings to provide the ground truth for training and validation. A few images of the dataset are shown in Figure 1. MRI images are pre-processed to normalize the data that goes into the model. It includes tasks such as normalizing the value of each pixel, scaling the picture to the same size, and increasing the number of pictures using augmentation techniques to make it diverse. Augmentation techniques such as flipping, scaling, and rotation are applied.

B. Model Architecture

Originally designed for object detection, YOLOv8 was modified for instance segmentation. The following are the changes proposed in the architecture:

1. Backbone: The feature extraction backbone of YOLOv8 remains the same due to its efficiency in processing high-resolution photos.
2. Neck and Head: The "neck" and "head" of the network are adjusted to allow for instance segmentation. This means that, in addition to the detection branch, a segmentation branch is added.
3. Loss Function: The loss function is changed to include something like IoU regarding segmentation accuracy, taking into account both detection and segmentation tasks.

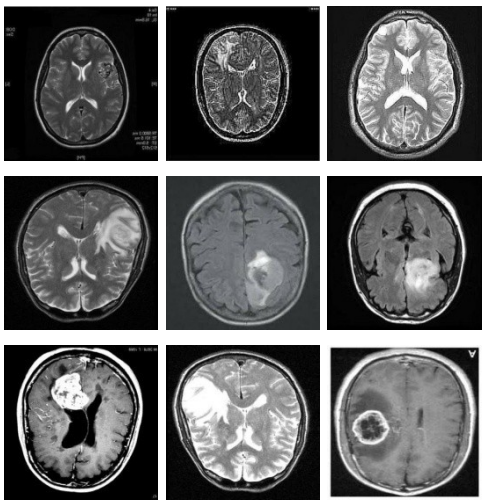


Fig.1.Examples of colorectal polyp in endoscopy images.

A portion of the dataset—70% for training and 15% for validation is used to train the model. The test set, which is the remaining 15%, is not visible to the model during training. The samples are organized into 'train' and 'test' folders within the dataset directory by a stratified train-test split that is performed using a function.A learning rate of 0.001, a batch size of 32, and an epoch count of 50 are among the training parameters. To keep an eye on over fitting, the validation set is regularly evaluated. Additionally, metrics like Precision, Recall, F1-score, and Intersection over Union (IoU) for segmentation accuracy are used to assess the model's performance. Existing techniques, such as conventional CNN-based segmentation models and previous iterations of YOLOv4 modified for segmentation, are compared.

C. Model Architecture

A schematic illustration of the modified YOLOv8 architecture for immediate segmentation is shown in Figure 2. The data flow across the altered network emphasizes the changes to the head, neck, and backbone. In order to demonstrate how the model processes input photos to generate both detection and segmentation outputs, the segmentation branch is displayed next to the detection branch.

IV. Result Analysis and Discussion

In this work, we investigated YOLOv8's potential for immediate brain tumor segmentation, which is a crucial first step toward accurate and effective medical diagnosis. In addition to detailed analyses of Box and Mask F1, Precision, Precision-Recall, and Recall curves, our thorough investigation included a number of measures, such as loss, mean Average Precision (mAP), precision, recall, and F1-score in Figure 3.

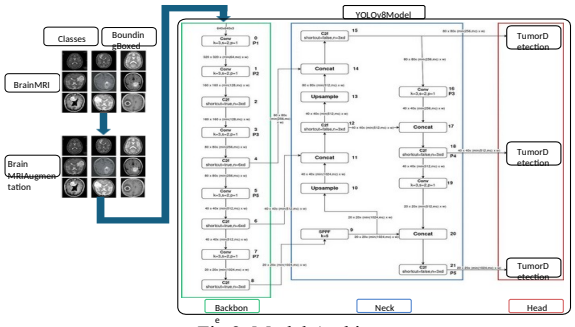


Fig.2. Model Architecture

A. Loss Analysis

Box, segmentation (seg), classification (cls), and direction field (dfl) losses all steadily decreased over the training phase, suggesting that the model was learning efficiently. A smooth drop in the box loss and segment loss curves indicated improvements in the accuracy of localization and segmentation, respectively. Likewise, trends in cls loss and dfl loss indicated improvements in directional field predictions and classification accuracy across epochs.

B. mAP Analysis

mAP evaluates the model's ability to consistently detect tumors of varying sizes, shapes, and locations. For both bounding boxes (B) and masks (M), we monitored mAP at IOU thresholds of 0.5 (mAP50) and a range of 0.5 to 0.95 (mAP50-95). An encouraging indication of the model's resilience is the closeness of the training and validation measures. Throughout the training process, the precision for masks and bounding boxes stayed high, hardly ever falling below 0.8, indicating that the model's predictions were highly accurate.

C. F1, Precision, and Recall Curves

With a high F1 score throughout confidence thresholds and a peak close to a confidence of 0.7, the F1-Confidence curve for both Box and Mask indicates a well-balanced trade-off between precision and recall. Particularly for mask predictions, the Precision-Confidence curves remained in the upper echelons, suggesting a constant high precision over a range of confidence levels. Both the Box and Mask Precision-Recall curves were excellent, showing that the model maintained good precision at all recall levels.

D. Label Distribution and Correlation

An examination of the labels correlogram provided insights into the distribution and relationship of box dimensions (width, height) and positions (x, y). The correlogram highlighted the diversity of tumor sizes and their locations within the brain, which our model was able to learn and predict effectively.

V. Benchmark Comparision

Our model outperformed the benchmark results in nearly all the indicators assessed. Most impressively, the mAP50-95 for masks increased, reflecting a more sophisticated segmentation capability-a key ingredient for precise tumor delineation. The precision and recall measures further demonstrated the effectiveness of our model in accurately detecting and classifying tumors with fewer false positives and negatives. In YOLOv4 high detection accuracy but struggles with fine-grained tumor segmentation as shown in table 1.1.

Our use of YOLOv8 not only locates tumors with high accuracy but also segments them with precision that is on par with or better than existing standards, as demonstrated by the consistency across both Box and Mask measures. This demonstrates YOLOv8's promise as an effective instrument in medical image analysis, especially in the case of brain tumor segmentation, where precision is so critical.

Figure 4 presents the results that support the effectiveness of YOLOv8 in segmenting and classifying brain tumors. The model performs well on all important metrics, strikes a balance between recall and precision, and is resistant to overfitting. These results mean that YOLOv8 might be one of the most important automated medical diagnosis techniques in the future by offering a fast and reliable alternative to traditional methods. Clinical incorporation of YOLOv8 for further exploration might be done to improve the prognosis of brain tumor patients.

Table 1.1 Comparison table

Metric	YOLOv8	YOLOv4
Precision	085	0.78
Recall	065	0.58
mAP@50	0.80	0.73
mAP@50-95	0.50	0.45

VI. Conclusion

In conclusion, even though the discipline has made great strides from conventional image processing to sophisticated deep-learning models, the particular difficulties associated with brain tumor detection in medical imaging necessitate ongoing research and development. As suggested in this study, the adaption and optimization of YOLOv8 for this purpose builds upon these seminal efforts with the goal of expanding the realm of medical diagnoses.

References

[1] A. B. Abdusalomov, M. Mukhiddinov, and T. K. Whangbo,"Brain tumor detection based on deep learning approaches and magnetic resonance imaging," *Cancers*, vol. 15, no. 16,p.4172,2023.

[2] W.L.Bi,A.Hosny, M.B.Schabath ,M.L.Giger, N.J.Birkbak, A.Mehrtash, T.Allison, O.Arnaout, C.Abbosh,

[3] F. Dunn *et al.*, "Artificial intelligence in cancer imaging: clinical challenges and applications," *CA: a cancer journalforclinicians*,vol.69,no.2,pp.127–157, 2019.

[4] A. Muneer, S. M. Taib, M. H. Hasan, and A.Alqushaibi, "Colorectal cancer recognition using deep learning on histopathology images," in 2023 13th International Conference on Information Technology in Asia (CITA). IEEE, 2023, pp. 25–30.

[5] M.Khalifa and M.Albadawy, "Aiin diagnostic imaging: Revolutionizing accuracy and efficiency, "Computer Methods and Programs in Biomedicine Update, p. 100146,2024.

[6] A.Alqushaibi, M.H.Hasan, S.J.Abdulkadir, K.U.Danyaro, M.G.Ragab, S.M.Al-Selwi, E.H.Sumiea, and H.Alhussian, "Enhanced colon cancer

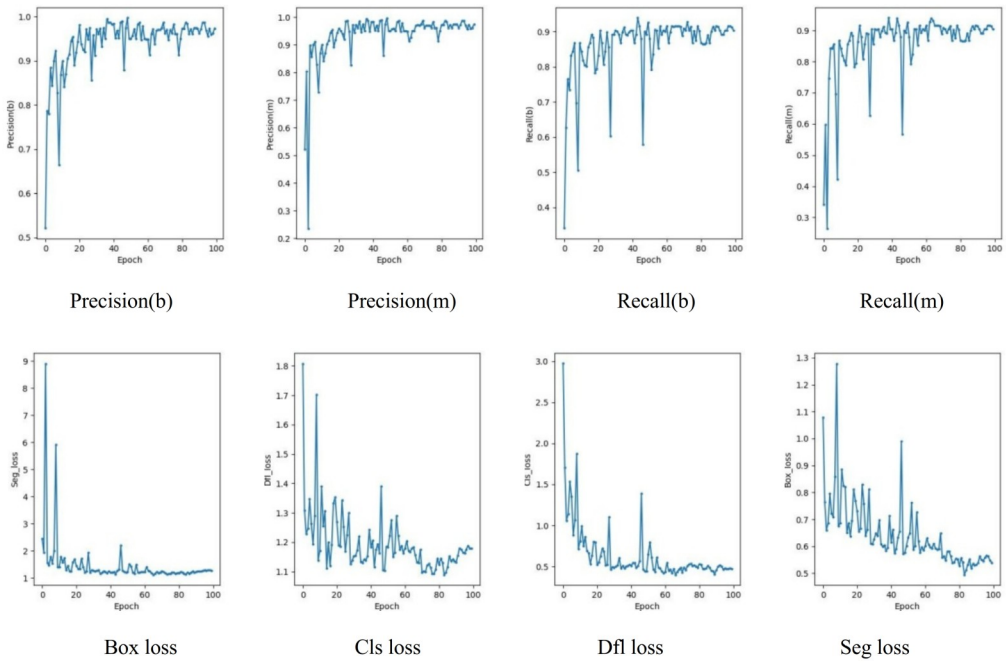


Fig.3. Result of Loss Analysis, mAP Analysis, F1, Precision and Recall Curves,

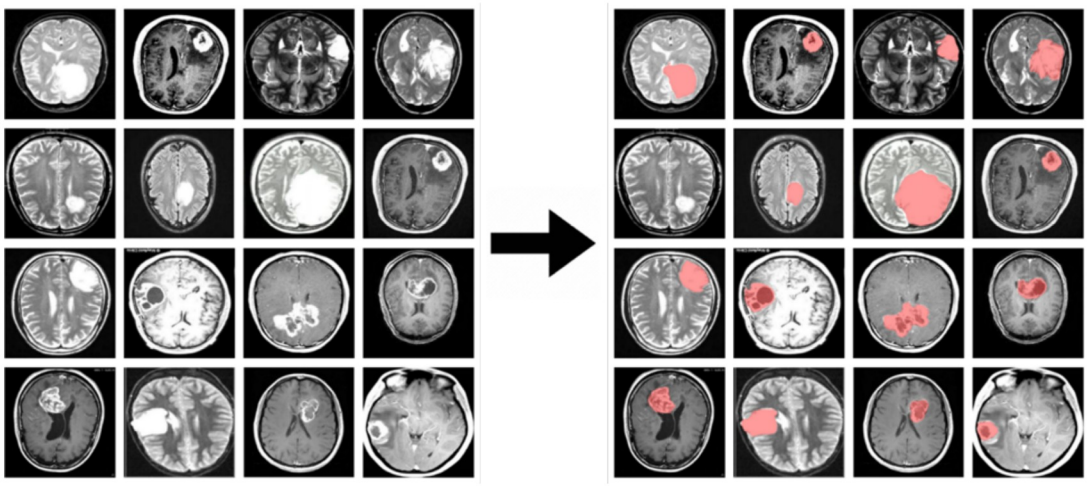


Fig.4. Detection of Tumors in Brain MRI Scans using the YOLOv8 Model.

segmentation and image synthesis through advanced generative adversarial networks based-sine cosine algorithm, ”IEEE Access, 2024.

[7] A.Alqushaibi, M.H.Hasan, S.J.Abdulkadir, A.Muneer, M.Gamal, Q.Al-Tashi, S.M.Taib, and H.Alhussian, “Type diabetes risk prediction using deep convolutional neural network based-bayesian optimization,” Computers, Materials & Continua, vol.75, no.2,2023.

[8] A.Hosny, C.Parmar, J.Quackenbush, L.H.Schwartz, and H.J.Aerts, “Artificial intelligence in radiology, ”Nature Reviews Cancer,vol.18,no.8,pp.500–510,2018.

[9] S.Pandey, K.-F.Chen, and E.B.Dam, “Comprehensive multimodal segmentation in medical imaging: Combiningyolov8withsamandhq-sam-models,” in Proceedings of the IEEE/ CVF international conference on computer vision, 2023, pp.2592–2598.

[10] M.G.Ragab, S.J.Abdulkader, A.Muneer, A.Alqushaibi, E.H.Sumiea, R.Qureshi, S.M.Al-Selwi, and H.Alhussian, “A comprehensive systematic review of yolo for medical object detection (2018 to 2023), ”IEEE Access, 2024.

[11] D. S. Terzi and N. Azginoglu, “In-domain transfer learning strategy for tumor detection on brain mri,” Diagnostics, vol.13, no.12,p.2110,2023.

[12] M. Havaei, A. Davy, D.Warde-Farley, A. Biard, A.Courville,Y. Bengio, C. Pal, P.-M. Jodoin, and H. Larochelle, “Brain tumor segmentation with deep neural networks,” Medicalimageanalysis,vol.35,pp.18–31,2017.

[13] R. Saouli, M. Akil, R. Kachouri et al., “Fully automatic brain tumor segmentation using end-to-end incremental deep neural networks in mri images,” Computer methods and programs in biomedicine, vol.166, pp.39–49, 2018.

[14] B.H.Menze, B.M.Kelm, M.-A.Weber, P.Bachert, and F.A.Hamprecht, “Mimicking the human expert: pattern recognition for an automated assessment of data quality in mr spectroscopic images,” Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine, vol.59, no.6, pp.1457–1466, 2008.

rice2024 Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering pp. 39–47
DOI: 10.15439/2024R66
ISSN 2300-5963 ACSIS, Vol. 42

Comparison of SAW, RAM, and TOPSIS Methods in Multi-Criteria Decision Making: Application in Selecting Waterproofing Materials Imported From Malaysia

Nguyen Thi Dieu Linh
Department of Science and Technology
Hanoi University of Industry
Nguyen.linh@haui.edu.vn

Nguyen Hong Son
Department of Science and Technology
Hanoi University of Industry

Nguyen Van Thien
School of Mechanical and Automotive Engineering
Hanoi University of Industry

***Abstract*—SAW is the oldest method among the multi-criteria decision-making (MCDM) approaches. On the other hand, RAM is known to be the newest method. TOPSIS is a highly renowned method and is the most widely used among MCDM methods. A question arises as to which method is deemed superior to the other two. The answer to this question is first found in this study. The selection of waterproofing materials is the problem used to compare the three aforementioned methods. The results indicated that RAM and TOPSIS are equally effective and superior to the SAW method.**

***Index Terms*—SAW method, RAM method, TOPSIS method, waterproofing material selection.**

I. Introduction

SELECTING of an option from among many alternatives is a common problem across all fields. To make a choice, various parameters (criteria) of the options must be evaluated. This means that choosing a particular option is a multi-criteria decision-making action [17]. Multi-criteria decision-making is carried out with the assistance of Multi-Criteria Decision-Making (MCDM) methods. There are over 200 different MCDM methods currently in use across various fields [18]. SAW is known to be the oldest method among MCDM approaches [9]. Despite having been around for a long time, its simplicity of application has kept it and its variations favored by scientists. The concept of SAW variations involves combining SAW with fuzzy theory to create the Fuzzy-SAW method for solving problems related to fuzzy sets, while fundamentally based on the original SAW method. In 2023, many studies continue to apply this method in various fields, such as selecting machining processes, milling processes, and evaluating indoor air quality [10], choosing rental cars [11], evaluating online learning platforms [12], selecting medical equipment suppliers [13], etc. RAM is the most recent MCDM method, introduced on September 7, 2023 [4]. According to the proponents of the RAM method, it overcomes the shortcomings of existing MCDM methods. The advantage lies in its ability to balance between beneficial and non-beneficial criteria. Overcoming the issue of reversal is also a strength of RAM [4]. Despite these mentioned advantages, due to its recent introduction, there have been no published studies on its application to date. TOPSIS is one of the most famous methods among MCDM approaches and is considered the most widely applied method [16]. In 2023, numerous studies have applied the TOPSIS method and its variations (Fuzzy-TOPSIS) in various fields, such as selecting businesses for mining investment [14], choosing solutions for grinding the surface of carbide cutting tools [15], selecting the defense strategy of the Serbian army [19], selecting logistics service providers [20], choosing locations for solar energy station construction [21], etc.

The analyses above lead to a question of which method—SAW, RAM, or TOPSIS—should be used. To decide which method to choose, a comparison of these methods is necessary and should be carried out initially. Unfortunately, such a comparison has not been conducted in any documented work. The objective of this article is to address this question. These three methods were simultaneously used to solve the problem of selecting waterproofing materials imported from Malaysia to Vietnam. Comparing the aforementioned three MCDM methods using a single weighting method for criteria may lead to biased conclusions. To achieve generalizable conclusions, the weights of the criteria have been determined using various methods. The summary of the steps for using the SAW, RAM, and TOPSIS methods will be presented in Chapter 2. Chapter 3 will summarize the steps for applying the weighting methods. The comparison of the three MCDM methods in selecting waterproofing materials will be discussed in Chapter 4. The final section of this article contains the scientific conclusions reached and directions for future research.

II. Multi-criteria decision-making methods used

A matrix with m rows and n columns will be established, where m is the number of alternatives to be ranked and n is the number of criteria used to describe each alternative. The value of criterion j for alternative i is denoted as xij, with i = 1 to m and j = 1 to n. Letters B and C are used to signify correspondingly that the higher the criterion, the better (B), and the lower the criterion, the better (C). The weight of criterion j is denoted as wj. The sequence of applying MCDM methods is as follows.

A. The SAW method

The sequence for ranking alternatives using the SAW method is as follows [3]:

Determine the normalized values using the following formula.

$$n_{ij}=\frac{x}{x_{ij}}, if\ j \in B \quad (1)$$

$$n_{ij}=\frac{x_{ij}}{x_{ij}}, if\ j \in C \quad (2)$$

The score Vi for each alternative is calculated using formula (3).

$$V_i=\sum_{j=1}^{n} w_j \cdot n_{ij} \quad (3)$$

The alternative with the highest Vi score is ranked 1. Conversely, the alternative with the lowest Vi score is ranked m.

B. The RAM method

To rank the alternatives using the RAM method, the following steps need to be carried out [4].

Normalize the data using formula (4).

$$n_{ij}=\frac{x_{ij}}{\sum_{i=1}^{m} x_{ij}} \quad (4)$$

Calculate the normalized values considering the weights of the criteria according to (5)

$$y_{ij}=w_j \cdot n_{ij} \quad (5)$$

Calculate the sum of normalized scores considering the weights of the criteria as per (6) and (7).

$$S_{+i}=\sum_{j=1}^{n} y_{+ij}, if\ j \in B \quad (6)$$

$$S_{-i}=\sum_{j=1}^{n} y_{-ij}, if\ j \in B \quad (7)$$

Calculate the score for each alternative according to (8).

$$RI_i=\sqrt[2+S_{-i}]{2+S_{+i}} \quad (8)$$

The alternative with the highest RIi score is ranked 1. Conversely, the alternative with the lowest RIi score is ranked m.

C. The TOPSIS Method

The TOPSIS method ranks alternatives in the following order [5]:

Determine the normalized values using formula (9).

$$n_{ij}=\frac{x_{ij}}{\sqrt{\sum_{i=1}^{n} x_{ij}^2}} \quad (9)$$

Calculate the normalized values considering the weights using formula (10).

$$y_{ij}=w_j . n_{ij} \quad (10)$$

Determine the best solution A^+ and the worst solution A^- for the criteria using the following two formulas.

$$A^+=\left\{y_1^+, y_2^+, \ldots, y_j^+, \ldots, y_n^+\right\} \quad (11)$$

$$A^-=\left\{y_1^-, y_2^-, \ldots, y_j^-, \ldots, y_n^-\right\} \quad (12)$$

Where: y_j^+ and y_j^- are the best and worst values of the normalized value y for criterion j.

Determine the values S_i^+ and S_i^- using the following two formulas.

$$S_i^+=\sqrt{\sum_{j=1}^{n}\left(y_{ij}-y_j^+\right)^2}, \quad i = 1, 2, \ldots, m \quad (13)$$

$$S_i^-=\sqrt{\sum_{j=1}^{n}\left(y_{ij}-y_j^-\right)^2}, \quad i = 1, 2, \ldots, m \quad (14)$$

Calculate the score C_i^* of the alternatives using formula (15).

$$C_i^*=\frac{S_i^-}{S_i^+ + S_i^-}, \quad i = 1, 2, \ldots, m; 0 \le C_i^* \le 1 \quad (15)$$

The alternative with the highest score is ranked 1, and the alternative with the lowest score is ranked m.

III. Used Weight Determination Methods

Three different methods were used in the article to calculate weights for the criteria, including the Equal method, the Entropy method, and the MEREC method. The Equal weight method was used due to its simplicity. The Entropy and MEREC methods were used because they are encouraged to be used [22].

Applying formula (16) to calculate the weights of the criteria using the Equal weight method [6].

$$w_j=\frac{1}{n} \quad (16)$$

The sequence for determining the weights of the criteria using the Entropy method is as follows [7]:

Determine the normalized values for the criteria using formula (17).

$$n_{ij}=\frac{x_{ij}}{m+\sum_{i=1}^{m} x_{ij}^2} \quad (17)$$

Calculate the Entropy measure for the criteria using formula (18).

$$e_j=\sum_{i=1}^{m}\left[n_{ij} \times \ln(n_{ij})\right] - \left(1 - \sum_{i=1}^{m} n_{ij}\right) \times \ln\left(1 - \sum_{i=1}^{m} n_{ij}\right) \quad (18)$$

Calculate the weights for the criteria using formula (19).

$$w_j = \frac{1-e_j}{\sum_{j=1}^{m}\left(1-e_j\right)} \tag{19}$$

The sequence for determining the weights for the criteria using the MEREC method is as follows [8]:

Calculate the normalized values using the following two formulas.

$$n_{ij} = \frac{min\, x_{ij}}{x_{ij}}, if\ j \in B \tag{20}$$

$$n_{ij} = \frac{x_{ij}}{max\, x_{ij}}, if\ j \in C \tag{21}$$

The values Si, S'ij, and Ej are calculated using the respective three formulas (22), (23), and (24)

$$S_i = \ln\left[1+\left(\frac{1}{n}\sum_{j}^{n}\left|\ln\left(n_{ij}\right)\right|\right)\right] \tag{22}$$

$$S_{ij}^{'} = \ln\left[1+\left(\frac{1}{n}\sum_{k,k\neq j}^{n}\left|\ln\left(n_{ij}\right)\right|\right)\right] \tag{23}$$

$$E_j = \sum_{i}^{m}\left|S_{ij}^{'}-S_i\right| \tag{24}$$

The weights for the criteria are determined using formula (25).

$$w_j = \frac{E_j}{\sum_{k}^{m} E_k} \tag{25}$$

IV. Waterproofing Material Selection

Vietnamese import a number of waterproofing materials from Malaysia, which have corresponding product codes: Solmax 440-900, Solmax 420-900, Solmax 480-900, Solmax 430-900, and Solmax 460-900. Many details about these products have been provided by the manufacturer, such as waterproofing capability, durability, flexibility, adhesion, chemical resistance, etc. There are several parameters with identical values across all product codes. Therefore, comparing options does not require consideration of those parameters. Only the parameters with varying values across the options need to be examined. Six technical parameters have been selected from the options, including average thickness, minimum thickness, tensile strength at flexure, tensile strength at break, tear strength, and puncture resistance. All six parameters fall under category B. Selecting a type of waterproofing material based solely on technical criteria would be a limitation. Procurement costs, processing costs, and time are factors that significantly impact both the economic and technical aspects of the project. Hence, factors related to processing time and processing costs should also be considered. A field survey identified three parameters: construction time, processing cost, and price. All three parameters are calculated per square meter of waterproofing material and fall under category C. Table 1 summarizes the data for the various options.

The Solmax 480-900 waterproofing material meets all six initial criteria and ranks highest compared to the other four remaining products. On the other hand, the Solmax 420-900 has the lowest values for all three criteria among the rest of the options. This necessitates the application of the MCDM technique to select the best waterproofing material. Firstly, determining the weights for the criteria is essential.

According to the Equal weight method, each criterion has an equal weight of 0.1111. When using the Entropy method, the normalized values calculated according to (17) have been synthesized in Table 2. The values Ej and weights wj were calculated using the respective formulas (18) and (19), and the results have been summarized in Table 3.

When using the MEREC method, the normalized values were calculated using the formulas (20) and (21), and the results are shown in Table 4.

Table 1. Types of waterproofing materials [1, 2]

Order	Average thickness (mm)	Minimum thickness (mm)	Tensile strength at flexure (kN/m)	Tensile strength at break (kNm)	Tear strength (N)	Puncture resistance (N)	Construction time (h)	Processing cost (Thousand Vietnamese dong)	Price (Thousand Vietnamese dong)
	C1	C2	C3	C4	C5	C6	C7	C8	C9
Solmax 440-900	1	0.9	15	28	130	355	0.37	34	272
Solmax 420-900	0.5	0.45	8	14	65	176	0.26	20	165
Solmax 480-900	2	1.8	31	57	250	705	0.43	42	366
Solmax 430-900	0.75	0.68	11	21	93	265	0.32	21	200
Solmax 460-900	1.5	1.35	23	43	187	540	0.39	36	285

Table 2. Normalized values in the Entropy weight method.

	C1	C2	C3	C4	C5	C6	C7	C8	C9
Solmax 440-900	0.0766	0.0780	0.0079	0.0043	0.0010	0.0003	0.0656	0.0067	0.0008
Solmax 420-900	0.0383	0.0390	0.0042	0.0021	0.0005	0.0002	0.0461	0.0040	0.0005
Solmax 480-900	0.1531	0.1560	0.0163	0.0087	0.0020	0.0007	0.0762	0.0083	0.0010
Solmax 430-900	0.0574	0.0589	0.0058	0.0032	0.0007	0.0003	0.0567	0.0041	0.0006
Solmax 460-900	0.1148	0.1170	0.0121	0.0066	0.0015	0.0005	0.0691	0.0071	0.0008

Table 3. Ej values and weights of the criteria calculated using the Entropy method.

	C1	C2	C3	C4	C5	C6	C7	C8	C9
Ej	-0.6968	-0.7049	-0.1661	-0.1049	-0.0323	-0.0135	-0.6056	-0.1234	-0.0224
wj	0.1479	0.1486	0.1017	0.0963	0.0900	0.0884	0.1400	0.0979	0.0891

Table 4. Normalized values in the MEREC weight method.

	C1	C2	C3	C4	C5	C6	C7	C8	C9
Solmax 440-900	0.5000	0.5000	0.5333	0.5000	0.5000	0.4958	0.8605	0.8095	0.7432
Solmax 420-900	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.6047	0.4762	0.4508
Solmax 480-900	0.2500	0.2500	0.2581	0.2456	0.2600	0.2496	1.0000	1.0000	1.0000
Solmax 430-900	0.6667	0.6618	0.7273	0.6667	0.6989	0.6642	0.7442	0.5000	0.5464
Solmax 460-900	0.3333	0.3333	0.3478	0.3256	0.3476	0.3259	0.9070	0.8571	0.7787

Table 5. Si and S'ij values in the MEREC weight method.

	Si	S'ij								
		C1	C2	C3	C4	C5	C6	C7	C8	C9
Solmax 440-900	0.4246	0.4377	0.4377	0.4353	0.4377	0.4377	0.438	0.4115	0.4152	0.4201
Solmax 420-900	0.2045	0.5457	0.5457	0.5457	0.5457	0.5457	0.5457	0.5708	0.5788	0.5804
Solmax 480-900	0.6515	0.3878	0.3878	0.3872	0.3881	0.387	0.3878	0.3296	0.3296	0.3296
Solmax 430-900	0.3602	0.4567	0.457	0.4524	0.4567	0.4544	0.4569	0.4512	0.4683	0.4651
Solmax 460-900	0.5788	0.3847	0.3847	0.3836	0.3853	0.3837	0.3853	0.3404	0.3443	0.3505

Table 6. Ej values and weights of the criteria calculated using the MEREC method.

	C1	C2	C3	C4	C5	C6	C7	C8	C9
Ej	0.9085	0.9089	0.9036	0.9076	0.9081	0.9084	1.0308	1.0483	1.0356
wj	0.1061	0.1062	0.1056	0.1060	0.1061	0.1061	0.1204	0.1225	0.1210

Table 7. Weights of the criteria.

Weight method	C1	C2	C3	C4	C5	C6	C7	C8	C9
Equal	0.1111	0.1111	0.1111	0.1111	0.1111	0.1111	0.1111	0.1111	0.1111
Entropy	0.1479	0.1486	0.1017	0.0963	0.0900	0.0884	0.1400	0.0979	0.0891
MEREC	0.1061	0.1062	0.1056	0.1060	0.1061	0.1061	0.1204	0.1225	0.1210

Table 8. Normalized values in the SAW method.

	C1	C2	C3	C4	C5	C6	C7	C8	C9
Solmax 440-900	0.5000	0.5000	0.4839	0.4912	0.5200	0.5035	0.7027	0.5882	0.6066
Solmax 420-900	0.2500	0.2500	0.2581	0.2456	0.2600	0.2496	1.0000	1.0000	1.0000
Solmax 480-900	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.6047	0.4762	0.4508
Solmax 430-900	0.3750	0.3778	0.3548	0.3684	0.3720	0.3759	0.8125	0.9524	0.8250
Solmax 460-900	0.7500	0.7500	0.7419	0.7544	0.7480	0.7660	0.6667	0.5556	0.5789

Table 9. Scores and rankings of the options using the SAW method.

	Equal weight		Entropy weight		MEREC weight	
	Vi	rank	Vi	rank	Vi	rank
Solmax 440-900	0.5440	3	0.5461	3	0.5480	4
Solmax 420-900	0.5015	5	0.4966	5	0.5243	5
Solmax 480-900	0.8369	1	0.8444	1	0.8218	1
Solmax 430-900	0.5349	4	0.5304	4	0.5501	3
Solmax 460-900	0.7013	2	0.7049	2	0.6966	2

Table 10. Normalized values in the RAM method.

	C1	C2	C3	C4	C5	C6	C7	C8	C9
Solmax 440-900	0.1739	0.1737	0.1705	0.1718	0.1793	0.1739	0.2090	0.2222	0.2112
Solmax 420-900	0.0870	0.0869	0.0909	0.0859	0.0897	0.0862	0.1469	0.1307	0.1281
Solmax 480-900	0.3478	0.3475	0.3523	0.3497	0.3448	0.3454	0.2429	0.2745	0.2842
Solmax 430-900	0.1304	0.1313	0.1250	0.1288	0.1283	0.1298	0.1808	0.1373	0.1553
Solmax 460-900	0.2609	0.2606	0.2614	0.2638	0.2579	0.2646	0.2203	0.2353	0.2213

Table 11. Normalized values considering the weights of the criteria in the RAM method.

	C1	C2	C3	C4	C5	C6	C7	C8	C9
Solmax 440-900	0.0193	0.0193	0.0189	0.0191	0.0199	0.0193	0.0232	0.0247	0.0235
Solmax 420-900	0.0097	0.0097	0.0101	0.0095	0.0100	0.0096	0.0163	0.0145	0.0142
Solmax 480-900	0.0386	0.0386	0.0391	0.0389	0.0383	0.0384	0.0270	0.0305	0.0316
Solmax 430-900	0.0145	0.0146	0.0139	0.0143	0.0143	0.0144	0.0201	0.0153	0.0173
Solmax 460-900	0.0290	0.0290	0.0290	0.0293	0.0287	0.0294	0.0245	0.0261	0.0246

Table 12. Some parameters in the RAM method and rankings of the options (when weights were calculated using the Equal weight method).

	S_{+i}	S_{-i}	RI_i	rank
Solmax 440-900	0.1159	0.0714	1.4360	3
Solmax 420-900	0.0585	0.0451	1.4234	5
Solmax 480-900	0.2319	0.0891	1.4686	1
Solmax 430-900	0.0860	0.0526	1.4307	4
Solmax 460-900	0.1744	0.0752	1.4540	2

The values of Si and S'ij were calculated using the respective formulas (22) and (23), and the results are shown in Table 5.

The Ej values and weights wj were calculated using the respective formulas (24) and (25), and the results have been synthesized in Table 6.

Weight determination for the criteria using three methods - Equal, Entropy, and MEREC - has been completed. In Table 7, the data from these calculations have been compiled.

The normalization of data for ranking the options using the SAW method was performed by applying formulas (1) and (2), and the results have been summarized in Table 8.

The Vi scores for the options were calculated using formula (3). These scores were used for ranking the options. These two steps were repeated three times corresponding to three different weight sets, and the results are presented in Table 9.

The normalization of data when using the RAM method for ranking the options was carried out by applying formula (4), and the results were compiled in Table 10.

The normalized values considering the weights of the criteria were calculated using formula (5). First, the weights of the criteria calculated using the Equal weight method were used, and the results are presented in Table 11.

The values S+i, S-i, and RIi were calculated using the respective formulas (6), (7), and (8). The RIi values were also used for ranking the options and are summarized in Table 12.

When the weights of the criteria were determined using the Entropy and MEREC methods, the ranking of the options using the RAM method was also carried out similarly. In Table 13, the RIi scores and rankings of the options are summarized for all three weight determination methods.

TABLE 13. SCORES AND RANKINGS OF THE OPTIONS USING THE RAM METHOD.

	Equal weight		Entropy weight		MEREC weight	
	RI_i	rank	RI_i	rank	RI_i	rank
Solmax 440-900	1.4360	3	1.4367	3	1.4326	3
Solmax 420-900	1.4234	5	1.4236	5	1.4215	5
Solmax 480-900	1.4686	1	1.4701	1	1.4631	1
Solmax 430-900	1.4307	4	1.4311	4	1.4282	4
Solmax 460-900	1.4540	2	1.4549	2	1.4496	2

TABLE 15. NORMALIZED VALUES CONSIDERING THE WEIGHTS OF THE CRITERIA IN THE TOPSIS METHOD.

	C1	C2	C3	C4	C5	C6	C7	C8	C9
Solmax 440-900	0.0391	0.0391	0.0382	0.0385	0.0405	0.0391	0.0512	0.0531	0.0506
Solmax 420-900	0.0196	0.0196	0.0204	0.0193	0.0202	0.0194	0.0360	0.0312	0.0307
Solmax 480-900	0.0783	0.0782	0.0790	0.0784	0.0779	0.0777	0.0595	0.0656	0.0681
Solmax 430-900	0.0293	0.0296	0.0280	0.0289	0.0290	0.0292	0.0443	0.0328	0.0372
Solmax 460-900	0.0587	0.0587	0.0586	0.0592	0.0582	0.0595	0.0540	0.0562	0.0530

TABLE 16. A+ AND A- VALUES IN TOPSIS.

	C1	C2	C3	C4	C5	C6	C7	C8	C9
A+	0.0783	0.0782	0.079	0.0784	0.0779	0.0777	0.036	0.0312	0.0307
A-	0.0196	0.0196	0.0204	0.0193	0.0202	0.0194	0.0595	0.0656	0.0681

TABLE 17. SOME PARAMETERS IN THE TOPSIS METHOD AND RANKINGS OF THE OPTIONS (WHEN WEIGHTS WERE CALCULATED USING THE EQUAL WEIGHT METHOD).

	Si+	Si-	Ci*	rank
Solmax 440-900	0.1015	0.0528	0.3420	3
Solmax 420-900	0.1433	0.0560	0.2809	5
Solmax 480-900	0.0560	0.1433	0.7191	1
Solmax 430-900	0.1211	0.0527	0.3034	4
Solmax 460-900	0.0610	0.0975	0.6154	2

TABLE 18. SCORES AND RANKINGS OF THE OPTIONS USING THE TOPSIS METHOD.

	Equal weight		Entropy weight		MEREC weight	
	RI_i	rank	RI_i	rank	RI_i	rank
Solmax 440-900	0.3420	3	0.3385	3	0.3448	3
Solmax 420-900	0.2809	5	0.2589	5	0.3091	5
Solmax 480-900	0.7191	1	0.7411	1	0.6909	1
Solmax 430-900	0.3034	4	0.2810	4	0.3278	4
Solmax 460-900	0.6154	2	0.6207	2	0.6031	2

When applying the TOPSIS method to rank the options, data normalization was conducted using formula (9), and the results are shown in Table 14.

The normalized values, considering the weights of the criteria, were calculated using (10). The weight set calculated using the Equal weight method was used first, and the results are summarized in Table 15.

The values A+ and A- were calculated using the respective formulas (11) and (12), and the results are shown in Table 16.

The values Si+, Si-, and Ci* were calculated using the respective formulas (13), (14), and (15). The ranking of the options was based on their Ci* scores. The results are shown in Table 17.

Ranking the options using the TOPSIS method when the weights of the criteria were determined using the Entropy and MEREC methods was also performed in a similar manner. In Table 18, the Ci* scores and rankings of the options are summarized for all three weight determination methods.

Table 19. Ranking of the options using different methods.

	SAW & Equal	SAW & Entropy	SAW & MEREC	RAM & Equal	RAM & Entropy	RAM & MEREC	TOPSIS & Equal	TOPSIS & Entropy	TOPSIS & MEREC
Solmax 440-900	3	3	4	3	3	3	3	3	3
Solmax 420-900	5	5	5	5	5	5	5	5	5
Solmax 480-900	1	1	1	1	1	1	1	1	1
Solmax 430-900	4	4	3	4	4	4	4	4	4
Solmax 460-900	2	2	2	2	2	2	2	2	2
S	0.9÷1			1			1		

Table 20. Ranking of the options after excluding Solmax 440-900.

	SAW & Equal	SAW & Entropy	SAW & MEREC	RAM & Equal	RAM & Entropy	RAM & MEREC	TOPSIS & Equal	TOPSIS & Entropy	TOPSIS & MEREC
Solmax 420-900	4	4	4	4	4	4	4	4	4
Solmax 480-900	1	1	1	1	1	1	1	1	1
Solmax 430-900	3	3	3	3	3	3	3	3	3
Solmax 460-900	2	2	2	2	2	2	2	2	2
S	1			1			1		

Table 21. Ranking of the options after excluding Solmax 420-900.

	SAW & Equal	SAW & Entropy	SAW & MEREC	RAM & Equal	RAM & Entropy	RAM & MEREC	TOPSIS & Equal	TOPSIS & Entropy	TOPSIS & MEREC
Solmax 440-900	4	3	4	3	3	3	4	4	4
Solmax 480-900	1	1	1	1	1	1	1	1	1
Solmax 430-900	3	4	3	4	4	4	3	3	3
Solmax 460-900	2	2	2	2	2	2	2	2	2
S	0.8÷1			1			1		

Thus, the ranking of the options using the three methods - SAW, RAM, and TOPSIS - has been completed. To facilitate result analysis, the data in Tables 9, 13, and 18 have been compiled in Table 19.

Observing Table 19, it can be seen that when using the RAM and TOPSIS methods, the rankings of the options are entirely consistent and independent of the weighting method used. These rankings also match exactly with the two cases of using the SAW method combined with the Equal weight method and when using the SAW method combined with the Entropy weight method. If the SAW method is combined with the MEREC weight method, the rankings of the options do not entirely match with the other combinations. This shows that the stability in ranking the options using the SAW method is slightly lower compared to using the RAM and TOPSIS methods.

The Spearman's rank correlation coefficient has also been used to analyze sensitivity [23-25], which is calculated using formula (26). Here, Di represents the rank difference of the options for a specific scenario compared to another scenario.

$$S = 1 - \frac{6\sum_{i=1}^{m} D_i^2}{m\left(m^2-1\right)} \tag{26}$$

The values of the Spearman coefficient have been calculated and placed in the last row of Table 19. It is observed that the Spearman coefficient is 1 when using both the RAM and TOPSIS methods. However, when using the SAW method, the coefficient ranges from 0.9 to 1. This further confirms the perception that ranking the options using the SAW method is slightly less stable compared to using the RAM and TOPSIS methods. Nevertheless, in all cases studied, Solmax 480-900 has been identified as the best waterproofing material.

TABLE 22. RANKING OF THE OPTIONS AFTER EXCLUDING SOLMAX 480-900.

	SAW & Equal	SAW & Entropy	SAW & MEREC	RAM & Equal	RAM & Entropy	RAM & MEREC	TOPSIS & Equal	TOPSIS & Entropy	TOPSIS & MEREC
Solmax 440-900	2	2	2	2	2	2	2	2	2
Solmax 420-900	4	4	4	4	4	4	4	4	4
Solmax 430-900	3	3	3	3	3	3	3	3	3
Solmax 460-900	1	1	1	1	1	1	1	1	1
S	1			1			1		

TABLE 23. RANKING OF THE OPTIONS AFTER EXCLUDING SOLMAX 430-900.

	SAW & Equal	SAW & Entropy	SAW & MEREC	RAM & Equal	RAM & Entropy	RAM & MEREC	TOPSIS & Equal	TOPSIS & Entropy	TOPSIS & MEREC
Solmax 440-900	3	3	3	3	3	3	3	3	3
Solmax 420-900	4	4	4	4	4	4	4	4	4
Solmax 480-900	1	1	1	1	1	1	1	1	1
Solmax 460-900	2	2	2	2	2	2	2	2	2
S	1			1			1		

TABLE 24. RANKING OF THE OPTIONS AFTER EXCLUDING SOLMAX 460-900.

	SAW & Equal	SAW & Entropy	SAW & MEREC	RAM & Equal	RAM & Entropy	RAM & MEREC	TOPSIS & Equal	TOPSIS & Entropy	TOPSIS & MEREC
Solmax 440-900	2	2	3	2	2	2	2	2	2
Solmax 420-900	4	4	4	4	4	4	4	4	4
Solmax 480-900	1	1	1	1	1	1	1	1	1
Solmax 430-900	3	3	2	3	3	3	3	3	3
S	0.8÷1			1			1		

To accurately conclude that the stability of ranking the options using the SAW method is slightly lower than using the other two methods, further investigation in different scenarios is necessary. Five different scenarios were created, and in each scenario, one option was excluded from the list. The rankings of the options in these five scenarios are presented in Tables 20 to 24. In each of these tables, the Spearman coefficient has also been calculated and placed in the last row of each table.

In all five scenarios mentioned above, when using both the RAM and TOPSIS methods, the Spearman coefficient is always 1. In this case, when using the SAW method, the Spearman coefficient ranges from 0.8 to 1. This once again affirms that the RAM and TOPSIS methods have very high correlation consistency in ranking the options and perform better than SAW method.

V. CONCLUSION

An investigation to compare three methods - SAW, RAM, and TOPSIS - has been conducted in this article. The comparison of these three methods was carried out in ranking various waterproofing materials imported to Vietnam from Malaysia. Some conclusions are drawn as follows:

A. The rankings of the options completely match when ranked using both the RAM and TOPSIS methods, regardless of the weighting method used.

B. When using the SAW method to rank the options, the rankings also completely match when usingeither the Equal or Entropy weight determination methods.

C. The option identified as the best does not depend on the MCDM method or the weighting method used.

D. The stability in ranking the options using the SAW method is slightly lower compared to using the RAM and TOPSIS

REFERENCES

[1] https://vngeo.com/thong-so-ky-thuat-mang-chong-tham-hdpe-solmax/
[2] https://vaidiakythuat.com/thong-so-ky-thuat-va-gia-ban-mang-chong-tham-hdpe-day-0-5mm.html
[3] S. Kusumadewi, S. Hartati, A. Harjoko, and R. Wardoyo, *Fuzzy Multi-Attribute Decision Making (FUZZY MADM)*, Yogyakarta: Penerbit Graha Ilmu, 2006.

[4] Alireza Sotoudeh-Anvari, Root Assessment Method (RAM): A novel multi-criteria decision making method and its applications in sustainability challenges, Journal of Cleaner Production 423 (2023) 138695, https://doi.org/10.1016/j.jclepro.2023.138695

[5] C. –L. Hwang, Y. –J. Lai, Ting_Yun Liu, *A new approach for multiple objective decision making*, Computers & Operations Research, Vol. 20, No. 8, pp. 889–899, 1993.

[6] R. M. Dawes, B. Coorigan, *Linear Models in DecisionMalking*, Psychological Bulletin, Vol. 81, pp. 95–106, 1974.

[7] YUXIN Z., DAZUO T., FENG Y., *Effectiveness of Entropy Weight Method in Decision-Making*, Mathematical Problems in Engineering, Vol. 2020, pp. 1–5, 2020.

[8] Keshavarz-Ghorabaee, M., Amiri, M., Zavadskas, E.K., Turskis, Z., Antucheviciene, J, *Determination of objective weights using a new method based on the removal effects of criteria (MEREC)*, Symmetry, Vol. 13, No. 4, pp. 1-20, 2021.

[9] Hamed Taherdoost, Analysis of Simple Additive Weighting Method (SAW) as a MultiAttribute Decision-Making Technique: A Step-by-Step Guide, Journal of Management Science & Engineering Research, Vol. 6. No. 1, 2023, 21-24, https://doi.org/10.30564/jmser.v6i1.5400

[10] Tran Van Dua, Combination of design of experiments and simple additive weighting methods: a new method for rapid multi-criteria decision making. EUREKA: Physics and Engineering, Vol. 2023, No. 1, 120–133, https://doi.org/10.21303/2461-4262.2023.002733

[11] Kamala Aliyeva, Aida Aliyeva, Rashad Aliyev, Mustafa Ozdeser, Application of Fuzzy Simple Additive Weighting Method in Group Decision-Making for Capital Investment, Axioms, Vol. 12, No. 797, 2023, 1-13, https://doi.org/10.3390/axioms12080797

[12] Ahmed E. Youssef, Kashif Saleem, A Hybrid MCDM Approach for Evaluating Web-Based E-Learning Platforms, IEEE Access, Vol. 11, 2023, 72436-72447, https://doi.org/10.1109/ACCESS.2023.3294798

[13] Neven Saleh, Mohamed N. Gaber, Mohamed A. Eldosoky, Ahmed M. Soliman, Vendor evaluation platform for acquisition of medical equipment based on multi-criteriadecision-making approach, Scientific reports, Vol. 13, No. 12746, 2023, https://doi.org/10.1038/s41598-023-38902-3

[14] Jiahui Su and Yidi Sun, An Improved TOPSIS Model Based on Cumulative Prospect Theory: Application to ESG Performance Evaluation of State-Owned Mining Enterprises, Sustainability, Vol. 15, No. 10046, 2023, 1-20, https://doi.org/10.3390/su151310046

[15] Kien Huy Nguyen, Dong Van Pham, Quoc Ve Tran, A multi-criteria decision-making in relieving grinding process of surface of gear milling tooth based on the archimedean spiral using TAGUCHI-AHP-TOPSIS method, EUREKA: Physics and Engineering, Vol. 2023, No. 4, 2023, 87–103, https://doi.org/10.21303/2461-4262.2023.002795

[16] Loubna Lamrini, Mohammed Chaouki Abounaima, Mohammed Talibi Alaoui, New distributed-topsis approach for multi-criteria decision-making problems in a big data context, Journal of Big Data, Vol. 10, No. 97, 2023, 1-21, https://doi.org/10.1186/s40537-023-00788-3

[17] Sushil Kumar Sahoo, Shankha Shubhra Goswami, A Comprehensive Review of Multiple Criteria Decision-Making (MCDM) Methods: Advancements, Applications, and Future Directions, Decision Making Advances, Vol.1, No. 1, 2023, 25-48, https://doi.org/10.31181/dma1120237

[18] Mahmut Baydaş, Tevfik Eren, Željko Stević, Vitomir Starčević, Raif Parlakkay, Proposal for an objective binary benchmarking framework that validates each other for comparing MCDM methods through data analytics, *PeerJ Computer Science*, Vol. 9, No. e1350, 2023, 1-24, https://doi.org/10.7717/peerj-cs.1350

[19] Dusko Tesic, Darko Bozanic, Dejan Stojkovic, Adis Puska, Ilija Stojanovic, DIBR–DOMBI–FUZZY MAIRCA Model for Strategy Selection in the System of Defense, Hindawi Discrete Dynamics in Nature and Society, Vol. 2023, No. 4961972, 2023, 1-14, https://doi.org/10.1155/2023/4961972

[20] Agnieszka TUBIS, Sylwia Werbinska-Wojciechowska, Fuzzy TOPSIS in selecting logistic handling operator: case study from POLAND, Transport, Vol. 38, No. 1, 2023, 12-30, https://doi.org/10.3846/transport.2023.17074

[21] Ibrahim badi, Ali Abdulshahed, Emhemed Alghazel, Using Grey-TOPSIS approach for solar farm location selection in Libya, Reports in Mechanical Engineering, Vol. 4, No. 1, 2023, 80-89, https://doi.org/10.31181/rme040129062023b

[22] Trung, D.D., Thinh, H.X., A multi-criteria decision-making in turning process using theMAIRCA, EAMR, MARCOS and TOPSIS methods: A comparative study, Advances in Production Engineering & Management, Vol. 16, No. 4, 2021, 443-456, https://doi.org/10.14743/apem2021.4.412

[23] Duc Trung Do, Application of FUCA method for multi-criteriadecision making in mechanical machining processes, Operational Research in Engineering Sciences: Theory and Applications, Vol. 5, Issue 3, 2022, pp. 131-152, https://doi.org/10.31181/oresta051022061d

[24] Mahmut Baydas, Comparison of the Performances of MCDM Methods under Uncertainty: An Analysis on Bist SME Industry Index, OPUS Journal of Society Research, Vol. 19, No. 46, 2021, 308-326, https://doi.org/10.26466//opusjsr.1064280

[25] Mahmut Baydas, Orhan Emre Elma, An objective criteria proposal for the comparison of mcdm and weighting methods in financial performance measurement: anapplication in BORSA ISTANBUL, Decision Making: Applications in Management and Engineering, Vol. 4, Issue 1, 2021, pp. 257-279, https://doi.org/10.31181/dmame210402257b

rice2024 Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering pp. 49–62
DOI: 10.15439/2024R28
ISSN 2300-5963 ACSIS, Vol. 42

A Survey on Sentiment Analysis in Tamil: Critical Analysis

S. Manoj
Alliance College of Engineering and Design
Bangalore, India
manoj632004s@gmail.com

Moumita Pal
Stanley College
Hyderabad, India
moumitafdp@gmail.com

***Abstract*—The review paper delves into the methodologies, techniques, and challenges specific to sentiment analysis and opinion extraction within the Tamil language. As the digital landscape continues to expand, the ability to comprehend sentiments and opinions expressed in Tamil across diverse online platforms has grown increasingly vital. The paper traces the evolution of sentiment analysis techniques tailored for Tamil, covering essential components such as feature extraction, lexicon creation, and the applications of various algorithms. Special attention is given to the distinct details of the Tamil language, encompassing its linguistic complexities, codeswitching, and the expression of sentiment in informal contexts. A critical analysis has been conducted to compare different models. Moreover, the review explores strategies for opinion extraction and provides insightful suggestions for potential areas for future research and development.**

***Index Terms*—Lexicon, Sentiment Analysis, Opinion Extraction, Transformers**

I. Introduction

SENTIMENT analysis, one of the most popular and significant subsets of Natural Language Processing (NLP), requires the analysis of textual data to find and classify sentiments as neutral, positive, or negative. Sentiment analysis tasks have significant implications across various domains, from market research and brand management to social and political analysis. Sentiment analysis in various languages is important as the amount of data and media in these languages continues to grow. This review paper aims to provide a comprehensive survey of methodologies, dataset preprocessing steps, challenges, and recent advancements in sentiment analysis and opinion extraction specific to the Tamil language. Tamil, being a Dravidian language, has its own distinctive syntactic and semantic complexities, posing unique challenges and opportunities in the field of NLP.

Opinion extraction in Tamil requires a nuanced understanding of subjective language and sentiment that depends on context. The review explores strategies for identifying and consolidating opinions from various of sources, including social media, online reviews, and forums, within the linguistic context of Tamil.

Furthermore, this review seeks to highlight new development and how they impact on the accuracy and precision of sentiment analysis task in Tamil. By integrating insights from existing research, this survey aims to provide a valuable resource for researchers, practitioners, and enthusiasts engaged in sentiment analysis and opinion extraction in the Tamil language. Different methodologies used across languages are compared which can serve as a guide for future developments in sentiment analysis tasks for Tamil. The analysis also explores potential areas for future advancements and aims to encourage further research and innovation in this rapidly evolving field, contributing to computational linguistics and natural language processing field of study.

II. Review of Literature

The survey discusses three fundamental sentiment analysis techniques: supervised learning, lexicon-based methods, and transformer-based approaches. Supervised learning involves training a model with labelled text data. Lexicon-based methods use a sentiment-scored word dictionary to assess text sentiment, while transformer-based approaches utilize pre-trained neural networks to understand context and deduce sentiment.

Sentiment analysis is well-established in English but poses challenges in Tamil and other resource-poor languages like Bengali, Malayalam, Kannada, and Telugu due to the scarcity of datasets and lexicons, compounded by the complexity and English code-mixing in these languages. Efforts are ongoing to enhance sentiment analysis capabilities in these languages.

The survey highlights specific studies aimed at improving sentiment analysis. In study [62], Sentiwordnet was enhanced for tweet classification by expanding the lexicon and training a classifier for polarity estimation. In study [29], domain-specific ontology was used to refine sentiment analysis, while [38] focused on lexicon expansion for Tamil through lexical similarity and rule-based analysis.

The Forum for Information Retrieval Evaluation (FIRE) hosts workshops to advance multilingual information access research, including sentiment analysis in code-mixed languages. These workshops have revealed insights into the effectiveness of various algorithms for sentiment analysis tasks. Researchers are employing diverse machine learning algorithms, such as Recurrent Neural Networks(RNN), transformer models, and genetic algorithms, for tasks like aspect-based sentiment analysis and word-level Natural language understanding, showcasing the potential of these algorithms to improve sentiment analysis system performance.

III. Dataset

The survey highlights the utilization of diverse datasets for sentiment analysis, each with unique preprocessing techniques and sources. Romanized Bangla and Bangla text samples, cited in studies [1] and [12], include 9,337 social media posts from platforms like YouTube, Twitter, and Facebook, with preprocessing that removes emoticons, hashtags, and proper nouns, and includes Part-Of-Speech(POS) tagging and manual sentiment categorization. The Bengali Cricket Commentary dataset, referenced in [2] and [25], comprises 2,489 Facebook comments, processed

by removing punctuation, digits, and stop words, and then tokenized and stemmed for a Bag of Words representation. An English corpus from Amazon Reviews, used in [12], contains over 68,356 reviews, with preprocessing that eliminates stop words and non-Bangla words, and adjusts for negation. The IMDB and Polarity Detection dataset, mentioned in [28], undergoes stop words, special characters, and URL removal, with additional lowercasing, stemming, and 10-fold cross-validation. Tamil SentiWordNet, discussed in [29] and [30], evolves from the English SentiWordNet 3.0 using various lexicons, classified into positive and negative sentiments. Code-mixed datasets for Dravidian languages, from studies [24], [60], [64], and [68], include YouTube comments processed to remove extraneous char- acters and symbols. FIRE datasets from 2020, 2021, and 2022, cited in [55], [56], [58], and [61], offer a rich source of bilingual and native texts in Malayalam-English, Tamil-English, and Kannada-English, with comprehensive preprocessing for analysis readiness. Each dataset's preprocessing is meticulously tailored to its linguistic features and format, ranging from simple cleaning to advanced tokenization and sentiment classification, providing a foundational basis for sentiment analysis research across various languages and contexts.

TABLE 1 SUMMARY OF SENTIMENT ANALYSIS INVESTIGATED ON DIFFERENT LANGUAGES

Title	Publication Year	Datasets used	Pre-processed steps
Sentiment Analysis on Bangla and Romanized Text using deep recurrent model [1]	2016	Bangla and Romanized Bangla text samples (Written in English) collected from product review pages, YouTube, Twitter, Facebook, and online news portals.	Remove emoticons, hashtags, proper nouns and applied POS tagging. Manually categorized into positive, negative and ambiguous samples
A Sentiment Classification in Bengali and Machine Translated English Corpus [2]	2019	Two Bengali datasets from cricket commentary and other from Drama review (scrapped from YouTube)	English comments were removed leaving only Bengali. Bengali corpora are converted to English using google machine translation. Class balancing using SMOTE. The dataset is stemmed and tokenized, Tokenized and applied term frequency-inverse document frequeny(tf-idf) method. Manual rating is given to each translation representing the accuracy from 1 to 5
Performing Sentiment Analysis in Bangla Microblog Posts [4]	2014	Bangla tweets using twitter API	Bangla lexicon translated to English and applied tokenization, normalization and POS tagging.
Multilingual Sentiment Analysis: An RNN-Based Frame- work for Limited Data [6]	2018	English reviews and restaurant reviews from Spanish, Dutch, Russian and Turkish	SentiWordNet lexicon used to obtain a positive and negative sentiment score and they're aggregated to classify review as positive or negative
Evaluation of Naıve Bayes and Support Vector Machines on Bangla Textual Movie Reviews [7]	2018	Texts from Bangla movie reviews sites, Facebook, websites, tweets and Movie Database (IMDb).	Punctuation characters, URLs, stop words emoticons were removed. Applied tokenization, stemming and vectorization.
fully Automatic Lexi- con Expansion for Do- main oriented Senti- ment Analysis dataset used in this paper [9]	2006	Japanese text reviews from the following domains Consumer electronics, Travel, Food, Books, Movies	Applied tokenization, stemming, stop words removal and POS tagging
Detecting Multilabel Sentiment and Emotions from Bangla YouTube Comments [11]	2018	Comments (Bangla and Romanized Bangla) obtained from YouTube video	Applied tokenization and Reduction of stop words, links, URLs, user tags and mentions from YouTube
Sentiment Mining from Bangla Data using Mutual Information [12]	2016	Product reviews collected from Amazon.	Removed stop words and Non-Bangla alphabetic words from translation. For sentences having presence of a negation word, a negation word is added before each word in the sentence
Opinion-Polarity Identification in Bengali [13]	2010	Bengali news corpus	Applied tokenization, stemming and POS tagging
Supervised Approach of Sentimentality Extraction from Bengali Facebook Status [15]	2016	User's status comments from Facebook	Manually tagging the Facebook data into positive and negative. Removed symbols like hashtags, websites URLs and applied stemming.
Opinion Mining and Analysis for Arabic Language [16]	2014	Arabic reviews and comments collected from different social media resources	Removed digits, punctuations, special symbols and non-letters. Applied normalization and tokenization

Title	Publication Year	Datasets used	Pre-processed steps
Sarcasm Detection followed by Sentiment Analysis for Bengali Language: Neural Network & Supervised Approach [17]	2023	News headlines from twitter containing English, Bengali and Romanised Bengali	Removing punctuations, non-alphabetical characters, stop words. Applied Stemming and Lemmatization, Word tokenization
Cross-Lingual Sentiment Analysis Without (Good) Translation [19]	2017	English reviews on Yelp, Chinese hotel reviews, Spanish billion-word corpus	Performed data cleaning, tokenization, stemming, normalization to remove diacritics.
Sentiment Analysis in Czech Social Media Using Supervised Machine Learning [20]	2013	Czech movie reviews and product review dataset	Used tokenization, POS tagging stemming, lemmatization and removed stop words
Aspect-Based Opinion Mining from Customer Reviews [21]	2016	Customer reviews from amazon	Removal of symbols and performed sentence splitting, lemmatization, POS tagging, dependency parsing and dependency analysis
Analysis and Tracking of Emotions in English and Bengali Texts: A Computational Approach [22]	2011	News stories and blog corpora in Bengali	Applied tokenization, stemming, lemmatization, POS tagging, stop word removal, named entity recognition
Datasets for Aspect- Based Sentiment Analysis in Bangla and Its Baseline Evaluation [23]	2018	Cricket dataset and Restaurant dataset	Punctuations, digits and stop words were removed. Performed manual data annotation and tokenization on both datasets and represented as bag of words(BOW)
Corpus creation for sentiment analysis in code mixed Tamil -English text [24]	2022	Tamil English mixed YouTube comments	If comment is fully written in Tamil or English it is discarded. Removed emoticons and applied sentence length filters (less than 5 words and more than 15-word sentences are removed) and performed data annotation
Tamil English language sentiment analysis system [25]	2016	Tamil user reviews from domains like books, DVDs and music	Opinionated words from dataset are extracted. Used google translate to turn Tamil comments to English.
Cross-Lingual Sentiment Analysis with Machine Translation [26]	2013	Turkish and English product review dataset	Google translate to translate English data to Turkish and performed manual annotation for the sentence polarity dataset.
Sentiment Analysis Using Machine Learning Techniques [28]	2017	IMDB movie review dataset and polarity dataset, tweets collected from Twitter	Removed Stop words, Numeric and special characters; Count Vectorization and tf-idf vectorization; 10-fold cross validation
Sentiment Analysis: An Approach for Analysing Tamil Movie Reviews Using Tamil Tweets [29]	2021	Tamil language movie tweets	Data cleaning – removal of retweets, URLs, special characters and applied stemming and tf-idf.
Towards Building a SentiWordNet for Tamil [30]	2016	English SentiWordNet 3.0, AFINN-111, Subjectivity Lexicon and Opinion Lexicon.	SentiWordNet and subjectivity lexicons are merged and filtered to strongly subjective data and removed duplicate words. AFINN-111 and Opinion Lexicon added to this. removed words with ambiguous sense
Rough Set Based Opinion Mining Tamil [31]	2017	Tamil product review dataset	Performed Sentence extraction, anaphora resolution. Applied tokenization, stemming, lemmatization and removal stop words.
Corpus Based Senti- ment Classification of Tamil Movie Tweets Using Syntactic Patterns [33]	2017	Tamil movie tweets	Removal of any external links, retweets, characters that repeat more than once and applied tokenization
Sentiment Analysis on Tamil Reviews as Products in Social Media Using Machine Learning Techniques: A Novel Study [35]	2020	Mobile phone user Tamil reviews from e-commerce websites	Performed tokenization, stemming, lemmatization, POS tagging, stop words removal.
Sentiment Extraction for Tamil Political Reviews[36]	2016	Political reviews are gathered from social media websites such as twitter and Facebook	Applied tokenization and POS tagging.

Title	Publication Year	Datasets used	Pre-processed steps
Analysing Sentiment in Tamil Tweets using Deep Neural Network [37]	2020	Tamil tweets data	Removal of symbols, special characters, dates, diacritics, punctuations and emoticons. Tweets converted into word embedding using word2vec.
Sentiment Lexicon Ex- pansion using Word2vec and fastText for Sentiment Prediction in Tamil Tweets [38]	2020	Data from movie review websites - Cineulagam, Filmibeat, Maalaimalar, Samayam, IndiaTimes, Behind-woods, as well as from Twitter, Facebook, Noolaham.com, and Ta.wikipedia.ord	removal of html tags, English words, repeated characters, symbols and emoticons. word embedding is created from this corpus using word2vec.
Sentiment Mining: An Approach for Bengali and Tamil Tweets [41]	2016	999 Bengali tweets and 1103 Tamil tweets	Performed Tokenization and texts converted lowercase. Removal of URLs, usernames, punctuations.
Sentiment Analysis of Tamil-English Codeswitched Text on Social Media Using Sub-Word Level LSTM [42]	2020	English and Tamil mixed comments from Facebook	Manual data annotation. splitting into train, validation and test set
Sentiment Analysis of Dravidian Code Mixed Data [43]	2021	Tamil and Malayalam code mixed dataset	Replace emojis with corresponding description in English. non-Tamil and non-Malayalam characters replaced with roman script representation. Applied tf-idf vectorization
Multilingual Senti- ment Analysis in Tamil, Malayalam and Kannada Code Mixed Social Media Posts using MBERT [44]	2021	Posts from YouTube of Tamil, Malayalam, Kannada code mixed languages	Removal of symbols, special characters, hashtags, punctuations, URLs, emojis and numerals. Texts converted to lowercase
Sentiment Analysis on Tamil Code Mixed Text using Bi LSTM [45]	2021	Tamil code mixed data from FIRE 2021	Removal of emojis, special characters, non- ASCII characters. Texts conversion to lowercase. Maximum size of message fixed to 30 to 70 characters. Applied Tokenization
A Study on the Perfor- mance of Supervised Algorithms for Classifi- cation in Sentiment Analysis [48]	2019	Twitter review data and US airline dataset	Conversion of texts to lowercase. Applied Tokenization, stemming, removal of stop words. filter tokens that exceed length of 15 and below 3.
Unsupervised Self Training for Sentiment Analysis of Code-Switched Data [50]	2021	Data set of four different languages - Hinglish(tweets), Spanglish (tweets), Tanglish (YouTube comments) and Malayalam – English (YouTube comments)	Removed URLs, special characters and use data embedding
Transformer based Sentiment Analysis in Dravidian Languages [52]	2021	FIRE-2021 dataset	Removed emojis and punctuation. Applied tokenization. All sequences are padded with same length.
Analyzing Sentiment in Indian Language Micro Text Using Re-current Neural Network [53]	2016	Twitter data in three languages—Tamil, Hindi, and Bengali, given by SAIL in 2015.	Tweet id is removed. labels of the tweet are merged along with the tweet
Sentiment Analysis in Tamil Texts: A study on Machine Learning Techniques and Feature Representation [54]	2019	Tamil texts from twitter, YouTube, Facebook on topics such as movie, product, news, sports and tv shows	Removal of URLs, hashtags and non-Tamil words. Applied Tokenization and vectorization using fastText, tf-idf and BOW.
Sentiment Analysis and Homophobia de- tection of YouTube comments in Code-Mixed Dravidian Lan- guages using machine learning and Trans- former models [55]	2022	FIRE 2022 dataset	Executed tokenization and data cleaning. Removed URLs, numerals, and tags. Data embed- ded using tf-idf, count vectorizer, and XLM, MPNet, BERT.
An ensemble-based model for sentiment analysis of Dravidian code-mixed social media posts [56]	2021	FIRE 2021 dataset	Extracting tf-idf features using 1–6-gram characters.
Sentiment Analysis in Tamil Language Using Hybrid Deep Learning Approach [57]	2022	Ratings and reviews of Tamil movies from Kaggle.	Performed data cleaning, removed stop words, punctuations, and special characters. Transform the given data(multiclass) into binary class data (into positive and negative)

Title	Publication Year	Datasets used	Pre-processed steps
Sentiment Classifica- tion of Code-Mixed Tweets using Bi-Directional RNN and Language Tags [58]	2021	English-Tamil code-mixed data from FIRE 2020	Eliminating references, Removing the punctua- tions, taking off URLs, removing excess white space, extracting words from hashtags and applied data embedding using fastText.
Sentiment Analysis on Code-Switched Dra- vidian Languages with Kernel Based Extreme Learning Machines [60]	2022	YouTube comments using three codemixed datasets	Removing stop words and emoticons, lemmatizing. the pre written labels for the data is altered. word embedding applied using fastText.
Deep Learning Based Sentiment Analysis for Malayalam, Tamil and Kannada Languages [61]	2021	FIRE2021 dataset	Removal of special Characters, emojis, URLs, and hashtags. Applied tokenization, stop word removal, word embedding and post padding
Hateful Sentiment De- tection in Real-Time Tweets: An LSTM-Based Comparative Approach [65]	2023	Scraped twitter comments	Remove symbol, punctuations. Text conversion to lowercase. Applied tokenization and tf- idf vectorization
PANDAS@TamilNLPAC L2022: Abusive Com- ment Detection in Tamil Code-Mixed Data Using Custom Embeddings with LaBSE [66]	2022	Tamil English YouTube comments	Performed text normalisation, removal of punctuations, extra unwanted characters and stop words. Feature extraction using tf-idf and LaBSE.
A Computational Approach to the Analysis and Generation of Emotion in Text [69]	2011	The corpus contains 815,494 blog posts from LiveJournal	Feature Extraction using Bag of words and sentiment orientation

IV. Methodology

A. Supervised algorithms

Supervised algorithms have been majorly used in sentiment classification tasks. Support vector machine (SVM) is opted in by most researchers as shown in Table 2, most of the time, texts are linearly separable which allows for faster processing and fewer parameters to optimize. SVM views the problem as a pattern-matching task that involves learning symbolic patterns that depend on a phrase's lexical and syntactic semantics [13]. SVM with Sequential minimal optimization (SMO) used in [18], [26] and [5] is known for its scalability that is to perform consistently in large datasets. SVM SMO system has been used to develop aspect-based, Word-level and documents-level sentiment analysis systems.

Ensemble models used in the survey are listed in Table 3. Decision tree classifier provides a hierarchical decomposition of the data space in which a condition on the attribute value is used to divide the data [10].

The research that has used naïve bayes model is listed and analyzed in Table 4, the model computes the posterior probability of a class, depending upon the distribution of the words in the dataset. The model works with the BOWs feature extraction which ignores the position of the word in the document [10]. Multinomial naive bayes [12], MNB uses multinomial distribution for all pairs where it uses the word counts and rectify the underlying calculations to act within.

[4], [10], [20], [26], [28] and [34] have used max entropy classifier, The Maximum entropy Classifier transforms labelled feature sets into vectors. Now, by combining the weights that are computed for each feature, the most likely label for a feature set can be found. [10].

[7], [23] and [25] have used K-Nearest Neighbor (KNN) classifier. KNN performs classification by finding k nearest (in Euclidean distance) data objects repetitively with trial and error in classifying, until it the data points are finally classified, and majority vote determines final classification. [3], [19], [24], [43] and [51] used logistic regression, it works by fitting a sigmoid function to the training data that outputs 0 and 1. It is popular for its simplicity and easy interpretation, and it generally achieves good accuracy in various datasets. It is relatively light weight and can be deployed with minimum resources.

[28] Linear Discriminant Analysis(LDA), using a discriminant analysis technique, LDA classifies reviews by representing the dependent variables as a linear combination of the independent variables. This approach focuses on creating a linear combination of the dependent variable based on the independent variables. After that, these linear equations will be processed to produce the necessary categorization outcome [28].

B. Genetic Algorithm Based model

In Genetic Algorithms (GA), the text reviews are classified after being depicted as chromosomes. While applying Neuro GA, GA is employed to select the best features from a large pool of features. Subsequently, neural network is used to classify the reviews based on the selected features. The papers that have opted for GA based model are listed in Table 5.

C. Fuzzy classifier

[8], [35] and [40] used fuzzy classifier. Sentiment polarity is vague about its conceptual reach. There is not a clear boundary between the concepts of "+ve", "-ve" and "neutral". To better handle such fuzziness in sentiment polarity, fuzzy set classifiers are used.

D. Roughset based classifier

Rough set theory-based classification, the fundamental aim of rough set analysis is to derive upper and lower approximations from the available data. This theory assigns a level of affiliation to each object, with a central focus on resolving ambiguity that stems from distinguishing objects within a specific domain [31]. The analysis and comparison of rough set-based classifiers identified in the survey are presented in Table 6.

E. Lexicon mased models

The lexicon-based model identified in the survey is presented in Table 7. The Lexicon based approach uses predefined dictionaries and assigned manually annotated sentiment scores. The sentiment is founded from aggregating sentiment scores of all words in each data and determines overall sentiment of the data. Lexicon based approach provides rich linguistic information which helps in improving accuracy and requires less processing time compared to supervised learning method.

F. Recurrent Neural Network models

Long short-term memory (LSTM) is an extension of simple RNN which reduces vanishing gradient problem and can remember dependencies over a large gap [1]. A bidirectional long short-term memory (BLSTM) processes the input sequence in both forward and backward directions, with both directions feeding into the same output layer. So, one-layer processes the input in one direction while the other LSTM layer processes the sequence in the opposite direction. [43] used sub word level LSTM model as it accounts for words that have a similar morpheme. For example, in the Tamil dataset, 'aval', 'avanga' and 'avala' have similar meanings

TABLE 2 CRITICAL ANALYSIS ON SVM BASED MODEL

Title	Model	Result	Critical Analysis
Performing Sentiment Analysis in Bangla Microblog Posts [4]	SVM and Maximum Entropy classifiers	SVM scores highest with accuracy of 93%	SVM has been majorly preferred for sentiment analysis tasks and it mostly scores better accuracy than other supervised algorithms. These experiments that have primarily used the SVM algorithm shows that results of the Tamil dataset didn't reach high accuracy as much as other languages and had scored poorly in code mixed data[24]. In other cases, SVM has performed very well, and it has the worked best with n-gram features.
Evaluation of Naive Bayes and Support Vector Machines on Bangla Textual Movie Reviews [7]	Naive Bayes (NB) and Support Vector Machines (SVM)	SVM performed slightly better than NB with precision of 0.86.	
Aspect Level Opinion Mining on Customer Reviews using Support Vector Machine [14]	SVM	precision is calculated as 83.34%, recall is calculated as 92.87% and F-measure is calculated as 87.34%.	
Comparative experiments using super- vised learning and machine translation for multilingual sentiment analysis [18]	Translate using translators of google, Moses and Bing to German, French and Spanish and trained using SVM SMO classifier, naive bayes and Random Forest(RF).	The better results were obtained with the SVM SMO classifier in most languages	
Datasets for Aspect- Based Sentiment Analysis in Bangla and Its Baseline Eval- uation [23]	SVM, KNN, Random Forest	SVM obtained the highest precision rate for both of the datasets 0.71 and 0.77 for cricket and restaurant dataset respectively.	
Sentiment Analysis Using Machine Learning Techniques [28]	Naive Bayes, Support Vector Machine, Maximum Entropy (ME), and Stochastic Gradient Descent (SGD)	SVM showed highest accuracy of 88.94% with unigram + bigram + trigram features	
Multilingual Sentiment Analysis using Machine Translation [5]	Classifiers used are SVM SMO, adaboost and bagging classifier on each language with unigram and bigram features	A low quality of the translation led to features extracted being not informative enough thus performing with less accuracy of average accuracy 0.564.	

due to their root word 'aval'. Further evaluation and comparison of the RNN model are provided in Table 8.

G. Transformer models

A Transformer based models are state of art models in sentiment analysis tasks were BERT, known for its deep contextual representation, can be expanded by incorporating a classification head to refine the model for downstream NLP tasks. it's primarily used in [24], [46] and [52]. RoBERTA was trained using an approach called Masked Language Modelling (MLM). Through this method, the model can grasp the connections between words in a sentence and comprehend the contextual meaning of the text. [32], [47], [49], [50] and [52] have used XLM ROBERTa, an unsupervised cross- lingual representation approach, was trained on Wikipedia data of 100 languages and fine-tuned on different downstream tasks for evaluation and inference. This involves samples from text sources in a variety of languages extracted, and the model is then trained to predict masked tokens in the input. [44], [49] have used multilingual BERT (MBERT), it has been pretrained in 104 languages with largest Wikipediaes.

Distil BERT is used in [46], [49], [52], which is 60% faster than BERT and includes 40% less parameter. It employs a triple loss language modelling approach, integrating cosine distance loss with the process of knowledge distillation. When compared to MLM loss, the two distillation losses in the triple loss exert a substantial influence on the model's performance [52]. In [49] and [46] character BERT have been used, it reduces the complexity and removes word piece tokenization entirely and instead employs a Character-CNN to represent entire words at the character level rather than at a sub-word level [46]. MuRIL stands as an Indic language model, having undergone extensive training and enhancements to excel in Indian languages. It provides support English and 16 other Indian languages. MuRIL surpassed multilingual BERT on all benchmark data sets of Indic languages in [52]. The papers that have adopted transformer-based models are listed and examined in Table 9.

TABLE 3 CRITICAL ANALYSIS ON ENSEMBLE MODEL

Title	Model	Result	Critical Analysis
Sentiment Analysis Using Machine Learning Techniques [28]	Naive bayes, SVM, random forest and Linear Discriminant Analysis(LDA)	random forest scored highest accuracy in both datasets of 88.88% in IMDB dataset and 95% in polarity dataset	The performance of the tree based model highly varies depending on the features used for the classification task and has surpassed the popular SVM in some cases. Ensemble models have been fairly well tested in Tamil and have shown promising results. The model can be used as a baseline for future research in sentiment analysis tasks
Sentiment Mining an Approach for Bengali and Tamil Tweets [41]	Features extracted: tf-idf, score of unigrams and bigram, tweets specific features - emoticons, hashtags. classifiers used are naive bayes and decision tree classifier	Bengali Tweets: Naive Bayes: (+ve = 0.52, -ve = 0.76, neutral = 0.79); Decision Tree: (+ve = 0.52, -ve = 0.88, neutral = 0.81) Tamil Tweets: Naive Bayes: (+ve = 0.51, -ve = 0.78, neutral = 0.73) Decision Tree: (+ve = 0.50, -ve = 0.82, neutral= 0.77)	
Sentiment Analysis andHomophobia detection of YouTube comments in Code-Mixed Dravidian Languages using machine learning and Transformer models [55]	SVM, Multilayer Perceptron (MLP), random forest classifier, Ada boost, Gradient Boosting, and Extratrees classifiers have been used.	For SA task in Tamil- English Count Vectorizer with the Random Forest model fetched the best F1-score of 0.61.	
Sentiment Analysis on Tamil Reviews as Products in Social Media Using Machine Learning Techniques: A Novel Study [35]	Decision Tree, naïve bayes, NBTree, Rough set, Fuzzy rough set, SVM, Fuzzy SVM, Rough Fuzzy SVM, bagging (random forest), stacking (LDA, KNN, SVM), stacking (C5.0, CART(Classification and Regression Tree), RF)	Bagging and stacking algorithms show accuracy of 91%, Rough Fuzzy SVM show 87% and Decision tree shows 81% in 5 class analysis	
sentiment classification of online consumer reviews using word vector representations [3]	classifiers used are SVM, naive bayes, logistic regression (LR), random forest	random Forest outperforms all the algorithms when used with word2vec representations with 90.21% accuracy	
Corpus creation for sentiment analysis in code mixed tamil-english text [24]	algorithms used for classifying polarities are LR, naïve bayes, decision tree, random forest, SVM, dynamic meta embedding, conv-LSTM and BERT	LR, Random Forest and decision tree performed fairly better with bothscoring the same f-score of 0.68	

TABLE 4 CRITICAL ANALYSIS ON BAYES MODEL

Title	Model	Result	Critical analysis
Cross-Lingual Sentiment Analysis with Machine Translation [26]	Classifiers used were SVM SMO and naive bayes classifier with n-gram features	Naive bayes shows higher accuracy with unigram, bigram features of 75.57%	Naive Bayes algorithms have achieved satisfactory results for opinion extraction across various languages. However, they haven't outclassed SVMs in this domain. Both methods exhibit comparable performance levels. Notably, the observed accuracy for Tamil falls short of that achieved in languages like Bengali and English. This discrepancy might cause from limited resources for Tamil language processing.
Predicting the Sentimental Reviews in Tamil Movie using Machine Learning Algorithms [34]	Classifiers used were SVM, naive bayes, Maxent Classifier, decision tree. features: The punctuations and apostrophe, TamilSentiwordnet	SVM gives an accuracy of 75.9% performing better than other methods	
Sentiment Mining from Bangla Data using Mutual Information [12]	Multinomial Naive Bayes (MNB)	For English, using testing data 85.1% accuracy without using negation and 85.8% accuracy with negation. For Bangla dataset, using testing data 84.78% accuracy without using negation and 83.77% accuracy with negation	
Sentiment Mining: An Approach for Bengali and Tamil Tweets [41]	Naive bayes and decision tree classifier	Bengali Tweets: Naive Bayes: (+ve = 0.52, -ve = 0.76, neutral = 0.79) Decision Tree: (+ve = 0.52, -ve = 0.88, neutral= 0.81) Tamil Tweets: Naive Bayes: (+ve = 0.51, -ve = 0.78, neutral= 0.73) Decision Tree: (+ve = 0.50, -ve = 0.82, neutral= 0.77)	
Machine Learning Technique to Detect and Classify Mental Illness on social media Using Lexicon- BasedRecommender System [63]	SVM and naive bayes	Results show that SVM model could better classify the genre of film with 65.73% accuracy	

TABLE 5 CRITICAL ANALYSIS ON GA BASED MODEL

Title	Model	Result	Critical analysis
Findings of the Shared Task on Offensive Language Identification in Tamil, Malayalam, and Kannada [47]	Tamil and Malayalam dataset used genetic algorithm technique. Kannada dataset used ensemble of mBERT and XLMRoBERTa models	obtained F1-score of Malayalam dataset is 0.97 obtained F1-score of Tamil dataset is 0.78 obtained F1-score of Kannada dataset is 0.75	The GA-based approach has not performed quite well for Tamil-English mixed data compared to English, Bengali, and Malayalam. The feature selection for this classification task has been a crucial aspect that impacts the results.A notable difference of the GA based classifier from other statistical systems is its ability to encode a whole sentence in GA and use it as a feature whereas in most classifier systems, the n-gram method has been followed[8].
Opinion Extraction and Summarization from Text Docu- ments in Bengali [8]	(i) Rule based approach, CRF (conditional random field) based approach, hybrid approach and GA based technique (ii) SVM classifier with features	(i) the GA based approach gives an accuracy of 90.22 and 90.6 for English and Bengali corpus respectively (ii) The model gives an accuracy of 70.04% with all 7 features	
Sentiment Analysis Using Machine Learning Techniques [28]	Using Genetic Algorithm and classification using KNN algorithm and Neuro-Genetic Algorithm	accuracy of proposed approach GA = 0.93, NeuroGA = 0.963	

H. Clustering models

Clustering functions by grouping similar unlabeled data, eliminating the need for extracting supervised informational features. There are two types of clustering: hierarchical and partition. Models for both types of clustering have been tested in [28] and the analysis is presented in Table 10.

TABLE 6 CRITICAL ANALYSIS ON ROUGH SET THEORY-BASED MODEL

Title	Model	result	critical analysis
Rough Set Based Opinion Mining Tamil [31]	Rough set theory-based Classification	The algorithm achieved accuracy of 0.99, 0.80, 0.92, 0.99 and 0.93 for most positive, positive, neutral, most negative, negative respectively	Although the method yielded good results, it has fallen behind comparing with results of other models. Rough set-based feature selection offers a more computationally efficient approach to selecting feature so it can be utilised as a hybrid model alongside other algorithm
Sentiment Analysis on Tamil Reviews as Proucts in Social Media Using Machine Learning Techniques: A Novel Study [35]	Decision Tree, naïve Bayes & NBTree, Rough set, Fuzzy rough set, SVM, Fuzzy SVM, Rough Fuzzy SVM, bagging (random forest), stacking LDA, KNN, SVM), stacking (C5.0, CART, RF)	Bagging and stacking algorithms show accuracy of 91%, Rough Fuzzy SVM show 87% and Decision tree shows 81% in 5 class analysis	

TABLE 7 CRITICAL ANALYSIS ON LEXICON BASED MODEL

Title	Model	Result	Critical analysis
fully Automatic Lexicon Expansion for Domain oriented Sentiment Analysis [9]	Lexicon based approach- sentence delimitation; proposition detection; polarity assignment	The precision of polarity assignment using the automatically acquired lexicon averaged 94%	While lexicon-based approaches offer an alternative to supervised methods, their effectiveness has been found limited in Tamil. where lexical resources are still evolving, the algorithm is fully dependent on quality and variety of words in the lexicon. Dictionary based approach takes less processing time than supervised learning approach, but their accuracy often falls short [10].
A Survey on Sentiment Analysis Algorithms for Opinion Mining [10]	(i) Supervised techniques: Decision tree classifier; SVM; rule-based classifier; naive bayes; max entropy (ii) Dictionary-based approach	Supervised techniques provide better accuracy compared to dictionary based approach	
Opinion extraction from online blogs and public reviews [27]	Lexicon repositories used are Senti- wordnet, wordnet, slang is used for polarity classification	Proposed method achieves an average accuracy of 79% on word level and 81% at sentence level	
Sentiment Analysis: An Approach for Analysing Tamil Movie R views Using Tamil Tweets [29]	Lexicon apparoach - TamilWordNet (TWN) and Tamil SentiWordNet (TSWN).	The proposed model given the best accuracy of 77.89%	

TABLE 8 CRITICAL ANALYSIS ON RNN BASED MODEL

Title	Model	Result	Critical analysis
Sentiment Analysis on Bangla and Romanized Text using deep recurrent model [1]	LSTM with word2vec	Bangla dataset attaining highest accuracy of 70%	RNN is specially used for serialized data making it ideal for sentiment analysis tasks. The RNN model, Bi-LSTM which has an advantage of understanding the meaning of the sentence in bi-directional propagation mechanism [57]. from the survey it looks that RNN model performed better than stand alone CNN model and has shown really high accuracy for Tamil than in other results [61]. The hybrid version of CNN- BiLSTM has shown leading results and finds potential for further tests [57].
Multilingual Sentiment Analysis: An RNN- Based Framework for Limited Data [6]	Lexicon based baseline model and RNN in all four datasets	Results showed that the RNN model outperforms in all four datasets	
Detecting Multilabel Sentiment and Emo- tions from Bangla YouTube Comments [11]	LSTM method, CNN method and Baseline using – SVM, NB classifiers	LSTM model performs slightly better than CNN. The highest achievable accuracy for 3 and 5 class sentiment analysis is 65.97% and 54.24%.	
Analyzing Sentiment in Tamil Tweets using Deep Neural Network [37]	Deep bi directional LSTM model with n-gram feature	The model scored an accuracy of 86.2%	
Sentiment Analysis of Online Tamil Con- tents using Recursive Neural Network Models Approach for Tamil Language [39]	RNN model using a binary tree model	The model scored an accuracy of 71.1% in long phrases, 73% in intra sentential negation and 70.8% in inter sentential Negation	

TABLE 8 CRITICAL ANALYSIS ON RNN BASED MODEL (CONTINUATION)

Title	Model	Result	Critical analysis
Sentiment Analysis of Tamil-English Code-Switched Text on Social Media Using Sub-Word Level LSTM [42]	6-layer RNN model including convolutional layer and LSTM layers	the proposed sub word level LSTM model was recorded an accuracy of 75%	
Sentiment Analysis on Tamil Code Mixed Text using Bi LSTM [45]	The datasets are Embedded using GLoVe and passed to bidirectional LSTM model	The framework gives an f1-score of 0.552	
Hateful Sentiment Detection in Real-Time Tweets: An LSTM-Based Comparative Approach [65]	Long-short term memory (LSTM)	The accuracy score was found to be 97%	
Analyzing Senti- ment in Indian Languages Micro Text Using Recurrent Neural Network [53]	Simple RNN model	F-score obtained by the RNN model in Tamil, Hindi and Bengali are 88.23,72.01 and 65.16 respectively	
Sentiment Analysis in Tamil Language Using Hybrid Deep Learning Approach [57]	feature extraction using fastText models used to classify - CNN-LSTM, CNN-BiLSTM and CNN- BiGRU	CNN-BiLSTM has achieved the higher accuracy of 80.2% and highest f1-score of 0.64	
Sentiment Classifi- cation of Code-Mixed Tweets using Bi-Directional RNN and Language Tags [58]	Bi-Directional LSTM model	The performance of the developed algorithm, garnered precision, re- call, and F1 scores of 0.59, 0.66, and 0.58 respectively	
Deep Learning Based Sentiment Analysis for Malayalam, Tamil and Kannada Languages [61]	Model-1: Convolutional network with LSTM Model-2: Bi-directional LSTM model Model-3: contains an Embedding layer, a Flatten, a hidden and, a Dense layer.	For Malayalam – English the best performance was given by Model-2 of accuracy 0.9482. For Kannada - English the best performance was given by Model-3 of accuracy 0.9896. For Tamil - English the best performance was given by Model-3 of accuracy 0.9905	

TABLE 9 CRITICAL ANALYSIS ON TRANSFORMER MODEL

Title	Model	Result	Critical Analysis
Multilingual Senti- ment Analysis in Tamil, Malayalam and Kannada Code Mixed Social Media Posts using MBERT [44]	MBERT model	the precision, recall and weighted f1 score for Tamil are 0.59, 0.60 and 0.60 respectively. The precision, recall and weighted f1 score for Kannada is 0.61, 0.61 and 0.61 respectively. The precision, recall and weighted f1 score for Malayalam are 0.72, 0.72and 0.72 respectively.	Transformers, a pretrained unsupervised approach, Numerous studies have been conducted to integrate deep learning and machine learning models to achieve optimal sentiment analysis. These models now are being widely used and models such as BERT, Distil- BERT, and fast-Text show very decent performance, yet there remains room for improvement and fine-tuning for each language. Transformer-based models have shown competitive performance compared to supervised methods[50]. Class imbalance significantly impacts the model's performance in low-support classes[52].
Findings of the Sentiment Analysis of Dravidian Languages in Code-Mixed Text [47]	XLM−RoBERTa	Tamil English, Malayalam-English and Kannada-English scored weighted average F1−score of 0.711, 0.804, and 0.630, respectively	
Unsupervised Self- Training for Senti- ment Analysis of Code-Switched Data [50]	RoBERTa	Hinglish showed f1 score of 0.32 and accuracy 0.36; Spanglish showed f1 score of 0.31 and accuracy 0.32; Tanglish showed f1 score of 0.15 and accuracy 0.16; Malayalam- English showed f1 score of 0.17 and accuracy 0.14	

TABLE 9 CRITICAL ANALYSIS ON TRANSFORMER MODEL (CONTINUATION)

Title	Model	Result	Critical Analysis
Transformer based Sentiment Analysis in Dravidian Languages [52]	MuRIL, vBERT, XLM-RoBERTa, DistilBERT	Using soft voting technique the average F1- Score are 0.708, 0.626, and 0.609 in Malayalam, Tamil, and Kannada respectively	
sentiment Analysis on Dravidian Code- Mixed YouTube Comments using Paraphrase XLM- RoBERTa Model [59]	XLM-RoBERTa model	the model on Tamil, Malayalam, and Kannada code-Mixed language datasets, and achieve F1-scores of 71.1, 75.3, and 62.5 respectively.	
Overview of Abusive Comment Detection in Tamil - ACL 2022 [64]	ML algorithms - Logistic Regression, Linear Support Vector Machines, Gradient Boost classifier, and KNN classifier. Deep learning algorithm - Multilayered perceptron, Vanilla LSTM, Recurrent Neural Networks (RNN) Transformers - mBERT, MuRIL BERT, XLM RoBERTa, and ULMFit.	MuRIL BERT model have shown the best performance the highest F-score of 0.41.	
NLP-CUET @DravidianLang- Tech-EACL2021: Offensive Language Detection from Multilingual Code- Mixed Text using Transformers [68]	SVM, LR, ensemble, LSTM with fastText, LSTM with word2vec and LSTM with attention, MBERT, indic-BERT, XLM-R	Transformer based models show best results for all languages: Tamil F1-score 0.76 by XLM-R, Malayalam F1-score 0.93 by XLM-R and Kannada F1-score 0.71 by M-BERT	

TABLE 10 CRITICAL ANALYSIS ON CLUSTERING MODEL

Title	Model	Result	Critical analysis
Sentiment Analysis Using Machine Learning Techniques [28]	Clustering method(unsupervised) − K means, mini batch K means, Affinity Propagation, and DBSCAN	DBSCAN performed the best with 0.95 adjusted rand index	Due to the semantic complexities involved, unsupervised methods are not widely used in sentiment analysis. For resource-poor languages like Tamil, it is early to expect unsupervised algorithms to perform well.

V. CHALLENGES IN SENTIMENT ANALYSIS

Sentiment analysis faces several challenges related to dataset and model development. The breakdown of some key challenges is the following

A. Dataset Oriented Challenges

Sentiment analysis models rely heavily on training data. However, language evolves, and the sentiment associated with words can shift over time. Training data from a specific period might not accurately reflect current sentiment. This can lead to models misinterpreting the emotions expressed in newer data.

The meaning of a word can be highly contextual and domain dependent. A model trained on general language data might struggle with domain-specific sentiment. For example, the word ‘rock’ in context of mu- sic refers to a genre but when used in casual sentence like ‘she rocks’, the word means to be amazing or great.

Sentiment analysis models need to effectively handle negation (e.g., "not," "no"). Negation can completely flip the sentiment of a sentence. Models require careful feature selection and processing techniques to accurately identify negation and its impact on sentiment.

Sentiment analysis across languages presents unique challenges. Sarcasm, humour, and even positive/negative words can vary significantly due to cultural differences. Furthermore, languages like Tamil, which often mix with English, require understanding the lexical nuances of both languages for accurate sentiment analysis. This increases computational complexity and can lead to ambiguity in some cases.

Sentiment can be conveyed through non-linguistic cues like emojis, hashtags, and capitalization. Integrating these cues into sentiment analysis models requires additional processing power and learning capacity. Models need to be efficient in handling both linguistic and non-linguistic features to provide a comprehensive sentiment analysis.

B. Algorithm Oriented Challenges

Text data needs to be converted into vectors for machine learning algorithms to process it. Choosing the right vectorization technique and feature selection methods significantly impacts the model's ability to capture sentiment-bearing information from the text.

For languages with limited sentiment analysis resources like the Tamil, translation to a well-resourced language might be necessary. Translation accuracy is crucial. Any loss of meaning or sentiment during translation can negatively impact the model's performance.

Sentiment in a sentence can be influenced by words far apart from each other. Traditional sentiment analysis models

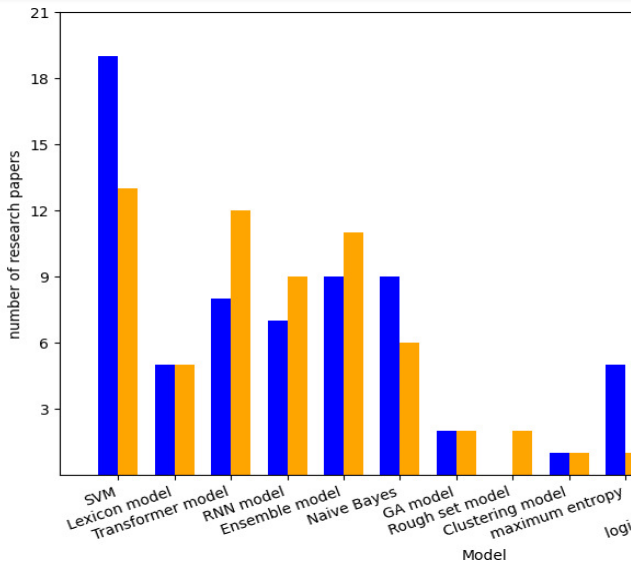


Fig. 1 comparison graph of number of papers that have used the respective model to perform SA task for Tamil and other languages

may struggle to capture these long-range dependencies. Techniques like recurrent neural net- works (RNNs) and transformers are more effective at modelling these relationships.

Sentiment data often exhibits class imbalance to a majority class outweighing the other. This imbalance can bias the model towards the majority class, leading to inaccurate classification of neutral sentiment.

VI. Conclusion

This survey has analysed sentiment analysis research in Tamil and other languages, comparing their performance. The findings reveal that supervised methods, LSTM, and transformer-based models generally outperform other approaches as discussed in the critical analysis. However, their results in Tamil lag behind those in other languages due to the imbalanced distribution of Tamil datasets and the complexities of the language, including prevalent code-mixed data.

Algorithms like lexicon-based and clustering methods are highly dependent on corpus quality. Existing datasets for Tamil sentiment analysis are outdated and lack the sophistication required for effective benchmarking. Figure 1 highlights the areas explored in Tamil sentiment analysis so far.

While transformer models have shown moderate performance in Tamil [50], they remain state-of-the-art and, hold potential for further exploration. Research has primarily focused on a few popular datasets, so future studies should venture into less explored areas. Hybrid models combining transformers with algorithms like GA could improve accuracy.

Improvements in Tamil lexicons, such as adding more words and focusing on adverbs, could enhance lexicon-based methods. Further research on linguistic models is crucial to better capture contextual and domain-specific nuances, paving the way for advancements in Tamil sentiment analysis.

References

[1] Hassan, Asif, Mohammad Rashedul Amin, Abul Kalam Al Azad, and Nabeel Moham- med. ”Sentiment analysis on bangla and romanized bangla text using deep recurrent models.” In 2016 International Workshop on Computational Intelligence (IWCI), pp. 51- 56. IEEE, 2016.

[2] Sazzed, Salim, and Sampath Jayarathna. ”A sentiment classification in bengali and machine translated english corpus.” In 2019 IEEE 20th international conference on infor- mation reuse and integration for data science (IRI), pp. 107-114. IEEE, 2019.

[3] Bansal, Barkha, and Sangeet Srivastava. ”Sentiment classification of online consumer reviews using word vector representations.” Procedia computer science 132 (2018): 1147-1153.

[4] Chowdhury, Shaika, and Wasifa Chowdhury. ”Performing sentiment analysis in Bangla microblog posts.” In 2014 International Conference on Informatics, Electronics & Vision (ICIEV), pp. 1-6. IEEE, 2014.

[5] Balahur, Alexandra, and Marco Turchi. “Multilingual sentiment analysis using ma- chine translation?.” In Proceedings of the 3rd workshop in computational approaches to subjectivity and sentiment analysis, pp. 52-60. 2012.

[6] Can, E. F., A. Ezen-Can, and F. Can. “Multilingual sentiment analysis: an RNN-based framework for limited data (2018).” arXiv preprint arXiv:1806.04511.

[7] Banik, Nayan, and Md Hasan Hafizur Rahman. “Evaluation of naive bayes and support vector machines on bangla textual movie reviews.” In 2018 international conference on Bangla speech and language processing (ICBSLP), pp. 1-6. IEEE, 2018.

[8] Das, A. M. I. T. A. V. A. “Opinion Extraction and Summarization from Text Documents in Bengali.” Kolkata, India (2011).

[9] Kanayama, Hiroshi, and Tetsuya Nasukawa. ”Fully automatic lexicon expansion for domainoriented sentiment analysis.” In Proceedings of the 2006 conference on empirical methods in natural language processing, pp. 355-363. 2006.

[10] Pradhan, Vidisha M., Jay Vala, and Prem Balani. ”A survey on sentiment analysis al- gorithms for opinion mining.” International Journal of Computer Applications 133, no. 9 (2016): 7-11.

[11] Tripto, Nafis Irtiza, and Mohammed Eunus Ali. ”Detecting multilabel sentiment and emotions from bangla youtube comments.” In 2018 International Conference on Bangla Speech and Language Processing (ICBSLP), pp. 1-6. IEEE, 2018.

[12] Paul, Animesh Kumar, and Pintu Chandra Shill. ”Sentiment mining from bangla data using mutual information.” In 2016 2nd international conference on electrical, computer & telecommunication engineering (ICECTE), pp. 1-4. IEEE, 2016.

[13] Das, Amitava, and Sivaji Bandyopadhyay. ”Opinion-polarity identification in bengali.” In International conference on computer processing of oriental languages, pp. 169-182. California, USA: Chinese and Oriental Languages Computer Society, 2010.

[14] Joshi, Anju, and Anubhooti Papola. ”Aspect Level Opinion Mining on Customer Re- views using Support Vector Machine.” International

Journal of Advanced Research in Computer and Communication Engineering (2017).
[15] Islam, Md Saiful, Md Ashiqul Islam, Md Afjal Hossain, and Jagoth Jyoti Dey. "Super- vised approach of sentimentality extraction from bengali facebook status." In 2016 19th international conference on computer and information technology (ICCIT), pp. 383-387. IEEE, 2016.
[16] Al-Kabi, Mohammed N., Amal H. Gigieh, Izzat M. Alsmadi, Heider A. Wahsheh, and Mohamad M. Haidar. "Opinion mining and analysis for Arabic language." IJACSA) Inter- national Journal of Advanced Computer Science and Applications 5, no. 5 2014: 181-195.
[17] Pal, Moumita, and Rajesh Prasad. "Sarcasm Detection followed by Sentiment Analysis for Bengali Language: Neural Network & Supervised Approach." In 2023 International Conference on Advances in Intelligent Computing and Applications (AICAPS), pp. 1-7. IEEE, 2023.
[18] Balahur, Alexandra, and Marco Turchi. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. Computer Speech & Language 28, no. 1 2014.
[19] Abdalla, Mohamed, and Graeme Hirst. "Cross-lingual sentiment analysis without (good) translation." arXiv preprint arXiv:1707.01626 (2017).
[20] Habernal, Ivan, Tom´a˘s Pt´a˘cek, and Josef Steinberger. "Sentiment analysis in czech social media using supervised machine learning." In Proceedings of the 4th workshop on computational approaches to subjectivity, sentiment and social media analysis, pp. 65- 74. 2013.
[21] Samha, Amani Khalaf. "Aspect-based opinion mining from customer reviews." PhD diss., Queensland University of Technology, 2016.
[22] Das, Dipankar. "Analysis and tracking of emotions in english and bengali texts: a com- putational approach." In Proceedings of the 20th international conference companion on World wide web, pp. 343-348. 2011.
[23] Rahman, Md Atikur, and Emon Kumar Dey. "Datasets for aspect-based sentiment analysis in bangla and its baseline evaluation." Data 3, no. 2 (2018): 15.
[24] Chakravarthi, Bharathi Raja, Vigneshwaran Muralidaran, Ruba Priyadharshini, and John P. McCrae. "Corpus creation for sentiment analysis in code-mixed Tamil-English text." arXiv preprint arXiv:2006.00206 (2020).
[25] Thilagavathi, R., and K. Krishnakumari. "Tamil english language sentiment analysis system." International Journal of Engineering Research & Technology (IJERT) 4, no. 16 (2016).
[26] Demirtas, Erkin. "Cross-lingual sentiment analysis with machine translation." (2013).
[27] Asghar, Muhammad Zubair. "Opinion Extraction From Online Blogs And Public Re- views." PhD diss., GOMAL UNIVERSITY DI KHAN, 2014.
[28] Tripathy, Abinash. "Sentiment Analysis Using Machine Learning Techniques." PhD diss., 2017.
[29] Ramanathan, Vallikannu, T. Meyyappan, and S. M. Thamarai. "Sentiment analysis: an approach for analysing tamil movie reviews using Tamil tweets." Recent Advances in Mathematical Research and Computer Science 3 (2021): 28-39.
[30] Kannan, Abishek, Gaurav Mohanty, and Radhika Mamidi. "Towards building a Senti- WordNet for Tamil." In Proceedings of the 13th International Conference on Natural Lan- guage Processing, pp. 30-35. 2016.
[31] Sharmista, Ramaswami, and M. Ramaswami. "Rough set based opinion mining in Tamil." International Journal of Engineering Research and Development (2017).
[32] Sean, Benhur. "Findings of the shared task on Emotion Analysis in Tamil." In Proceed- ings of the Second Workshop on Speech and Language Technologies for Dravidian Lan- guages, pp. 279-285. 2022.
[33] Ravishankar, Nadana, and Shriram Raghunathan. "Corpus based sentiment classifica- tion of tamil movie tweets using syntactic patterns." IIOAB Journal: A Journal of Multi- disciplinary Science and Technology 8, no. 2 (2017): 172-178.
[34] Se, Shriya, R. Vinayakumar, M. Anand Kumar, and K. P. Soman. "Predicting the senti- mental reviews in tamil movie using machine learning algorithms." Indian journal of sci- ence and technology 9, no. 45 (2016): 1-5.
[35] Sharmista, A., and Dr M. Ramaswami. "Sentiment Analysis on Tamil Reviews as Prod- ucts in Social Media Using Machine Learning Techniques: A Novel Study." Madurai Kama- raj University Madurai-625 21 (2020).
[36] Anish, D., and V. Sumathy. "Sentiment Extraction for Tamil Political reviews" (2016).
[37] Anbukkarasi, S., and S. Varadhaganapathy. "Analyzing sentiment in Tamil tweets us- ing deep neural network." In 2020 Fourth International Conference on Computing Meth- odologies and Communication (ICCMC), pp. 449-453. IEEE, 2020.
[38] Thavareesan, Sajeetha, and Sinnathamby Mahesan. "Sentiment lexicon expansion using Word2vec and fastText for sentiment prediction in Tamil texts." In 2020 Moratuwa engineering research conference (MERCon), pp. 272-276. IEEE, 2020.
[39] Padmamala, R., and V. Prema. "Sentiment analysis of online Tamil contents using re- cursive neural network models approach for Tamil language." In 2017 IEEE International conference on smart technologies and management for computing, communication, controls, energy and materials (ICSTM), pp. 28-31. IEEE, 2017.
[40] Mouthami, K., K. Nirmala Devi, and V. Murali Bhaskaran. "Sentiment analysis and classification based on textual reviews." In 2013 international conference on Information communication and embedded systems (ICICES), pp. 271-276. IEEE, 2013.
[41] Prasad, Sudha Shanker, Jitendra Kumar, Dinesh Kumar Prabhakar, and Sachin Tripa- thi. "Sentiment mining: An approach for Bengali and Tamil tweets." In 2016 Ninth Inter- national Conference on Contemporary Computing (IC3), pp. 1-4. IEEE, 2016.
[42] Raveendirarasa, Vidyapiratha, and C. R. J. Amalraj. "Sentiment analysis of tamil-eng- lish codeswitched text on social media using subword level lstm." In 2020 5th Interna- tional Conference on Information Technology Research (ICITR), pp. 1-5. IEEE, 2020.
[43] Mandalam, Asrita Venkata, and Yashvardhan Sharma. "Sentiment analysis of Dravid- ian code mixed data." In Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pp. 46-54. 2021.
[44] Kalaivani, Adaikkan, and Durairaj Thenmozhi. "Multilingual Sentiment Analysis in Tamil Malayalam and Kannada code-mixed social media posts using MBERT." In FIRE (Working Notes), pp. 1020-1028. 2021.
[45] Roy, Pradeep Kumar, and Abhinav Kumar. "Sentiment Analysis on Tamil Code-Mixed Text using Bi-LSTM." In Working Notes of FIRE 2021-Forum for Information Retrieval Eval- uation (Online). CEUR. 2021.
[46] Chakravarthi, Bharathi Raja, Ruba Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P. McCrae. "Dravidiancodemix: Sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text." Language Resources and Evaluation 56, no. 3 (2022): 765-806.
[47] Chakravarthi, Bharathi Raja, Ruba Priyadharshini, Sajeetha Thavareesan, Dhivya Chin- nappa, Durairaj Thenmozhi, Elizabeth Sherly, John P. McCrae et al. "Findings of the sen- timent analysis of dravidian languages in code-mixed text." arXiv preprint arXiv:2111.09811 (2021).
[48] Sunitha, P. B., Shelbi Joseph, and P. V. Akhil. "A study on the performance of super- vised algorithms for classification in sentiment analysis." In TENCON 2019-2019 IEEE Re- gion 10 Conference (TENCON), pp. 1351-1356. IEEE, 2019.
[49] Hande, Adeep, Siddhanth U. Hegde, Ruba Priyadharshini, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Sajeetha Thavareesan, and Bharathi Raja Chakravarthi. "Benchmarking multi-task learning for sentiment analysis and offensive language identi- fication in under-resourced dravidian languages." arXiv preprint arXiv:2108.03867 (2021).
[50] Gupta, Akshat, Sargam Menghani, Sai Krishna Rallabandi, and Alan W. Black. "Unsu- pervised self-training for sentiment analysis of code-switched data." arXiv preprint arXiv:2103.14797 (2021).
[51] Srinivasan, R., and C. N. Subalalitha. "Sentimental analysis from imbalanced code- mixed data using machine learning approaches." Distributed and Parallel Databases (2021): 1-16.
[52] Jada, Pawan Kalyan, D. Sashidhar Reddy, Konthala Yasaswini, Arunaggiri Pandian K, Prabakaran Chandran, Anbukkarasi Sampath, and Sathiyaraj Thangasamy. "Transformer based Sentiment Analysis in Dravidian Languages." In FIRE (Working Notes), pp. 926-938. 2021
[53] Seshadri, Shriya, Anand Kumar Madasamy, Soman Kotti Padannayil, and M. Anand Kumar. "Analyzing sentiment in indian languages micro text using recurrent neural net- work." IIOAB J 7 (2016): 313-318
[54] Thavareesan, Sajeetha, and Sinnathamby Mahesan. "Sentiment analysis in Tamil texts: A study on machine learning techniques and feature representation." In 2019 14th Conference on industrial and information systems (ICIIS), pp. 320-325. IEEE, 2019.

[55] Varsha, Josephine, B. Bharathi, and A. Meenakshi. ”Sentiment Analysis and Homo- phobia detection of YouTube comments in Code-Mixed Dravidian Languages using ma- chine learning and transformer models.” In Working Notes of FIRE 2022-Forum for Infor- mation Retrieval Evaluation (Hybrid). CEUR. 2022.

[56] Kumar, Abhinav, Sunil Saumya, and Jyoti Prakash Singh. ”An ensemble-based model for sentiment analysis of Dravidian code-mixed social media posts.” In Working Notes of FIRE 2021-Forum for Information Retrieval Evaluation (Online). CEUR. 2021.

[57] Ramesh Babu, Suba Sri. ”Sentiment Analysis In Tamil Language Using Hybrid Deep Learning Approach.” PhD diss., Dublin, National College of Ireland, 2022.

[58] Mahata, Sainik, Dipankar Das, and Sivaji Bandyopadhyay. ”Sentiment classification of codemixed tweets using bi-directional rnn and language tags.” In Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages, pp. 28-35. 2021.

[59] Babu, Yandrapati Prakash, and Rajagopal Eswari. ”Sentiment Analysis on Dravidian CodeMixed YouTube Comments using Paraphrase XLM-RoBERTa Model.” Working Notes of FIRE (2021).

[60] SR, Mithun Kumar, Lov Kumar, and Aruna Malapati. “Sentiment Analysis on Code- Switched Dravidian Languages with Kernel Based Extreme Learning Machines.” In Pro- ceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages, pp. 184-190. 2022.

[61] Pavan Kumar, P. H. V., B. Premjith, J. P. Sanjanasri, and K. P. Soman. “Deep Learning Based Sentiment Analysis for Malayalam, Tamil and Kannada Languages.” (2021).

[62] Bravo-Marquez, Felipe. “Acquiring and exploiting lexical knowledge for twitter senti- ment analysis.” PhD diss., University of Waikato, 2017.

[63] Sumathy, B., Anand Kumar, D. Sungeetha, Arshad Hashmi, Ankur Saxena, Piyush Ku- mar Shukla, and Stephen Jeswinde Nuagah. ”Machine Learning Technique to Detect and Classify Mental Illness on Social Media Using Lexicon-Based Recommender System.” Computational Intelligence and Neuroscience 2022 (2022).

[64] Priyadharshini, Ruba, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U. Hegde, and Prasanna Kumaresan. ”Overview of abusive comment detection in Tamil-ACL 2022.” In Proceed- ings of the Second Workshop on Speech and Language Technologies for Dravidian Lan- guages, pp. 292-298. 2022.

[65] Roy, Sanjiban Sekhar, Akash Roy, Pijush Samui, Mostafa Gandomi, and Amir H. Gan- domi. ”Hateful Sentiment Detection in Real-Time Tweets: An LSTM-Based Comparative Approach.” IEEE Transactions on Computational Social Systems (2023).

[66] Swaminathan, Krithika, K. Divyasri, G. L. Gayathri, Thenmozhi Durairaj, and B. Bhara- thi. ”PANDAS@ Abusive Comment Detection in Tamil Code-Mixed Data Using Custom Embeddings with LaBSE.” In Proceedings of the Second Workshop on Speech and Lan- guage Technologies for Dravidian Languages, pp. 112-119. 2022.

[67] Chakravarthi, Bharathi Raja, Ruba Priyadharshini, Navya Jose, Thomas Mandl, Prasanna Kumar Kumaresan, Rahul Ponnusamy, R. L. Hariharan, John Philip McCrae, and Elizabeth Sherly. “Findings of the shared task on offensive language identification in Tamil, Malayalam, and Kannada.” In Proceedings of the first workshop on speech and language technologies for Dravidian languages, pp. 133-145. 2021.

[68] Sharif, Omar, Eftekhar Hossain, and Mohammed Moshiul Hoque. “Nlp-cuet@ dravid- ianlangtech-eacl2021: Offensive language detection from multilingual code-mixed text using transformers.” arXiv preprint arXiv:2103.00455 (2021).

[69] Keshtkar, Fazel. A computational approach to the analysis and generation of emotion in text. University of Ottawa (Canada), 2011

rice2024 Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering pp. 63–69
DOI: 10.15439/2024R109
ISSN 2300-5963 ACSIS, Vol. 42

Integrating Computational Advertising with Guaranteed Display for Enhanced Performance in Wi-Fi Marketing

Bach Pham Ngoc, Linh Nguyen Duy, Bao Bui Quoc, Nhat Nguyen Hoang *
Faculty of Mathematics and Informatics
Hanoi University of Science and Technology
Hanoi, Vietnam
bachpn.hust@gmail.com, duyljnh@gmail.com, buiquocbao121198@gmail.com, hoangnhatdb@gmail.com

***Abstract*—Wi-Fi Marketing effectively engages potential customers by displaying advertisements before granting internet access through public or business Wi-Fi hotspots. However, the increasing number of advertising campaigns complicates resource allocation, necessitating optimized ad placement to achieve campaign goals while minimizing disruptions to the user experience. This paper examines principles of efficient resource allocation in Wi-Fi Marketing, focusing on fairness, demand optimization, and user satisfaction. We propose an allocation model that formalizes advertising contracts between advertisers and publishers managing Wi-Fi infrastructure, incorporating ad supply, ad requests, and contractual terms. The model employs an objective function to balance fairness, penalize unmet requirements, and maximize user engagement, while adhering to constraints such as minimum impressions and resource limits. Additionally, we introduce mathematical formulations to strategically distribute advertisements, ensuring quota fulfillment and catering to diverse audience segments. The proposed framework not only enhances campaign performance but also maintains a seamless and positive user experience. By implementing these principles and the proposed model, Wi-Fi Marketing can effectively manage resource allocation complexities, thereby maximizing the impact of advertising efforts.**

***Index Terms*—Wifi Marketing; Guaranteed Display Advertising; Digital Advertising and Ad Scheduling; Computational Advertising; Campaign Performance Maximization; Resource Allocation in Computer Networks.**

I. Introduction

WI-FI marketing has emerged as a powerful tool for businesses to monetize public Wi-Fi hotspots through advertising. Venue owners can offer premium paid access or advertising-sponsored free access to users [1]. This model is formulated as a three-stage Stackelberg game, where the ad platform's revenue-sharing policy affects Wi-Fi pricing but not advertising pricing. The effectiveness of this approach depends on factors like advertising concentration and user visiting frequency. To optimize ad scheduling, researchers have proposed algorithms that consider Wi-Fi communication constraints and user tolerance for viewing ads [2]. These algorithms aim to maximize user interest in displayed advertisements, potentially increasing revenue for venues and advertisers. Additionally, resource allocation challenges in Wi-Fi networks can be addressed using techniques like the decomposition algorithm to ensure fair distribution of ad impressions across targeted locations.

Wi-Fi marketing presents challenges in resource allocation, fairness, and user satisfaction. Researchers have proposed various approaches to address these issues. Son Anh Ta et al. [3] introduced a Dantzig-Wolfe decomposition algorithm to optimize fairness in ad allocation across targeted locations. Wanru Xu et al. [2] developed a greedy swap algorithm to maximize user interest in Wi-Fi advertisements while considering communication constraints and user tolerance. M. Bateni et al. [4] proposed a stochastic approximation scheme for fair resource allocation in dynamic marketplaces, achieving a balance between seller revenues and buyer fairness. Haoran Yu et al. [1] presented a Wi-Fi monetization model offering premium and advertising-sponsored access, analyzing the economic interactions among stakeholders as a three-stage Stackelberg game. These studies collectively address the multifaceted problem of effective advertising resource allocation in Wi-Fi marketing, considering fairness, demand optimization, and user satisfaction to enhance the effectiveness of this marketing channel.

This paper presents a comprehensive framework for addressing resource allocation challenges in Wi-Fi Marketing. The proposed model formalizes the contractual relationships between advertisers and Wi-Fi publishers, incorporating essential elements such as ad supply, ad requests, and predefined contractual terms. By employing a robust objective function, the model seeks to balance fairness, penalize unmet requirements, and optimize user engagement while adhering to constraints such as minimum impression quotas and resource limitations.

Furthermore, the framework introduces mathematical formulations for the strategic distribution of advertisements, ensuring diverse audience segments are effectively targeted and contractual quotas are fulfilled. These formulations are designed to maximize campaign performance without com-

* Corresponding author

promising user experience, thus addressing the dual priorities of advertisers and publishers.

By integrating these principles, the proposed allocation model offers a systematic solution to the complexities of resource management in Wi-Fi Marketing. This approach not only enhances the efficiency and impact of advertising efforts but also upholds a positive and uninterrupted user experience, cementing Wi-Fi Marketing as a sustainable and effective marketing strategy.

II. Related Works

A. Guaranteed Display Advertising

Guaranteed Display Advertising (GDA) is a crucial model in online advertising, allowing advertisers to secure a predetermined number of impressions for target audiences. Recent research has focused on optimizing GDA planning and allocation strategies. Turner [6] formulated the GDA planning problem as a transportation problem with a quadratic objective, developing algorithms for solving large-scale problems. Subsequent studies have proposed adaptive frameworks to improve both contract delivery and user engagement. Cheng et al. [7] introduced an adaptive unified allocation framework that optimizes contract delivery and user interest simultaneously. Fang et al. [8] developed a personalized delivery system that accounts for individual-level constraints and user-ad interactions. Dai et al. [9] proposed a fairness-aware allocation model that balances guaranteed delivery, impression quality, and traffic cost. These advancements have led to significant improvements in contract delivery rates, click-through rates, and overall advertising revenue for e-commerce platforms (Cheng et al., [7]; Fang et al., [8]; Dai et al., [9]).

B. Resource Allocation in Computer Networks

Resource allocation optimization in wireless networks is a crucial area of research for improving system performance. Various approaches have been explored, including cross-layer multiuser optimization (Zhu Han & K. J. R. Liu [10]) and utility-based resource management frameworks (Song & Li [11]). These methods aim to efficiently allocate resources for diverse traffic types with different QoS requirements. In the context of Wi-Fi marketing, a novel mathematical model has been proposed to optimize fairness in campaign allocation across targeted locations (Ta Anh Son et al. [3]). The Dantzig-Wolfe decomposition algorithm is suggested as an effective solution for this large-scale problem. Recent advancements include the integration of deep reinforcement learning with a multi-objective framework to develop periodic product recommendation systems, enabling resource optimization tailored to user preferences and system constraints [12]. Moreover, techniques such as optimistic linear support and user clustering have been combined with multi-objective reinforcement learning to build multi-objective periodic recommendation systems, further enhancing resource allocation efficiency [13]. Other techniques, such as power control, multiple access, and dynamic resource allocation, have also been studied for wireless resource allocation optimization (M. Mehrjoo et al. [14]; Zhu Han & K. J. R. Liu, [10]). These approaches collectively contribute to enhancing the performance of wireless systems, including Wi-Fi networks, and improving resource allocation strategies for various applications.

C. Digital Advertising and Ad Scheduling

Recent research on targeted advertisement distribution and scheduling has explored various approaches to improve ad delivery and effectiveness. Mobile social networks have been utilized for content dissemination, considering user location, mobility, and interests while accounting for resource limitations (Ravaei et al., [15]). In television advertising, combining mathematical programming and machine learning has led to revenue increases of 3-5 % for networks (Souyris et al., [16]). For offline advertising, a system called TARP uses convolutional neural networks to generate viewer demographics and queue scheduling algorithms to display relevant ads on billboards and screens (Malhotra et al., [17]). In the context of IPTV, a 0.502-competitive revenue maximizing scheduling algorithm has been developed to place targeted ads based on comprehensive user profiles derived from TV, broadband, and mobile usage (Kodialam et al., [18]). These advancements aim to enhance ad targeting, increase revenues, and improve viewer experience across various platforms.

III. Problem statement

Within the realm of Wi-Fi Marketing, a common strategy involves presenting advertisements to users prior to granting them internet access via public or business Wi-Fi hotspots. This technique proves effective for engaging potential customers. However, as the volume of advertising campaigns grows, the complexity of resource allocation intensifies. The key challenge lies in optimizing ad placement to fulfill campaign objectives while minimizing disruptions to the user experience.

Wi-Fi Marketing leverages a model where users must view or interact with an advertisement before accessing the internet. This paradigm is characterized by several distinct aspects:

- **Compulsory ad interaction for access:** Users are required to view or engage with an advertisement as a prerequisite to gaining Wi-Fi connectivity. This method ensures direct interaction with users but necessitates strategic ad distribution to maintain a balance between marketing effectiveness and user convenience.
- **Heterogeneous campaigns and audience segmentation:** Advertising campaigns cater to diverse audiences, segmented by factors such as demographics, geographic regions, and behavioral patterns. A dynamic resource allocation framework is critical to delivering relevant advertisements to the appropriate audiences, thereby maximizing campaign impact.
- **Optimization of campaign outcomes:** Ensuring that advertising objectives are met requires a resource allocation mechanism that prioritizes efficient ad delivery while enhancing user engagement.

Consequently, resource allocation in Wi-Fi Marketing centers on the strategic scheduling and delivery of advertisements to achieve campaign goals while ensuring a seamless and positive user experience during Wi-Fi access.

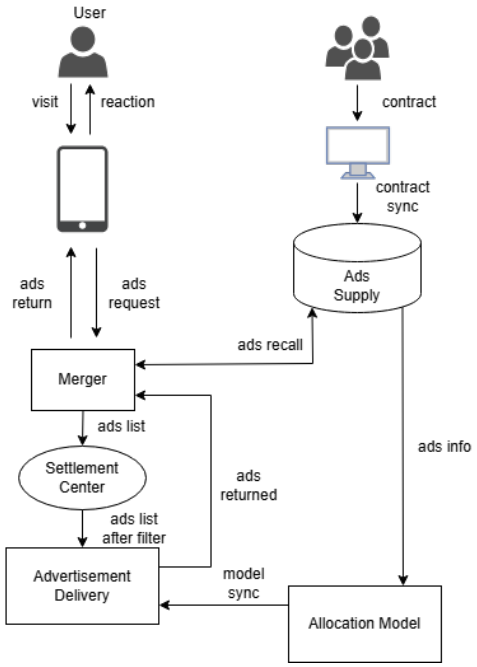


Fig. 1. Overall processing pipeline.

A. Principles of Efficient Resource Allocation

To achieve efficient resource allocation in Wi-Fi marketing, the system must adhere to fundamental principles that ensure fairness, optimize according to demand, and enhance user experience. Below are the key principles aimed at achieving these objectives.

1) Ensuring Fairness and Avoiding Over-Allocation: Fairness in resource allocation ensures that every user or advertisement has a reasonable opportunity for exposure. Additionally, avoiding over-allocation prevents excessive ad repetition, which can irritate users and waste advertising resources.

Suppose x_{ij} represents the resources allocated to advertisement j for user i. The objective is to ensure that the total allocation does not exceed the required limit, expressed as:

$$\sum_{j} x_{ij} \leq d_j, \quad \forall j. \tag{1}$$

where d_j is the maximum limit for the number of advertisements deployed by advertiser j.

To ensure fairness, resources for each advertisement j are allocated based on a priority weight w_j. The allocation model can be optimized using a fairness-maximizing objective function:

$$\max \sum_{i} \sum_{j} p_j \cdot x_{ij}. \tag{2}$$

subject to constraints that prevent exceeding resource limits and maintain fairness among different advertisements.

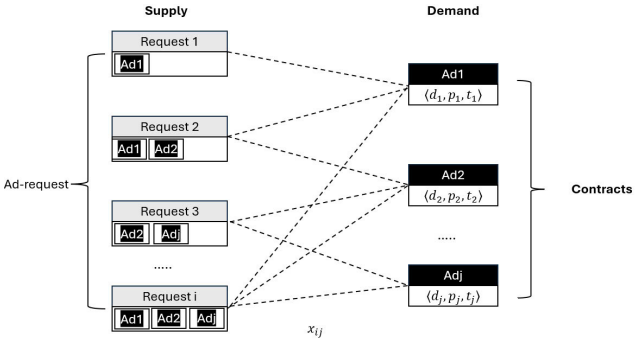


Fig. 2. Contracts between advertisers and publishers.

2) Optimizing Fairness in Resource Allocation: Fairness plays a critical role in ensuring that each advertisement has a reasonable and balanced opportunity for exposure, particularly when multiple campaigns compete for the same advertising space. To achieve fairness, an optimization objective function can be defined as follows:

$$\frac{1}{2} \sum_{i,j \in \Gamma(i)} \frac{1}{\theta_j} \left(x_{ij} - \theta_j\right)^2. \tag{3}$$

Where:

- x_{ij}: The actual amount of resources allocated to advertisement j for user i.
- θ_j: The ideal allocation ratio for advertisement j, representing the desired level of allocation for optimal effectiveness.
- $\Gamma(i)$: The set of advertisements requested by user i.

This objective function minimizes the disparity between the actual allocation x_{ij} and the ideal allocation θ_j, thereby optimizing fairness in resource distribution. A small value of this expression indicates that resource allocation has reached an optimal fairness level, ensuring balanced and reasonable distribution among advertisements.

B. Basic Allocation Model

This model formalizes advertising contracts between advertisers and publishers (the entities responsible for managing Wi-Fi infrastructure). These contracts ensure that advertisements are displayed to users a predetermined number of times within a specific timeframe while meeting additional constraints such as audience segmentation and display duration, as illustrated in Figure 2.

The key components of the model include:

1) *Ad Supply:* This represents a collection of advertisements, $Ad_1, Ad_2, \ldots, Ad_n$, provided by multiple advertisers. Each advertisement has unique requirements for impressions, geographic targeting, and audience segmentation. These requirements often specify the need to reach distinct user groups characterized by specific attributes.

2) *Ad Requests:* Ad requests are user-initiated actions that occur when accessing Wi-Fi services. Each request includes details about the user, the access location, and other contextual attributes, enabling the system to deliver the most relevant advertisements to users.
3) *Contracts:* Contracts between advertisers and publishers define the terms and conditions for advertisement delivery, including:
   - Impression Quota: Specifies the minimum number of times an advertisement must be displayed during a predefined timeframe.
   - Target Audience: Identifies specific demographic or user segments that the advertisement is intended to reach.

1) Advertisement Parameters: Each advertisement Ad_j is characterized by the following parameters:

- d_j: The maximum number of required impressions for advertisement Ad_j.
- p_j: The target audience segment designated for advertisement Ad_j.
- t_j: The timeframe during which advertisement Ad_j must be displayed.

These parameters are represented as a tuple (d_j, p_j, t_j), which serves as the basis for optimizing ad allocation.

2) Objective Function: The resource allocation problem in Wi-Fi marketing aims to optimize the number of ad impressions to maximize revenue for service providers while maintaining user satisfaction. The model satisfies advertising contract requirements, ensures fairness in resource allocation, and considers genuine user interest in advertisements.
The optimization problem is formulated as follows:

$$\arg\min \frac{1}{2} \sum_{j,i\in\Gamma(j)} \frac{V_j}{\theta_j}(x_{ij} - \theta_{ij})^2 - \sum_j p_j \sum_{i\in\Gamma(j)} x_{ij}. \quad (4)$$

where:

- x_{ij}: Represents the allocation of advertisement j to user i.
- θ_j: The ideal allocation ratio for advertisement j, calculated as $\theta_j = \frac{d_j}{\sum_{i\in\Gamma(j)}}$.
- V_j: A weight parameter associated with the importance of fairness for advertisement j.
- p_j: The penalty weight for failing to meet the required impressions for advertisement j.
- d_j: The minimum required impressions for advertisement j.
- s_i: Resource capacity at supply i, estimated from Wi-Fi traffic predictions.

3) Constraints: The model is subject to the following constraints:

1) Minimum Impressions:

$$\sum_{i\in\Gamma(j)} x_{ij} \le d_j, \quad \forall j. \quad (5)$$

This ensures that the number of impressions for advertisement j does not exceed the required quota d_j.

2) Resource Limits:

$$\sum_{j\in\Gamma(i)} x_{ij} \le s_i, \quad \forall i. \quad (6)$$

This constraint limits the allocation to the resource capacity s_i at supply i.
3) Non-Negativity:

$$x_{ij} \ge 0, \quad \forall i, j. \quad (7)$$

This guarantees that all allocation variables x_{ij} and penalty terms u_j remain non-negative, consistent with their physical interpretation in resource allocation.

4) Components of the Objective Function: The objective function balances the following three aspects:

1) Fairness in Allocation

$$\frac{1}{2} \sum_{j,i\in\Gamma(j)} \frac{V_j}{\theta_j}(x_{ij} - \theta_j)^2. \quad (8)$$

This term minimizes the deviation between actual allocation x_{ij} and the ideal allocation θ_j, promoting fair resource distribution.
2) Penalty for Unmet Requirements

$$\sum_j p_j \sum_{i\in\Gamma(j)} x_{ij}. \quad (9)$$

This penalizes scenarios where advertisement j fails to meet the required impressions d_j, helping to satisfy contractual obligations.

C. Model for Allocation Application

In the Wi-Fi Marketing system, the resource allocation problem optimizes two campaign types: Domain and Network. Domain ads focus on location-specific content, enhancing personalization, while Network ads extend brand reach regardless of location. The allocation process occurs in two stages: In Stage 1, resources are allocated evenly to ensure fairness between the campaigns. In Stage 2, unused resources are dynamically reallocated to maximize overall revenue. The model uses ratio constraints to adjust priorities and ensure efficient resource utilization.

1) Stage 1: Fair Distribution: In the first stage of the resource allocation model, the system determines the initial allocation of limited resources to Domain and Network campaigns. The goal is to ensure that each campaign has sufficient resources to meet its minimum display requirements, as outlined in contracts, while maintaining fairness in distribution. This initial allocation serves as a foundation for the next stage, where surplus resources can be reallocated to maximize overall revenue. Our new optimization problem is:

$$\arg\min \frac{1}{2} \sum_{j\in D,i\in\Gamma(j)} \frac{V_j}{\theta_j}(x_{ij} - \theta_{ij})^2. \quad (10)$$

s.t

$$\begin{aligned} \sum_{i\in\Gamma(j)} x_{ij} &\le d_j \quad \forall j. \\ \sum_{j\in\Gamma(i), j\in D} x_{ij} &\le s_i * ratio_i \quad \forall i. \\ \sum_{j\in\Gamma(i), j\in N} x_{ij} &\le s_i * (1 - ratio_i) \quad \forall i. \\ x_{ij} &\ge 0 \quad \forall i, j \in D. \end{aligned} \tag{11}$$

2) Stage 2: Revenue Maximization: In this stage, the objective is to optimize the utilization of the surplus resources from Stage 1. Specifically, if either the Domain or Network campaign does not fully use its allocated resources, the remaining surplus will be transferred to the other campaign. This approach ensures the maximum efficiency in resource utilization, thereby optimizing the overall revenue of the Wi-Fi marketing system.

Using the solution x^* from Stage 1 as the baseline data, the objective function is redefined to focus on maximizing the use of the remaining resources. The problem, then, becomes one of dynamically reallocating the surplus to maximize total revenue, ensuring the most efficient use of available resources across both campaigns. This leads to the mathematical model, which encapsulates these objectives and constraints.

$$\arg\min - \sum_j p_j \sum_{i\in\Gamma(j)} x_{ij}. \tag{12}$$

s.t

$$\begin{aligned} \sum_{i\in\Gamma(j)} x_{ij} &\le d_j \quad \forall j. \\ \sum_{j\in\Gamma(i)} x_{ij} &\le s_i \quad \forall i. \\ x_{ij} &\ge x^\star_{ij} \quad \forall i, j. \\ x_{ij} &\ge 0 \quad \forall i, j. \end{aligned} \tag{13}$$

IV. Experimental Evaluation

A. Experimental Setup

1) Datasets and evaluation metrics: We analyze data from AWING's advertising platform, encompassing 10,996 supply locations, 128 campaigns, and over 4 million edges within a 100-day timeframe. For simulation purposes, a representative sample of this network is used to reflect the allocation dynamics. To benchmark performance, a linear programming (LP) [19]optimizer is employed, establishing an upper bound for the allocation model and enabling a fair comparison of different optimization methods.

To evaluate the effectiveness of guaranteed display advertising allocation, we focus on two key metrics:

- **Number of allocated clicks**, which measures the total clicks generated through valid allocations. This metric directly reflects the system's ability to meet demand and maximize campaign effectiveness.
- **Over-allocation rate**, defined as the ratio of impressions exceeding the required demand to the total allocated impressions. This metric highlights inefficiencies in resource distribution and the system's capacity to minimize unnecessary surplus.

These metrics provide a comprehensive view of allocation efficiency, balancing campaign performance against resource optimization. By filtering out over-allocated impressions during serving, we ensure the insights remain actionable and aligned with operational goals.

2) Benchmark methods: For benchmarking purposes, we explore and compare three recent methods against our proposed approach in offline allocation experiments:

- SHALE: This method, introduced by Bharadwaj et al [20], is a dual-based optimal algorithm designed for the basic allocation model. It focuses on two main objectives: ensuring distribution fairness and maximizing impressions. Although effective in certain contexts, SHALE does not account for additional complexities that may arise in more nuanced allocation scenarios.
- ALI: Proposed by Fang et al [21], ALI is an allocation model that optimizes the Click-Through Rate (CTR). However, a key limitation of ALI is the lack of a constraint to ensure that the number of impressions allocated does not exceed the demand. Instead, the model incorporates a hyper-parameter known as the learning rate. This parameter is employed during the iterative updates of α to mitigate the risk of overloading, helping to stabilize the process during allocation, but it does not fully eliminate the possibility of exceeding demand limits.

These methods provide a baseline for comparison, highlighting the differences in approach and performance when applied to various allocation objectives. In our experiments, we evaluate the effectiveness of these models against the benchmarks, emphasizing how each addresses specific allocation challenges and their respective limitations. Important parameter values used in our experiments are summarized in Table I.

TABLE I
SUMMARY OF PARAMETER VALUES

Parameter	Value	Description
v_j	1	Weight of fairness objective
p_j	10	Weight of penalty objective
$t_{\max}$	50	Max iterations for all methods

B. Results

1) Phase 1 - Resource Allocation for Domain and Network Campaigns: In this phase, we independently solve the resource allocation problem for Domain and Network campaigns. The effectiveness of each method is evaluated based not only on the number of allocated impressions but also on over-allocation rates, the number of clicks received, and the L2 distance metric. Tables II and III summarize the results, allowing comparisons of the effectiveness of our method, SHALE, and ALI.

While SHALE achieves similar allocation rates to our method, the slight difference is negligible. However, ALI

exhibits significantly higher over-allocation rates of 0.138% and 1.21%, respectively, highlighting its weaker control over resource distribution compared to our method and SHALE.

Regarding clicks, our method demonstrates a significant advantage over SHALE and ALI. On the Network and Domain datasets, our method achieves 3.36% and 3.12% more clicks than SHALE, and 2.83% and 4.13% more clicks than ALI, respectively. These results reflect the ability of our approach to attract more clicks and enhance conversion rates, making a meaningful impact on advertising effectiveness.

In terms of the L2 distance metric, our method achieves the lowest values across both datasets, indicating more even ad distribution. Specifically, for the Network dataset, our L2 distance is 4.57% and 6.2% lower than SHALE and ALI, respectively. Similarly, for the Domain dataset, our method outperforms SHALE and ALI by 2.5% and 5.1%. This demonstrates that our method not only attracts users effectively but also maintains balanced resource distribution.

TABLE II
RESULTS FOR THE NETWORK DATASET.

Method	Allocated Impressions	Over-Allocation Rate	Clicks	L2 Distance
RAP	**943103.70**	**0**	**274934.66**	**1.0534e8**
SHALE	931176.32	0	265984.39	1.1034e8
ALI	935213.18	0.138	267378.21	1.1234e8

TABLE III
RESULTS FOR THE DOMAIN DATASET.

Method	Allocated Impressions	Over-Allocation Rate	Click	L2 Distance
RAP	**35139.03**	**0**	**11434.66**	**6.5764e6**
SHALE	33921.89	0	10980.61	6.7443e6
ALI	34019.32	0.125	11089.34	6.9321e6

2) Phase 2 - Maximizing Surplus Resources: Building on Phase 1, Phase 2 introduces an additional requirement that the solution values must meet or exceed the allocation levels from Phase 1.

In terms of allocated impressions, our method achieves 1,378,242.56, significantly exceeding the values from Phase 1 (943,103.70 for Network and 35,139.03 for Domain). This demonstrates our method's capability to utilize surplus resources effectively, expanding the advertising reach compared to Phase 1.

Regarding clicks, our method achieves 482,384.73 clicks in Phase 2, substantially higher than the 274,934.66 and 11,434.66 clicks from the Network and Domain datasets in Phase 1. This increase highlights the continued optimization of user engagement under evolving conditions.

For the L2 distance metric, our method records 2.4384e9 in Phase 2, ensuring even resource distribution. Although the value is slightly higher than Phase 1 due to the increased resource pool and broader distribution scope, it remains significantly better than other methods, maintaining balance and fairness.

TABLE IV
RESULTS FOR PHASE 2.

Method	Allocated Impressions	Over-Allocation Rate	Click	L2 Distance
RAP	**1378242.56**	**0**	**482384.73**	**2.4384e9**
SHALE	1336895.28	0	454544.39	2.9983e9
ALI	1350677.70	0.112	472737.19	3.0323e9

These results highlight the superiority of our method in maximizing surplus resources while maintaining efficiency, fairness, and user engagement.

V. CONCLUSION

In the field of Wi-Fi Marketing, efficient resource allocation plays a crucial role in optimizing advertising campaign performance while maintaining a positive user experience. By applying fundamental principles such as fairness, demand-based optimization, and enhancing user satisfaction, the resource allocation model proposed in this study has demonstrated its ability to balance advertising objectives with user convenience. The optimized objective functions are designed to ensure that advertisements are distributed effectively, achieve high interaction rates, and fully meet contractual requirements without causing user inconvenience.

The resource allocation model, with constraints on advertising quotas and user satisfaction optimization, not only helps maximize campaign effectiveness but also contributes to minimizing the risks of ad overload, thereby enhancing user trust and satisfaction. The research results indicate that optimizing ad allocation based on the established criteria can significantly improve the performance of advertising campaigns in the current diverse and complex Wi-Fi Marketing environment. In the future, research can be expanded and enhanced by:

- Developing Multi-Objective Models: Build multi-objective optimization models to simultaneously maximize various performance metrics such as ClickThrough Rate (CTR), conversion rate, and user satisfaction.
- Practical Experiments: Implement the proposed models in real-world Wi-Fi Marketing environments to evaluate their effectiveness and adjust the models based on practical feedback.
- Expanding Research Scope: Extend the research to encompass other factors such as advertising timing, and user behavior characteristics to create more comprehensive ad allocation strategies.

These research directions will not only help improve the performance of Wi-Fi Marketing campaigns but also contribute to developing smarter, more flexible, and user-friendly advertising solutions in the future.

REFERENCES

[1] H. Yu, M. H. Cheung, L. Gao and J. Huang, "Public Wi-Fi Monetization via Advertising," in IEEE/ACM Transactions on Networking, vol. 25, no. 4, pp. 2110-2121, Aug. 2017, doi: 10.1109/TNET.2017.2675944.

[2] W. Xu, X. Fan, T. Wu, Y. Xi, P. Yang and C. Tian, "Interest Users Cumulatively in Your Ads: A Near Optimal Study for Wi-Fi Advertisement Scheduling," IEEE INFOCOM 2021 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Vancouver, BC, Canada, 2021, pp. 1-6, doi: 10.1109/INFOCOMWKSHPS51825.2021.9484633.
[3] Son Ta Anh, Thuy Thi Nguyen. Solving Resource Allocation Problem in Wifi Network by Dantzig-Wolfe Decomposition Algorithm. JST: Smart Systems and Devices.
[4] Bateni, Mohammad & Chen, Yiwei & Ciocan, Dragos & Mirrokni, Vahab. (2016). Fair Resource Allocation in A Volatile Marketplace. 819-819. 10.1145/2940716.2940763.
[5] J. Xu, K. chih Lee, W. Li, H. Qi and Q. Lu, "Smart pacing for effective online ad campaign optimization," inProceedings of the 21th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD. 2015, pages 2217 –2226.
[6] Turner, John. "The planning of guaranteed targeted display advertising." Operations research 60.1 (2012): 18-33.
[7] Cheng, Xiao, et al. "An Adaptive Unified Allocation Framework for Guaranteed Display Advertising." Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining. 2022.
[8] Fang, Zhen, et al. "Large-scale personalized delivery for guaranteed display advertising with real-time pacing." 2019 IEEE International Conference on Data Mining (ICDM). IEEE, 2019.
[9] Dai, Liang, et al. "Fairness-aware Guaranteed Display Advertising Allocation under Traffic Cost Constraint." Proceedings of the ACM Web Conference 2023. 2023.
[10] Han, Zhu, and KJ Ray Liu. Resource allocation for wireless networks: basics, techniques, and applications. Cambridge university press, 2008.
[11] Song, Guocong, and Ye Li. "Utility-based resource allocation and scheduling in OFDM-based wireless broadband networks." IEEE Communications magazine 43.12 (2005): 127-134.
[12] Dat, Dang Tien, et al. "The periodic product recommendation system based on deep reinforcement learning and the multi-objective framework." 2023 12th International Conference on Awareness Science and Technology (iCAST). IEEE, 2023.
[13] Dat, Dang Tien, et al. "Building the multi-objective periodic recommendation system through integrating optimistic linear support and user clustering to multi-object reinforcement learning."
[14] Mehrjoo, Mehri, Mohamad Khattar Awad, and Xuemin Sherman Shen. "Resource allocation in OFDM-based WiMAX." WiMAX network planning and optimization (2009): 113-131.
[15] Ravaei, Bahman, et al. "Targeted content dissemination in mobile social networks taking account of resource limitation." Concurrency and Computation: Practice and Experience 29.18 (2017): e4207.
[16] Souyris, Sebastián, Sridhar Seshadri, and Sriram Subramanian. "Scheduling Advertising on Cable Television." Operations Research 71.6 (2023): 2217-2231.
[17] Malhotra, Ruchika, et al. "User targeted offline advertising using recognition based demographics and queue scheduling." Int. J. Eng. Adv. Technol.(IJEAT) 9.3 (2020).
[18] Kodialam, Murali, et al. "Online scheduling of targeted advertisements for IPTV." 2010 Proceedings IEEE INFOCOM. IEEE, 2010.
[19] Google or-tools. URL https://developers.google.com/optimization/lp/glop.
[20] Vijay Bharadwaj, Peiji Chen, Wenjing Ma, Chandrashekhar Nagarajan, John Tomlin, Sergei Vassilvitskii, Erik Vee, and Jian Yang. Shale: An efficient algorithm for allocation of guaranteed display advertising. In KDD'12, 2012. ISBN 9781450314626.
[21] Z. Fang, Y. Li, C. Liu, W. Zhu, Y. Zhang, and W. Zhou. Large-scale personalized delivery for guaranteed display advertising with real-time pacing. In ICDM'19, pages 190–199, Nov 2019.

rice2024 Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering pp. 71–77
DOI: 10.15439/2024R34
ISSN 2300-5963 ACSIS, Vol. 42

Harnessing AI for Enhanced Identity Management: Addressing Cybersecurity Challenges in the Digital Age

Suman Thapaliya
Department of Information Technology
Lincoln University College
mailsumanthapaliya@gmail.com

Sudan Jha
Department of Computer Science and Engineering
Kathmandu University
jhasudan@ieee.org

***Abstract*—The integration of Artificial Intelligence (AI) into cybersecurity has significantly advanced identity management, particularly in combating Account Takeover (ATO) and enhancing digital security. Traditional cybersecurity methods often fail to keep up with the dynamic nature of cyber threats, necessitating advanced AI-driven solutions to effectively protect digital identities. This article explores the transformative impact of AI on identity management within the cybersecurity field, focusing on its benefits, challenges, and future potential. A comprehensive review of current literature and empirical findings was conducted, analyzing the application of AI through machine learning, deep learning, and neural networks. The results highlight AI's capability to enable real-time anomaly detection, proactive defense mechanisms, and enhance the resilience of identity protection systems. AI-powered systems exhibit significant advantages in adapting to evolving security threats by providing real-time analysis and understanding the contextual nuances of user behavior. These systems effectively mitigate risks associated with unauthorized access, thereby strengthening overall cybersecurity posture. Key findings emphasize AI's continuous learning from emerging attack tactics, its role in the interpretability of security incidents, and the importance of collaborative frameworks between AI systems and human experts. Addressing challenges such as ethical considerations, algorithmic biases, and the need for transparency remains critical for the ethical deployment and successful integration of AI in cybersecurity.**

***Index Terms*—Artificial Intelligence (AI), Cyber Security, Data Protection, Identity Management, Threat Detection.**

I. Page Layout

THE FIELD of Cyber Security has seen a substantial transition with the introduction of Artificial Intelligence (AI), particularly in the vital areas of Account Takeover (ATO) prevention and Identity Management [1]. It might be difficult for traditional security measures to keep up with the more sophisticated attackers due to the constantly shifting landscape of cyber threats [2]. The field of artificial intelligence has evolved as a transformative instrument, offering sophisticated capabilities to promptly detect, avert, and mitigate identity-related risks [3]. Artificial intelligence (AI)-powered systems can safeguard digital identities and prevent unauthorized access through machine learning algorithms, behavioral analytics, and anomaly detection, among other methods [4]. This article explores the revolutionary effects of AI on identity management, outlining the technology's main benefits, drawbacks, and potential applications in the constantly changing field of cybersecurity [5].

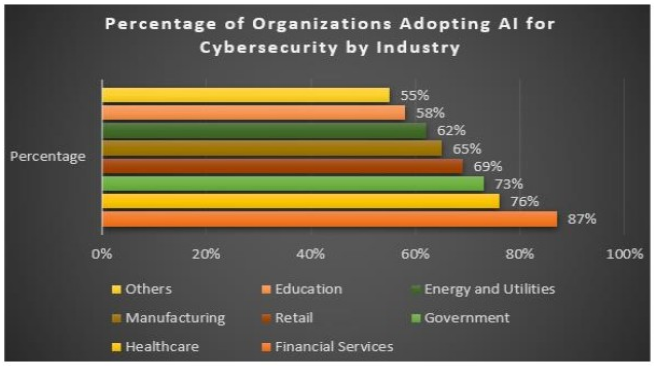


Figure I. Adoption of AI in Cybersecurity by Industry

II. Literature Review

- Camacho (2024) outlined the applications of AI in cybersecurity, including threat detection, vulnerability assessment, incident response, and predictive analysis. Artificial intelligence (AI) systems quickly analyzed enormous amounts of data to find unusual patterns suggestive of possible security breaches by utilizing machine learning techniques. Furthermore, proactive defense mechanisms were made possible by AI-driven technology, giving organizations the ability to proactively minimize risks and protect sensitive information. But the use of AI in cybersecurity also brought up important privacy and ethical issues, so its application must be approached with balance. After conducting a thorough analysis, Camacho emphasized how cybersecurity frameworks must include artificial intelligence (AI) in order to successfully mitigate threats in the digital age [17].
- Varney (2019) examined how artificial intelligence (AI) changed cybersecurity and safeguarded digital ecosystems, pointing out that AI performed better on cognitive tasks than humans, allowing for more complex attacks. The study made clear the necessity of continuing research to improve defenses against threats like malware and phishing that are

enabled by AI. It also covered the unethical differences in AI application between the US and its enemies. Expert systems have been proposed as a way to use AI's ability to identify patterns for defense.

- Chakraborty et al. (2023) talked about how the digital era's technical advancements automated daily tasks but lacked adequate security. They emphasized how difficult it is becoming to secure connected devices and how sophisticated AI-driven cyberthreats are becoming more prevalent. The authors looked at both traditional and sophisticated defense strategies against cyberattacks before offering some possible future uses for AI in cybersecurity.
- Mohammed (2020) talked about how AI may be used to solve cybersecurity problems while highlighting the risks associated with the digital revolution, like data mining and exploitation. It emphasized the value of managing digital identities as well as the possibilities of blockchain technology. In order to combat cybercrime, the report emphasized the necessity for more secure data storage techniques and the shortcomings of conventional systems. Lastly, it discussed how AI may help reduce cyberattacks and solve cybersecurity issues.
- Ansari et al. (2022) explored the use of artificial intelligence (AI) in cybersecurity, emphasizing the technology's expanding impact on the sector. They observed that, as information technology becomes more prevalent in enterprises, cybersecurity is becoming more and more important in the technology sector. The study covered how artificial intelligence (AI) has greatly impacted cybersecurity and how this has resulted in the notable inclusion of machine learning in new cybersecurity-related technologies. In order to examine the overall effects of artificial intelligence on cybersecurity, the writers reviewed the literature, examining both the advantages and disadvantages of the technology.

III. AI In identity management

The role of AI in securing digital identities

Artificial intelligence plays a crucial role in preserving the security of digital identities through the application of state-of-the-art techniques such as machine learning, deep learning, and neural networks [18]. Artificial intelligence (AI) systems possess the capacity to scrutinize copious amounts of data, encompassing user behavior patterns, gadget fingerprints, and contextual details, with the aim of constructing a comprehensive comprehension of every individual's digital identity [6]. Artificial intelligence (AI) can detect anomalies or unauthorized access attempts quickly by establishing a typical user behavior, which can help detect potential security breaches promptly [7].

Understanding contextual nuances and user behavior

AI systems are very good at understanding the nuances and complexity of human behavior, which is essential for identity management to work. By examining user activities such as device usage, network activity, and login times, AI may get a thorough grasp of each user's unique behavioral signature [8]. AI can distinguish between legitimate user behaviors and suspect activity thanks to this contextual understanding, even in situations where the latter may resemble the former quite a bit .

Real-time anomaly detection

The capacity of AI to detect anomalies in real-time is one of its primary benefits for identity management. AI systems monitor user behavior continuously, contrasting it with known threat patterns and behavioral baselines [9]. AI has the ability to detect anomalies, such as an odd login location or an abrupt shift in user behavior, and to set off pre-established security protocols and generate notifications. Preventing unwanted access and lessening the effects of any security breaches require the capacity to detect in real time.

Proactive defense against unauthorized access attempts

AI makes it possible to defend against unwanted access attempts in a proactive manner. Artificial intelligence (AI) can remain one step ahead of prospective adversaries by constantly learning and adapting to shifting danger scenarios. Security systems can proactively take preventive action by using machine learning algorithms that are able to recognize trends and indicators that could indicate possible threats. Furthermore, AI can automate the application of security guidelines and access restrictions, ensuring that only those with permission can access confidential information [10].

A. AI in Account Takeover (ATO) Prevention

Figure 2. Percentage of Successful Cyber Attacks Prevented by AI-based Security Systems

B. Recognizing patterns indicative of compromised accounts

The detection of patterns that indicate compromised accounts is a critical function of artificial intelligence, which aids in the prevention of Account Takeover (ATO). Large amounts of data, including login attempts, user behavior,

and device information, can be analyzed by machine learning algorithms to find potentially suspicious activity [11]. AI, for example, can detect unusual login locations, sudden changes in a user's behavior, or multiple failed attempts at authentication, all of which could be signs of an account compromise. When these trends are identified in a timely manner, AI-powered systems can initiate alerts and quickly take corrective action, like freezing an account or adding further verification processes.

C. Employing multi-factor authentication

The effectiveness of multi-factor authentication (MFA) in preventing unwanted access is greatly increased by AI. Conventional MFA relies on inflexible guidelines and preset challenges that proficient attackers can easily circumvent. In contrast, MFA that is powered by AI dynamically modifies authentication difficulties according to environmental circumstances and the user's risk profile [12]. For instance, AI may request additional authentication information from the user, such as biometric data or a one-time password, if it detects a login attempt from an unknown device or location. Organizations can utilise risk-based authentication with AI, enabling the deployment of the appropriate level of security in response to the perceived threat level [19].

D. Continuous learning from evolving attack tactics

AI gives security systems the ability to predict changing attack strategies and react continuously, keeping them one step ahead of their enemies. Cybercriminals are constantly coming up with new techniques to launch ATO attacks, such as social engineering, credential stuffing, and phishing. Artificial intelligence (AI)-enabled systems can learn from changing strategies by analyzing past attack data, seeing new trends, and then updating their algorithms. Organizations are able to maintain a robust defense against attempts to obtain unauthorized access and respond to evolving risks thanks to this continuous learning process. AI can also detect any security flaws in the system and recommend improvements to make it more secure overall.

E. Enhancing the resilience of identity protection mechanisms

By providing an additional layer of security, artificial intelligence fortifies the resilience of identity protection systems. Passwords and security questions are examples of traditional identity protection techniques that are commonly vulnerable to social engineering and brute-force attacks. By applying cutting-edge methods like behavioral biometrics and user profiling, AI can improve these mechanisms. To create a unique behavioral signature, AI, for example, might examine a variety of inputs, including mouse movements, typing habits, and device interactions. Along with more traditional authentication techniques, this signature can be used to confirm the user's identity. By combining artificial intelligence (AI) with well-established identity protection protocols, businesses may create a strong, adaptable security framework that effectively thwarts efforts by unauthorized individuals to gain access [21].

IV. Advantages of AI in Identity Management

TABLE I. Advantages of AI in Identity Management

Advantage	Description
Adaptability to evolving security tactics	AI algorithms can quickly learn and adapt to new security threats, ensuring that the system remains effective against the latest attack tactics.
Providing interpretable insights	AI-powered systems offer interpretable insights into security incidents, enabling security teams to understand the underlying causes and take appropriate actions.
Real-time analysis for efficient response	AI enables real-time analysis of identity-related data, facilitating efficient response mechanisms to security incidents and minimizing the potential impact [20].
Understanding contextual nuances	AI excels at understanding contextual nuances and learning from diverse cyber scenarios, enhancing its effectiveness in identifying and mitigating identity-related risks.
Proactive defense against emerging threats	AI enables a proactive defense against emerging threats by continuously monitoring user behavior and system activity, allowing organizations to take preventive measures.
Mitigating risks and safeguarding sensitive data	AI plays a crucial role in mitigating risks and safeguarding sensitive information by implementing robust access controls, monitoring mechanisms, and detecting anomalous activities.

A. Adaptability to evolving security tactics

Because AI is so flexible, it provides a huge advantage in identity management when it comes to adjusting to evolving security strategies. Artificial intelligence (AI) has the capacity to pick up on changes in cybercrimes' tactics swiftly, as they are constantly evolving [13]. AI algorithms may automatically update their threat detection models by analyzing large volumes of data and spotting new trends, which keeps the security system up to date against the most modern dangers. To stay ahead of opponents and keep up a strong defense against identity-related attacks, one must be flexible.

B. Providing interpretable insights into security incidents

Identity management systems driven by artificial intelligence (AI) offer insightful information about security events, enabling security teams to comprehend the type and source of risks. AI is capable of producing in-depth reports and visualizations that offer thorough justifications for the

choices it makes. These revelations may include the particular behavior information offered includes the precise patterns of behaviour or abnormalities that set off an alarm, the sequence of events that preceded the incident, and the possible effects on the organisations. Security teams are capable of prioritising response actions, making well-informed judgements, and successfully reducing risks [14].

C. Real-time analysis for efficient response mechanisms

AI makes it possible to analyse identity-related data quickly, which facilitates the development of efficient reaction plans for security events [22]. AI is capable of identifying abnormalities and possible threats in real-time by utilising its enormous data processing capabilities. Security teams may act quickly by blocking bogus login attempts, isolating affected accounts, or triggering incident response procedures thanks to this real-time analysis. AI is essential in decreasing the possible impact of identity-related attacks and helping organisations maintain the integrity of their systems by shortening the time lag between threat detection and reaction.

D. Understanding contextual nuances and learning from diverse cyber scenarios

AI exhibits a great capacity to learn from a variety of cyber scenarios and grasp the complex intricacies of identity management. AI has the capacity to generate comprehensive profiles of both common and uncommon actions by analysing a variety of data points, including user behaviour, device information, and network activity. AI can differentiate between legitimate user behaviours and possible dangers even in situations where the differences are subtle because of its contextual understanding [15]. Furthermore, AI continuously improves its algorithms and gains accuracy over time by learning from a variety of cyber scenarios it encounters. AI is a powerful tool for identifying and reducing identity-related hazards because of its ability to understand context and learn from experience.

E. Proactive defense against emerging threats

AI strengthens identity management's proactive defence against new threats [16]. AI can identify possible weaknesses and weak points in the security posture by continuously monitoring user behaviour and system activity. Businesses that adopt a proactive approach may be able to patch security flaws before hackers can exploit them. AI can also assess the effectiveness of existing security measures and simulate different attack scenarios, which helps organisations find and fix weaknesses in their identity management systems [23]. Organisations can keep ahead of adversaries and lower the likelihood of successful identity-related attacks by implementing a proactive approach with AI.

F. Mitigating risks and safeguarding sensitive information

In identity management, artificial intelligence (AI) is essential for reducing risks and safeguarding sensitive data. Artificial Intelligence successfully lowers the risk of unauthorised access, data breaches, and identity theft by accurate identification and quick reaction to possible threats. Strict access controls can be enforced with the use of AI-powered solutions, ensuring that only authorised users have access to sensitive resources [24]. Furthermore, AI can keep an eye on user behaviour and spot any potentially worrying activities, such unusual data access patterns or attempts to harvest private information. AI helps organisations protect their most important assets and maintain the security, dependability, and accessibility of sensitive data by notifying security staff in a timely manner and putting preventive steps in place [25].

V. Challenges and Considerations

Machine learning and deep learning algorithms are used by cybersecurity experts to perform tasks like intrusion detection, malware analysis, and anomaly identification. Every algorithm has pros and cons of its own, and the best approach relies on the particular security problem that needs to be solved. Here are a few instances of outcomes from various algorithms:

- *Support Vector Machines (SVM):* SVM is a popular intrusion detection technique that has proven to be effective at recognising known attacks. For example, SVM was used in conjunction with ant colony networks in a study by Feng et al. to detect network invasions. Using the KDD Cup 1999 dataset, the study attained an outstanding accuracy rate of 96.75%.
- *Deep Learning:* Deep learning methods such as Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN) have demonstrated promising outcomes in identifying intricate and unfamiliar attacks. Jiang et al. demonstrated a multichannel CNN in their work that was able to identify clever attacks with remarkable accuracy—99.98% —on the NSL-KDD dataset. CNN and RNN were combined to create a deep learning method for network intrusion detection by Shone et al. Their accuracy on the KDD Cup 1999 dataset was 97.85%.
- *Ensemble Methods:* Several learning techniques combined improve robustness and speed. Xin et al. integrated deep belief networks (DBN) and support vector machines (SVM) to efficiently detect intrusions. They achieved an amazing accuracy rate of 99.14% on the NSL-KDD dataset [26].
- *Transfer Learning:* Utilising expertise from one field to improve performance in another. Transfer learning is effective in detecting new attack patterns, as demonstrated by Gu et al. who used it to find weaknesses in the supply chain of the machine learning model [27].
- *Sequential Models:* Because sequential models are so good at modeling sequential data, they have been used extensively to identify IoT-botnet attacks. One example of this is Long Short-Term

Memory (LSTM) networks. On a customised dataset, the LSTM-based method suggested by Soe et al. effectively identified IoT-botnet attacks with an astounding accuracy of 99.23% [28].

These findings demonstrate how various algorithms perform differently in some cybersecurity tasks. It's crucial to recognise the difficulties and restrictions associated with any method, though. These include the need for large-scale datasets that have been annotated, the potential for adversarial attacks, and the models' interpretability. Taking into account the unique requirements and constraints of respective cybersecurity applications, researchers and practitioners should carefully evaluate the suitability of different algorithms [29].

TABLE II. CHALLENGES AND CONSIDERATIONS IN IMPLEMENTING AI FOR CYBERSECURITY

Challenge/ Consideration	Description
Ensuring the ethical use of AI	Organizations must establish clear guidelines and protocols to govern the collection, storage, and analysis of data used by AI systems, ensuring alignment with ethical principles and individual privacy rights.
Addressing potential biases in AI algorithms	To mitigate biases, organizations must ensure that the training data is diverse, representative, and free from discriminatory patterns, and regularly audit and test AI algorithms for fairness.
Maintaining transparency and accountability	Organizations should strive for explainable AI, where the reasoning behind AI decisions is clearly articulated, and establish accountability mechanisms, such as regular audits and oversight committees, to ensure responsible use of AI systems.
Collaborating with human experts	Effective collaboration between AI systems and human experts is essential for achieving optimal results, leveraging the strengths of both artificial and human intelligence to develop a comprehensive approach to cybersecurity.

A. Ensuring the ethical use of AI in Cyber Security

As AI grows more common in cybersecurity, ethical use becomes harder. AI systems may analyze massive amounts of sensitive data, raising privacy and abuse concerns. Clear policies and processes are needed to regulate AI data collection, storage, and processing. AI systems must also be responsible, open, and just. To maintain trust and avoid unanticipated consequences, AI systems must prioritize person rights and respect the law and ethics.

B. Addressing potential biases in AI algorithms

AI biases must be addressed in cybersecurity. When AI systems learn algorithms from biased data in the training set, biases may be reinforced. An AI system trained on a dataset of attacks from a given demographic may become biased in identifying threats from that demographic. This can cause false positives and unjust profiling. To reduce preconceptions, organizations must prioritize diverse, representative training data without bias. Audits and testing must be done often to discover and fix bias-related issues.

C. Maintaining transparency and accountability

Accountability and transparency are essential when employing AI in cybersecurity. AI systems' complicated and confusing decision-making mechanisms might make decisions hard for stakeholders to understand. AI-powered security systems without transparency may lose trust and raise issues about their impartiality. To solve this problem, organizations should aim toward explainable AI, which explains AI judgments to humans. Additionally, oversight committees and frequent audits can help ensure that AI systems are working as planned and meeting organizational and legal standards.

D. Collaborating with human experts for optimal results

AI in cybersecurity requires human involvement for optimum results. Despite their proficiency at processing huge volumes of data and spotting hidden patterns, AI algorithms are worse at contextual comprehension and intuition than humans [30]. Analysts must understand the security environment, analyze AI algorithm outputs, and make informed recommendations. Collaboration between artificial and human intelligence can boost cybersecurity [31]. Organizations must encourage cooperation and give training to integrate AI technology into security operations [32-25].

VI. FUTURE DIRECTIONS

A. Continuous advancements in AI technologies for Cyber Security

AI technology is improving cybersecurity and driving innovation. AI algorithms should improve our ability to solve complicated cybersecurity problems. Deep learning methods like generative adversarial networks (GANs) can provide more varied and realistic datasets to help AI models recognize and respond to new threats [22]. Advances in sentiment analysis and natural language processing (NLP) may allow AI systems to better interpret unstructured data like social media and forum postings to identify security issues. Advancements in AI technologies are expected to defend organizations against cyber-attacks.

B. Integration of AI with other emerging technologies (e.g., blockchain, quantum computing)

AI in cybersecurity may take an interesting turn when combined with blockchain and quantum computers. These technologies enable organizations to build resilient systems that can handle today's complicated threat scenario. Blockchain technology can securely log security issues and protect AI model training data. Quantum computing could improve traditional computer systems by helping AI algorithms detect and respond to threats faster. As AI and cybersecurity technologies advance, we may expect new and inventive uses.

C. Fostering interdisciplinary research and collaboration

Interdisciplinary research and collaboration are needed to maximize AI's cybersecurity potential. Cybersecurity is complex and requires knowledge in computer science, mathematics, psychology, and social science. Collaboration between scholars and practitioners from diverse fields can improve cybersecurity. This strategy considers complicated organizational-human connections that determine cyber risk. This may encompass interdisciplinary research centers and programs and opportunities for industry, government, and academia to collaborate and share knowledge. By combining stakeholders' perspectives and expertise, we can build a more secure digital future.

VII. Conclusion

In conclusion, the integration of artificial intelligence (AI) into cybersecurity has brought about significant transformations in identity management and threat detection. AI-powered systems offer unparalleled capabilities in securing digital identities, detecting anomalies, and preventing unauthorized access attempts through advanced machine learning algorithms and real-time analysis. Moreover, AI enables proactive defense mechanisms, continuous learning from evolving attack tactics, and the enhancement of resilience in identity protection mechanisms. Despite these advantages, challenges such as ensuring ethical use, addressing biases in AI algorithms, maintaining transparency and accountability, and collaborating effectively with human experts must be carefully considered and addressed.

Looking ahead, the future of AI in cybersecurity holds immense potential for further advancements. Continuous improvements in AI technologies are expected to enhance our ability to address complex cybersecurity issues, while integration with other emerging technologies like blockchain and quantum computing promises to create stronger and more resilient cybersecurity systems. Fostering interdisciplinary research and collaboration will be crucial in fully realizing AI's potential in cybersecurity, as it requires expertise from various fields to tackle the multidimensional nature of cyber risk effectively. By leveraging AI technologies and fostering collaboration across disciplines, we can pave the way for a more robust and secure digital future

References

[1] Benhadjyoussef, N., Karmani, M., & Machhout, M. (2021). Power-based Side Channel Analysis and Fault Injection: Hacking Techniques and Combined Countermeasure. International Journal of Advanced Computer Science and Applications, 12(5).

[2] Babu, V. H., & Balaji, K. (2020). Survey on modular multilevel inverter based on various switching modules for harmonic elimination. In Intelligent Computing in Engineering: Select Proceedings of RICE 2019 (pp. 451-458). Springer Singapore.

[3] Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: an overview from machine learning perspective. Journal of Big data, 7, 1-29.

[4] Arri, H. S., Singh, R., Jha, S., Prashar, D., Joshi, G. P., & Doo, I. C. (2021). Optimized task group aggregation-based overflow handling on fog computing environment using neural computing. *Mathematics, 9*(19), 2522. https://doi.org/10.3390/math9192522

[5] Rosenberg, I., Shabtai, A., Elovici, Y., & Rokach, L. (2021). Adversarial machine learning attacks and defense methods in the cyber security domain. ACM Computing Surveys (CSUR), 54(5), 1-36.

[6] Wang, Z., Wang, E., & Zhu, Y. (2020). Image segmentation evaluation: a survey of methods. Artificial Intelligence Review, 53(8), 5637-5674.

[7] Jha, S., Ahmad, S., Arya, A., Alouffi, B., Alharbi, A., Alharbi, M., & Singh, S. (2023). Ensemble learning-based hybrid segmentation of mammographic images for breast cancer risk prediction using fuzzy C-means and CNN model. *Journal of Healthcare Engineering, 2023*(1), 1491955. https://doi.org/10.1155/2023/1491955

[8] Camacho, N. G. (2024). The role of ai in cybersecurity: Addressing threats in the digital age. Journal of Artificial Intelligence General science (JAIGS) ISSN: 3006-4023, 3(1), 143-154.

[9] Varney, A. (2019). Analysis of the impact of artificial intelligence to cybersecurity and protected digital ecosystems (Master's thesis, Utica College).

[10] Chakraborty, A., Biswas, A., & Khan, A. K. (2023). Artificial intelligence for cybersecurity: Threats, attacks and mitigation. In Artificial Intelligence for Societal Issues (pp. 3-25). Cham: Springer International Publishing.

[11] Mohammed, I. A. (2020). Artificial intelligence for cybersecurity: A systematic mapping of literature. Artif. Intell, 7(9), 1-5.

[12] Ansari, M. F., Dash, B., Sharma, P., & Yathiraju, N. (2022). The impact and limitations of artificial intelligence in cybersecurity: a literature review. International Journal of Advanced Research in Computer and Communication Engineering.

[13] Jha, S., Jha, N., Prashar, D., Ahmad, S., Alouffi, B., & Alharbi, A. (2022). Integrated IoT-based secure and efficient key management framework using hashgraphs for autonomous vehicles to ensure road safety. *Sensors, 22*(7), 2529. https://doi.org/10.3390/s22072529

[14] Sharma et al., "Artificial Intelligence Techniques for Landslides Prediction Using Satellite Imagery," in IEEE Access, vol. 12, pp. 117318-117334, 2024

[15] Buczak, A. L., & Guven, E. (2015). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications surveys & tutorials, 18(2), 1153-1176.

[16] Jha, S., Kumar, R., Hoang Son, L., Abdel-Basset, M., Priyadarshini, I., Sharma, R., & Viet Long, H. (2019). Deep learning approach for software maintainability metrics prediction. *IEEE Access, 7,* 61840–61855. https://doi.org/10.1109/ACCESS.2019.2913349

[17] Feng, W., Zhang, Q., Hu, G., & Huang, J. X. (2014). Mining network data for intrusion detection through combining SVMs with ant colony networks. Future Generation Computer Systems, 37, 127-140.

[18] Jiang, F., Fu, Y., Gupta, B. B., Liang, Y., Rho, S., Lou, F., ... & Tian, Z. (2018). Deep learning based multi-channel intelligent attack detection for data security. IEEE transactions on Sustainable Computing, 5(2), 204-212.

[19] Ahmad, S., Jha, S., Eljialy, A. E. M., & Khan, S. (2021). A systematic review on e-wastage frameworks. *International Journal of Advanced Computer Science and Applications, 12*(12). https://doi.org/10.14569/IJACSA.2021.0121287

[20] Shone, N., Ngoc, T. N., Phai, V. D., & Shi, Q. (2018). A deep learning approach to network intrusion detection. IEEE transactions on emerging topics in computational intelligence, 2(1), 41-50.

[21] Rajagopal, S. Ahmad, S. Jha, H. A. M. Abdeljaber, and J. Nazeer, "AI Based Secure Analytics of Clinical Data in Cloud Environment: To-

wards Smart Cities and Healthcare," J. Adv. Inf. Technol., vol. 14, no. 5, pp. 1132-1142, 2023.

[22] Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., ... & Wang, C. (2018). Machine learning and deep learning methods for cybersecurity. Ieee access, 6, 35365-35381.

[23] Jha, S., Nkenyereye, L., Prasad Joshi, G., & Yang, E. (2020). Mitigating and monitoring smart city using internet of things. *Computers, Materials & Continua, 65*(2), 1059–1079. https://doi.org/10.32604/cmc.2020.011754

[24] Gu, T., Dolan-Gavitt, B., & Garg, S. (2017). Badnets: Identifying vulnerabilities in the machine learning model supply chain. arXiv preprint arXiv:1708.06733.

[25] Soe, Y. N., Feng, Y., Santosa, P. I., Hartanto, R., & Sakurai, K. (2020). Machine learning-based IoT-botnet attack detection with sequential architecture. Sensors, 20(16), 4372.

[26] Jha, S., Prasad Joshi, G., Nkenyereya, L., Wan Kim, D., & Smarandache, F. (2020). A direct data-cluster analysis method based on neutrosophic set implication. *Computers, Materials & Continua, 65*(2), 1203–1220. https://doi.org/10.32604/cmc.2020.011618

[27] Jha, S., Prashar, D., Long, H. V., & Taniar, D. (2020). Recurrent neural network for detecting malware. *Computers & Security, 99,* 102037. https://doi.org/10.1016/j.cose.2020.102037

[28] Saad, A., Faddel, S., Youssef, T., & Mohammed, O. A. (2020). On the implementation of IoT-based digital twin for networked microgrids resiliency against cyber-attacks. IEEE transactions on smart grid, 11(6), 5138-5150.

[29] Jha, S., Seo, C., Yang, E., & Joshi, G. P. (2021). Real-time object detection and tracking system for video surveillance. *Multimedia Tools and Applications, 80*(3), 3981–3996. https://doi.org/10.1007/s11042-020-09749-x

[30] Shaukat, K., Luo, S., Varadharajan, V., Hameed, I. A., & Xu, M. (2020). A survey on machine learning techniques for cyber security in the last decade. IEEE access, 8, 222310-222354.

[31] Usama, M., Qadir, J., Raza, A., Arif, H., Yau, K. L. A., Elkhatib, Y., ... & Al-Fuqaha, A. (2019). Unsupervised machine learning for networking: Techniques, applications and research challenges. IEEE access, 7, 65579-65615.

[32] S. Jha, S. Routray, H. A. M. Abdeljaber, and S. Ahmad, "A novel approach for decision support system in cricket using machine learning," Int. J. Comput. Appl. Technol., vol. 70, no. 2/3, pp. 86-92, 2022.

[33] Xiao, L., Wan, X., Dai, C., Du, X., Chen, X., & Guizani, M. (2018). Security in mobile edge caching with reinforcement learning. IEEE Wireless Communications, 25(3), 116-122.

[34] S. Jha, "Model Selection Procedure in Alleviating Drawbacks of The Electronic Whiteboard," 2021 3rd International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), Greater Noida, India, 2021, pp. 319-322

[35] V. Puri et al., "A Hybrid Artificial Intelligence and Internet of Things Model for Generation of Renewable Resource of Energy," in IEEE Access, vol. 7, pp. 111181-111191, 2019

rice2024 Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering pp. 79–87
DOI: 10.15439/2024R15
ISSN 2300-5963 ACSIS, Vol. 42

Enhancing Plant Disease Detection Through Image Analysis Using SSD Mobilenet V2 and ResNet-50

Paribartan Timalsina
Department of Computer Science and Engineering
Kathmandu University
Dhulikhel, Nepal
timalsinapari015@gmail.com

Subarna Bhattarai
Department of Computer Science and Engineering
Kathmandu University
Dhulikhel, Nepal
sbrnbhttr@gmail.com

Shaswot Paudel
Department of Computer Science and Engineering
Kathmandu University
Dhulikhel, Nepal
shaswotpaudel58@gmail.com

Sudan Jha*
0000-0003-0074-2584
Department of Computer Science and Engineering
Kathmandu University
Dhulikhel, Nepal
jhasudan@ieee.org

***Abstract*—Being an agrarian nation, Nepal possesses huge economic value from the agricultural sectors with more than half of its population contribution in farming, thus bringing a great potential share in the nation's Gross Domestic Product (GDP). Despite this substantial contribution, there are still issues with advancement that hinders growth in this field. This study, therefore presents a model for the detection of plants and diagnosis of diseases associated with the plants using real-time camera feeds or by analyzing captured images. Existing models often suffer from unrelated irrelevant images; confusing one plant for another which results in failure of proper diagnosis of diseases. Addressing this issue, our model first detects the presence of a plant within the frame or image before disease classification. Plant detection is performed using the Single-Shot Detector (SSD) Mobilenet V2 model. Disease classification process is initiated only when the plant is detected. The Residual Network (ResNet)- 50 model performs the plant disease classification taking the clipped image from detected plant. By focusing only on the detected plant, we reduce background complexity and improve classification accuracy. The model has demonstrated a high level of accuracy in both detecting plants and classifying diseases based on the prevalent diseases associated with specific plants.Our model achieved a mean Average Precision (mAP) score of 0.61 for the object detection model which includes 9 classes and achieved average accuracy of 98% for the classification of 9 plant species having a total of 33 classes. This innovative model is hence targeted at providing reliable and accurate plant detection and disease diagnosis to address some of the key challenges in agricultural technology in Nepal. On near future it has the potential for automated plant diseases detection and real time monitoring of plant diseases. Additionally, we can deploy it in mobile applications, allowing farmers to identify plant diseases using smartphone cameras, facilitating timely detection and reduction of crop losses.**

***Index Terms*—Deep Learning, SSD Mobilenet V2, ResNet-50**

This research was supported by Kathmandu University.
*Corresponding author: jhasudan@ieee.org

I. Introduction

Agriculture being one of the most critical macroeconomic sectors is making a substantial economic contribution in the nation's economy. Approximately 60.4% of the population is engaged in agriculture, which accounts for 27% of Nepal's Gross Domestic Product (GDP) [1]. Despite this fact, small-scale farmers often face challenges related to ignorance and lack of access to cutting-edge, affordable technologies such as precision agricultural tools and technology for crops diseases prediction. This technological gap shortfall affects the crop production and overall productivity. Likewise, the existing bio-molecular technologies like antigen-antibody reaction, DNA sequence amplification are very sensitive for small-scale use and are very costly.

To address the issue, we propose a model suitable for identification of plant species and classification of diseases using object detection and Deep Learning (DL) methodologies. The model utilizes either live camera feeding or image uploading for plant detection and disease classification. The proposed model through commonly available smartphones aims to make the solution accessible and user-friendly. For plant species detection, we employ an object detection model based on Convolutional Neural Networks (CNNs). Specifically, we utilize the Single-Shot Multibox Detector (SSD) with a MobileNet V2 architecture, allowing us to build a lightweight model that is computationally efficient and suitable for real-time applications. SSD MobileNet V2 is different as it uses the depthwise separable convolution thus reducing the parameters of the model. SSD MobileNet V2 is mostly used beacuse of it's lightweight characterstics and maintaining the balance between speed and accuracy. The SSD generates bounding boxes with confidence scores which allows the system to

correctly locate the position of a plant within an image.

For disease classification, we make use of the ResNet-50 architecture among the best CNN architectures in the strong extraction of deep features from images. This architecture enables the model to effectively find out the disease classification depending on the features in the detected region of the plant. The ResNet-50 architecture can effectively manage network depth, increase performance and generalization capacity. The key factor for using ResNet-50 lies in it's ability to learn set of residual functions to map the input to desired output. Balancing lightweight object detection for plant identification with a more complex classification model enables us to maintain good accuracy while minimizing resource consumption for fast and efficient system. The proposed method has demonstrated the potential for drastic improvement in agriculture. Productivity is conveyed by providing smallholder farmers with a reliable means and accessible tool for plant and disease identification. Using less complex SSD MobileNet V2 for detection followed by more complex ResNet-50 for classification, we achieve an optimal trade-off between performance and computational efficiency.

II. Literature Review

Numerous bio-molecular research studies have been conducted to increase the early and accurate detection of plant diseases like Enzyme Immunoassay/Enzyme-Linked Immunosorbent Assay(ELISA) and Real-Time Polymerase Chain Reaction (RT-PCR). ELISA, based on the antigen-antibody reaction to detect the diseases in plants by quantifying various kinds of molecules like proteins, hormones and toxins can give high false positive or negative results if the experiment isn't done with great care and also depending on various factors like freshness, concentration of solution and types of organism [2], [3]. Another method, Real-Time Polymerase Chain Reaction(RT-PCR) used to raise and quantify the targeted DNA molecule at the same time is effective because of it's rapid testing with greater sensitivity but is still very expensive for daily applications because of high costs of machines and reagents [4].

As the biological molecular methods are very sensitive and can be very costly researchers have come up with an idea of detecting the plants and it's related diseases using deep learning techniques. One early study implemented a multistep process, including color transformation, masking of green pixels and removal with specific thresholds, and segmentation through equal-sized patches, followed by classification using a database of 500 plants [5]. This approach used Support Vector Machine (SVM) classifier with an accuracy of 94.74%.

Remote Sensing technologies including airborne multispectral/hyperspectral imagery and high-resolution satellite sensors enable image acquisition using airborne or satellite sensors followed by pre-processing and feature extraction with mapping of crop diseases at last [6]. [7] allow for the analysis of spectral properties, revealing changes in reflectance patterns that indicate stress or disease in crops where healthy vegetation reflects significant light in the near-infrared region, while stressed or diseased plants show altered reflectance.

Several research studies were focused on isolating lesions for plant disease detection by removing the background from leaf images. An effective technique for image segmentation was presented based on a Chan-Vese model and Sobel operator [8]. This method includes three steps: extracting leaf contours via the Chan–Vese model and detecting edges using an enhanced Sobel operator, removing the background by identifying high green-level pixels, and isolating the target leaf in complex backgrounds by combining the Chan-Vese and Sobel results. A proposed model, combined traditional CNN with squeeze-and-excitation (SE) module and global pooling layer to identify the plant diseases and achieved an overall accuracy of 91.7% [9].

Another study involved CNNs for the classification and detection of diseases affecting the plant species potatoes, tomatoes, and peppers. Using the Plant Village dataset consisting of 20,636 images from fifteen classes and reported 98.29% training accuracy and 98.029% testing accuracy [10]. A study proposed a system to detect the varieties of plants like Apple, Grape, Potato, Corn, Sugarcane and Tomato and the diseases associated with them comprising of 35000 dataset with an overall average accuracy of 96.25% and the system being able to give 100% accuracy confidence in classification [11]. Lately, object detection methods have been widely used with aim to minimize loss on a given dataset with enhanced accuracy [12], [13]. Models such as Faster R-CNN and You Only Look Once(YOLO) were employed for object detection [12], [13] while ResNet [14] was utilized for image classification.

Object detection models have also been used in detecting plant diseases. [15] explored various deep learning object detection methods like Regions with Convolutional Neural Networks(R-CNN), Region-based Fully Convolutional Network(R-FCN), and SSD, which were employed in the identification of diseases in plant leaves. These models learn complex scenarios from the plants area and hence can have good accuracy. [16] proposed a multistage method used to detect and classify the leaves of the given plant using YOLOV3 as plant detection model and ResNet-18 as classification model. [17] proposed a model that leverages a Conditional Generative Adversarial Network for generation of synthetic data, a Convolutional Neural Network for feature extraction, and a Logistic Regression classifier for quick classification of the plant species. The model was trained for plants like apple, corn, grapes, potato, sugarcane, and tomato, and gave an accuracy of 96.5% on the combined dataset and 99% to 100% on the dataset of individual species.

The use of CNNs has had a significant impact on computer vision tasks, but the accuracy of CNNs can be significantly impacted if the images in the dataset are diverse [18]. Images and frames with unwanted objects besides the area of interest, called background noise, can greatly impact the efficacy of CNNs. The in-situ plant images have unwanted parts such as soil, rocks, and/or human body parts that result in cluttered backgrounds and hence reduced accuracy [18].

[19], [20] found that average accuracy of CNN models can worsen with a higher number of classes, while performance may increase with fewer categories. Similarly, class imbalance or the over representation of some classes relative to others, is a common issue in large-scale image classification datasets. This lopsided class distribution can negatively affect the model's performance [21].

ResNet-50 model was used for image classification due to its' superior performance in image classification task with respect to other models [22]. The architecture of ResNet-50 is of a deep residual in nature, thus preventing vanishing gradient descent. This ultimately helps the model learn effectively the complex features that make it accurate for classification tasks.Likewise, SSD MobileNet V2 architecture has been used for leaf detection since it can quickly process images [23]. Although this model is relatively lower as compared to other models regarding accuracy, such as inception and fasterRCNN, our object detection models do not require extreme complexity or exceptionally high accuracy.

Recent advancement has focused on developing and training classification models based on Vision Transformer(ViT) and CNN and checking if the recent advancements outperforms the previously given models. A research conducted by [24] demonstrated that the ViT2 model outperformed other models and accquired accuracy upto 99.7%. for identifying the diseases in tomato plants. Another research by [25] developed smartphone-based application using a ViT-based model that used self-attention mechanism and achieved accuracy of 90.99% during experimentaion. So, vision based transformers are being widely used in detection of diseases of plants in present time.

III. Methodology

A. Data Collection

The images of the different types of plant species required for research were collected from various sources like GitHub and Kaggle. Images were collected such that there is the inclusion of variety of plants and associated diseases. These images were used for plant leaf detection and image classification.

B. Dataset Preprocessing

All the images were checked for the acceptable files like (jpg, jpeg, png and bmp). This was done to ensure data integrity and to make sure that only acceptable files are processes as valid input. During training of plant detection model, the dataset was checked for distribution of data like class imbalance problem, and oversampling and undersampling was done to ensure that there is uniform distribution of data and making it suitable for training. For the plant leaf detection task the target number for each types of plant was set to 2000. Plants with fewer data samples were oversampled by generating additional samples and for large number of samples the number of samples were reduced. Our original file containing the annotated image information was then changed into csv format which is further changed into TensorFlow Record(TFRecord) format making it compatible with TensorFlow to read and write the large datasets.

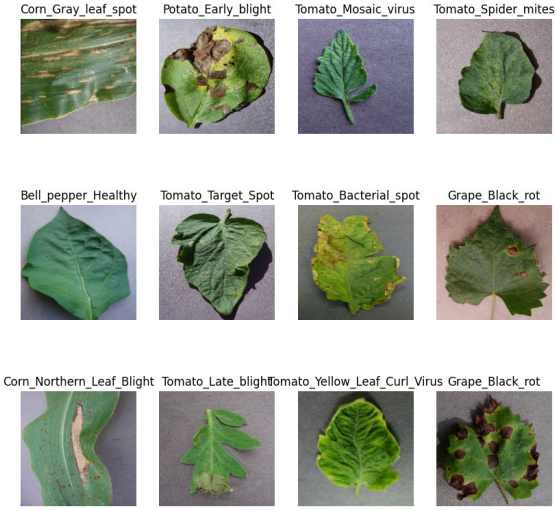


Fig. 1: Sample images from the dataset showing different plant species

During the training of image classification model, images were correctly labeled. The images size were resized to 256*256 to establish the uniformity during training. The images were then normalized in the range of [0,1] for getting stable gradient descent updates during back-propagation and getting faster convergence. The dataset were then splitted in the ratio of 80%, 10% and 10% for train, test and validation data. The sample of images used during training is shown in Figure 1.

C. Data Augmentation and Pipelining

For ensuring the diversity in data and to improve generalization data augmentation is performed. The augmentation was done by flipping the images with horizontal and vertical rotation upto 20%. This process was done during the training phase to ensure that our model learns from the broader dataset. In addition to it the dataset was cached, shuffled, and prefetched for optimization of data pipeline. Caching helps to reduce the data loading time by storing the dataset in memory after first epoch. Shuffling was done having a buffer size of 1000 for presenting the data in random order during training thus preventing overfitting. Prefetching was done to assure that the next batch of data is loaded during the model training and thus making the smooth and efficient training process. To improve the generalization ability of model plants images were collected from different lighting conditions and also the different stages of diseases in a particular plant was taken into consideration. In addition to it the adjustment of contrast and brightness were applied to ensure model in robust among varying environmental condition.

D. Data Analysis and Description

After completing data pre-processing and augmentation, the total number of images required for both models were finalized. The plant leaf detection model is trained on nine different plants, and a separate image classification model is trained for disease classification in each of these plants. The number of plants used in the dataset for the image classification model is enlisted in Table I along with the class names in classification are provided in Table II respectively.

TABLE I: Object Detection and Image Classification Dataset

Plant Species	Object Detection (Number of Images)	Image Classification (Number of Images)
Grape	500	5544
Potato	547	4400
Tomato	703	12523
Apple	562	5511
Corn	545	5643
Bell Pepper	560	3131
Strawberry	462	3012
Peach	604	3112
Cherry	220	3122
Total	**4693**	**40998**

TABLE II: Disease Classification for Different Plants

Plant	Diseases	Number of Classes
Apple	Apple_Black_rot, Apple_Healthy, Apple_Scab,Cedar_Apple_rust	4
Bell Pepper	Bell_pepper_Bacterial_spot, Bell_pepper_Healthy	2
Cherry	Cherry_Healthy, Cherry_Powdery_mildew	2
Corn	Corn_Common_rust, Corn_Gray_leaf_spot, Corn_Healthy, Corn_Northern_Leaf_Blight	4
Grape	Grape_Black_Measles, Grape_Black_rot, Grape_Healthy, Grape_Isariopsis_Leaf_Spot	4
Peach	Peach_Bacterial_spot, Peach_Healthy	2
Potato	Potato_Early_blight, Potato_Healthy, Potato_Late_blight	3
Strawberry	Strawberry_Healthy, Strawberry_Leaf_scorch	2
Tomato	Tomato_Bacterial_spot, Tomato_Early_blight, Tomato_Healthy, Tomato_Late_blight, Tomato_Leaf_Mold, Tomato_Mosaic_virus, Tomato_Septoria_leaf_spot, Tomato_Spider_mites, Tomato_Target_Spot, Tomato_Yellow_Leaf_Curl_Virus	10

IV. Proposed Model Design

The proposed model employs a two-tiered architecture for plant detection and disease classification. At first detection of the species of plant is done using SSD MobileNet V2 and then diseases associated with that plant is identified using ResNet-50 model architecture. The workflow of the proposed model is shown in Figure 2 with the description in the following subsections.

A. Plant Detection

Object detection works by classifying and detecting objects by creating bounding boxes around them. In the proposed plant detection model, SSD Mobilenet V2 is used to identify the plant species, mentioned in Table I. SSD Mobilenet V2 is selected due to its lightweight nature and ability to balance computational efficiency with accuracy, making it suitable for mobile devices. This step confirms that only the relevant part

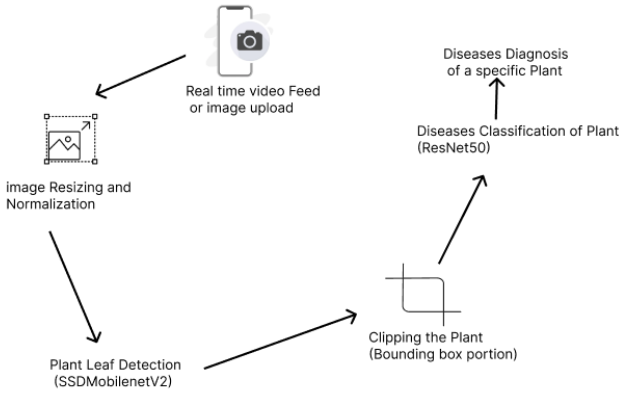


Fig. 2: Workflow of Proposed Model

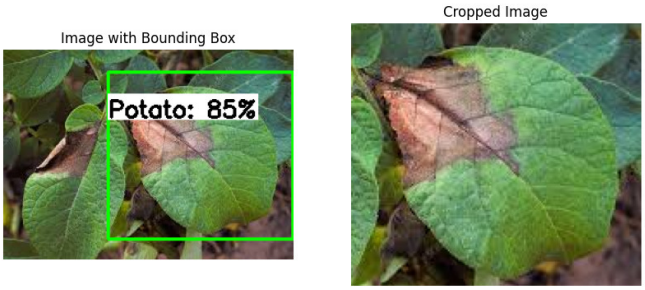


Fig. 3: Cropped region used for Potato disease classification

of the image is further processed for image classification and the clustered or irrelevant background is clipped out. This will make the image classification model only to focus on the foreground provided by clipping the bounding box portion given by first step. The process begins by detecting leaves in the image followed by selection of the identified leaf region.This is illustrated in Figure 3, Figure 4, and Figure 5, where each step in the detection and selection pipeline is visualized.

B. Diseases Classification

After the completion of above step the clipped portion of image is passed to ResNet-50 model to classify the diseases of a detected plant. ResNet-50 model architecture incorporates the concept of residual learning thus allowing the model to accurately classify complex pattern making it ideal to classify various plant diseases. It ensures the precise and reliable diseases diagnosis as it has to focus only on the relevant part of the image having plant leaf.

V. Results and Discussion

A. Plant Detection Model

The performance of plant detection model was evaluated using a method called mean average precision(mAP). This

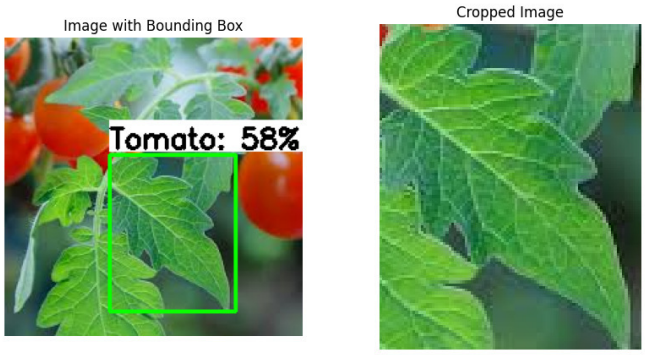


Fig. 4: Cropped region used for Tomato disease classification

Fig. 5: Cropped region used for Apple disease classification

metric is commonly used in the object detection model's performance evaluation. The mAP is combination of precision and recall for each class giving the average value of the model performance. The mAP @ 0.5:0.95 for different plants is presented in Table III.

TABLE III: mAP @ 0.5:0.95 for Plant Detection Model

Plant	**mAP (%)**
Tomato	40.57
Peach	54.74
Corn	78.64
Apple	69.45
Potato	53.43
Bell Pepper	38.50
Cherry	81.88
Grape	75.27
Strawberry	64.65
Overall	**60.91**

The overall mAP was obtained to be 0.62 resulting good accuracy for plant detection model. The plant detection model has to be lightweight making it run effectively on mobile devices in real-time. Hence, this model is made less complex and maintaining reasonable accuracy.

B. Diseases Classification model

The performance of classification model was evaluated using accuracy, f1-score and confusion matrix. The f1 score is the combination of precision and recall and tells us how well is our model performing. Additionally, the confusion matrix gives us the description of True Positive(TP), True Negative(TN), False Positive(FP) and False Negative(FN) allowing to compare the true labels with the predicted ones.

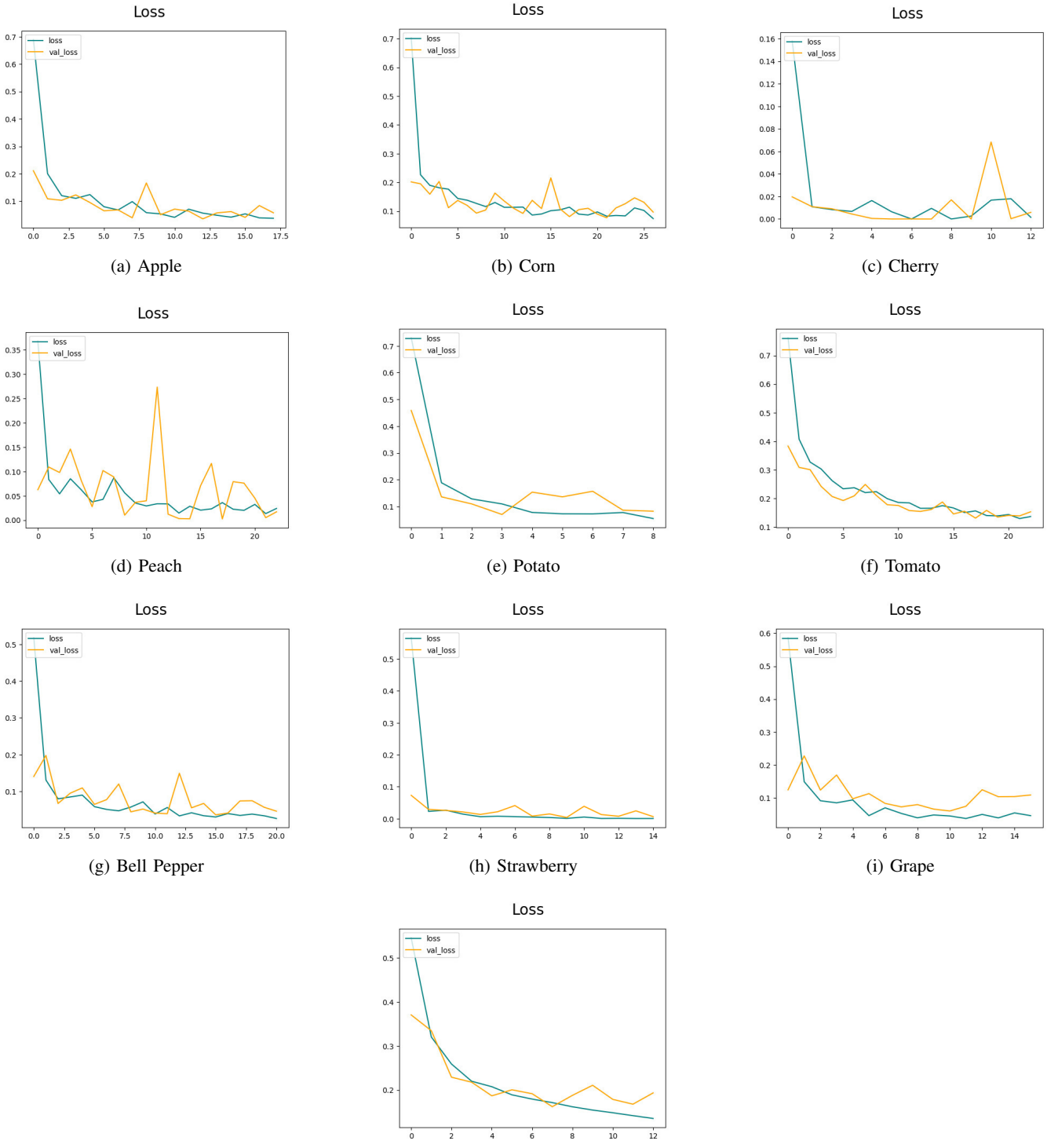
The Figure 6 illustrates the loss plots for the classification models of various plant species, including Apple, Corn, Cherry, Peach, Potato, Tomato, Bell Pepper, Strawberry, Grape, as well as the Combined Model. All the above mentioned loss plots are used as the performance evalutation metric of the ResNet-50 model. In each figures we can observe that the both training loss and validation loss are decreasing which assures that the model is performing well and is converging after each epochs. The fluctuations in the loss plots represent that the model is attempting to adjust it's weight parameters to minimize the error during training. But as the epoch are increasing the loss is decreasing indicating that model learns and improves. The fluctuations are common and indicate that during certain epochs the model is encountering challenges in certain epochs leading certain increase in loss during some epochs. To confirm there is no overfitting the early stopping techniques are applied with a patience of five. Likewise the best weights are taken as the final weights for classification model.

Figure 6j represents the combined model having 33 classes of all plants diseases. The loss value is decreasing after each epochs showing that the model is learning and converging. In comparison, the average test loss value for the individual plant models is 0.06937 with a standard deviation of 0.074, whereas the test loss value for the combined model is 0.182. This indicates that the study that we have done to isolate the foreground from the cluttered background and making the image classification model focus only on the relevant part of plant has the increased performance than the combined model having all thirty three classes. Additionally it suggests that the model with trained with less number of classes tends to perform more accurately than will higher number of classes in the plant detection task.

The Figure 7 presents the confusion matrices for various plant species, including Apple, Corn, Cherry, Peach, Potato, Tomato, Bell Pepper, Strawberry, and Grape. Additionally, the confusion matrix for the combined model is also included in Figure 7j. Each matrix provides insights into the model's performance in accurately classifying the respective plant species. The confusion matrix of different plants under study reveals that all the plant models are performing very well with high accuracy in testing datasets and effectively classifying the diseases related to a specific plant species. The high diagonal values indicate that models are very effective and the values outside the diagonal are lesser for most of the plants indicating a high accuracy is maintained. Likewise, after viewing the combined plant model having 33 classes it also has high diagonal values indicating the model performs well in the combined model as well but on average comparison the individual plant model outperforms the combined model.

The performance of the image classification model for different plants is summarized in Table IV. The table presents the test loss and accuracy along with the F1-score for each plant.

The proposed two-tier architecture for the plant leaf detection and it's diseases classification demonstrated high accuracy



(a) Apple (b) Corn (c) Cherry

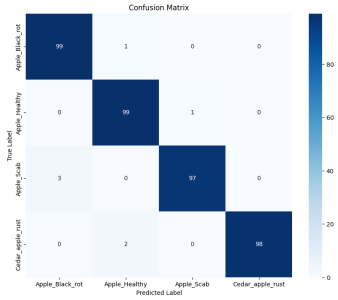
(d) Peach (e) Potato (f) Tomato

(g) Bell Pepper (h) Strawberry (i) Grape

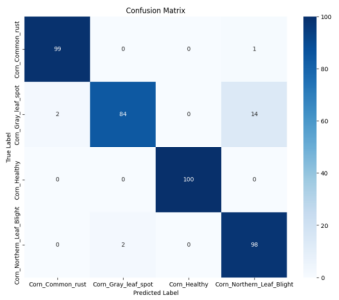
(j) Combined Model

Fig. 6: Loss plots for different plants and the combined model.

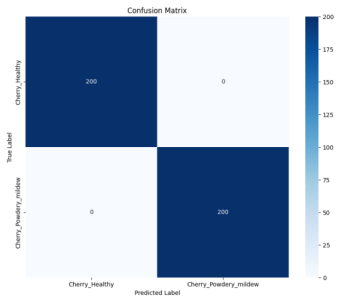
of 98.3% as compared to the combined model. The suggested approach narrowed down the classification task yielding high accuracy. In contrast, the plant species with combined 33 classes has an accuracy decreased to 94.70%. The reduction in accuracy in the combined model highlights the benefits of isolating plant species before disease classification, which simplifies the model's task and enhances overall accuracy. Similarly, the input images' cluttered and rough backgrounds



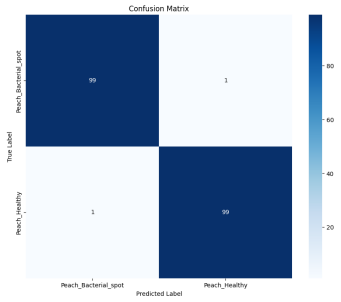
(a) Apple



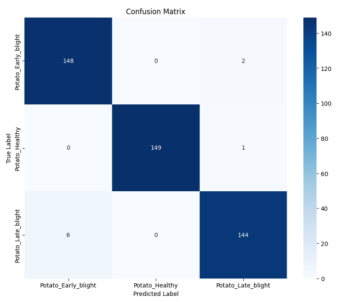
(b) Corn



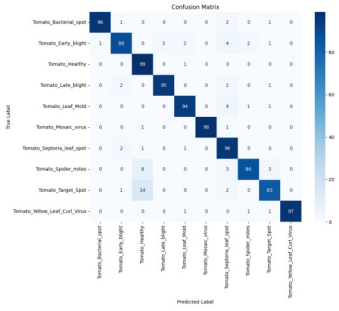
(c) Cherry



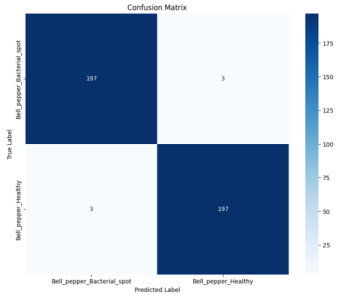
(d) Peach



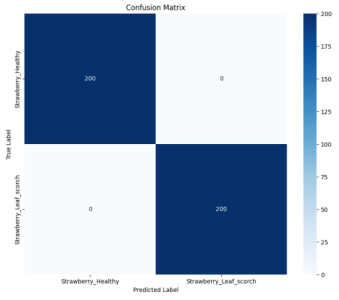
(e) Potato



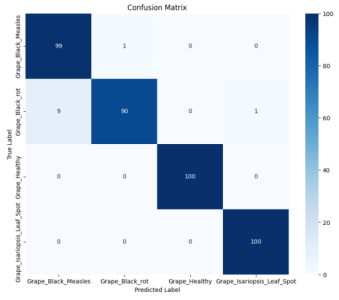
(f) Tomato



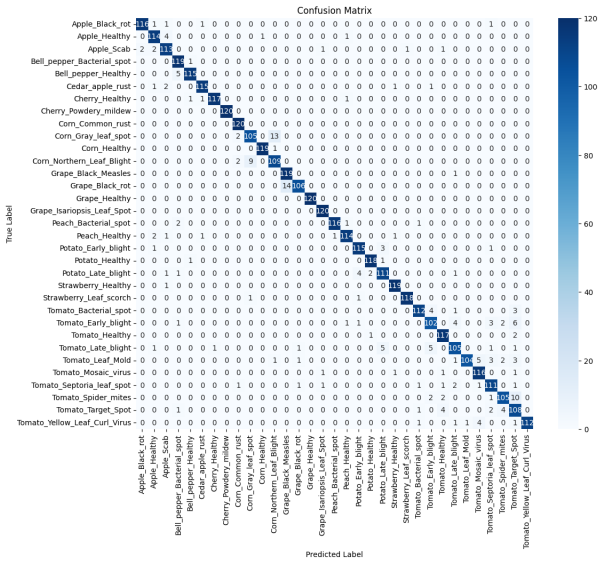
(g) Bell Pepper



(h) Strawberry



(i) Grape



(j) Combined Model

Fig. 7: Confusion matrices for different plants and the combined model.

TABLE IV: Image Classification Results

Plant	Test Loss	Test Accuracy (%)	F1 Score
Apple	0.0802	98.65	0.98
Corn	0.1123	97.5	0.96
Peach	0.0388	99.50	0.99
Cherry	0.000335	100.00	1.00
Grape	0.0605	97.25	0.97
Strawberry	0.0020	99.84	0.99
Bell Pepper	0.0282	98.50	0.98
Tomato	0.152837	95.70	0.95
Potato	0.0567	98.00	0.98
Average	–	**98.30**	**0.98**

made it difficult for the combined model to extract pertinent information, which decreased the accuracy of the classification. On the other hand, the two-tiered model successfully identified the plant in the picture and then carried out the disease classification. Using this method there is a less chance of misidentification of one plant's diseases as another thus increasing prediction accuracy.The ability to first isolate the plant before classification allowed the model to handle complex backgrounds more effectively, contributing to the overall improvement in performance.

Similarly it can also be observed that the individual classification models where there are higher number of classes tends to have lower accuracy when compared with the models with fewer classes. In addition to it there was plant species like Corn, Grape, Tomato where there was impact of background noise in the dataset and the spots on the leaves for diseases detection seems to be similar which made it challenging for the model to accurately classify the diseases. Because of these factors, the accuracy of certain plant diseases detection model has declined in comparision to others as seen in the confusion matrix.

C. Comparative Analysis

In this section we compare our model with the similar and existing systems. The comparison is based on the accuracy, the number of data samples, dataset augmentation used and the model architecture.The comparison is given in the following table.

The table V gives the comparative analysis of our proposed model with some of the existing models on the basis of the testing accuracy, number of samples used, number of classes and the model architecture. From the table it is clearly observable that our proposed model with an of SSD Mobilenet V2 and ResNet-50 outperforms other existing models either with the less dataset used or the high accuracy with the less number of training samples as compared to others. Our proposed model addresses the significant improvement in the image classification task specially in the complex and cluttered backgrounds.

Similarly in the comparative anaysis of the dataset augmentation [16] and [10] applied the advanced data augmentation technique like brightness variations and image shearing of images. Other models applied simple augmentation techniques like flipping and rotation and the absence of techniques like contrast variation, brightness variation can negatively impact the performance of model on new and unseen datatsets. Since our model incorporated advanced augmentation techniques like varying light conditions images and adjustment in contrast and brightness, it has demonstrated high accuracy and greater robustness as compared to other models shown in Table V.

Having the entire focus on relevant section of plant our model enhances the performance in terms of accuracy for the real-time applications in agriculture. The successful implementation of this can address the technological gap between the farmers and improved agricultural methods resulting in decreased diseases related loss and thereby improving plants management and increasing productivity.

VI. Conclusion

The proposed model effectively detects plants and classifies plant diseases using SSD Mobilenet V2 for object detection and ResNet-50 for disease classification. The performance and accuracy results achieved through the proposed model prove very promising in further enhancing agricultural practices. The promising results indicate the approach to be significantly benefit in plant health monitoring.

Despite the encouraging outcomes, there is still a room for further improvement that can be achieved by increasing the number of available training data on the plant species. This would further lead to a much-improved performance with regard to the lower detection and classification accuracy in plant species. Improvement can still be attained by trying other more advanced object detection algorithms for model's robustness under varying environmental conditions. While our model is typically deployed as mobile application it can work well in both the controlled environment and in various weather conditions. Only the factor that can impact the performance of model is because of the background noised in the real-time image frames and other factors such as motion blur, poor lighting conditions and occlusions.

The proposed model will also be further enhanced to support the extended range of plant species within the application and adaptation of the model for diverse geographical locations and farming practices. With these in consideration, our model aims to create a holistic agricultural management tool that strives for better health and productivity in crops, considering the factors of the above mentioned.

References

[1] B. R. Neupane, "Contribution of expenditure to agriculture growth in nepal," pp. 119–131, 2023. [Online]. Available: https://doi.org/10.3126/qjmss.v5i1.56502

[2] S. Sakamoto, W. Putalun, S. Vimolmangkang, W. Phoolcharoen, Y. Shoyama, H. Tanaka, and S. Morimoto, "Enzyme-linked immunosorbent assay for the quantitative/qualitative analysis of plant secondary metabolites," *Journal of natural medicines*, vol. 72, pp. 32–42, 2018.

[3] F. Martinelli, R. Scalenghe, S. Davino, S. Panno, G. Scuderi, P. Ruisi, P. Villa, D. Stroppiana, M. Boschetti, L. R. Goulart *et al.*, "Advanced methods of plant disease detection. a review," *Agronomy for sustainable development*, vol. 35, pp. 1–25, 2015.

[4] K. Alemu, "Real-time pcr and its application in plant disease diagnostics," *Adv. Life Sci. Technol*, vol. 27, pp. 39–49, 2014.

TABLE V: Comparison of Proposed Model with Existing Systems

Study	Model Architecture	Number of Sample Images	Dataset Used	Number of Classes	Accuracy
[10]	CNN	20,636	PlantVillage	12	98.02%
[16]	YOLOV3+ResNet-18	36,000	PlantVillage	29	96%
[11]	CNN	35,000	Not Mentioned	21	96.5%
[9]	CNN with squeeze-and-excitation(SE)	-	Dataset from challenger.ai	10	91.7%
Proposed Model	**SSD MobileNet V2 + ResNet-50**	**40,998**	**Custom + Augmented Data**	**33**	**98.3%**

[5] S. Arivazhagan, R. N. Shebiah, S. Ananthi, and S. V. Varthini, "Detection of unhealthy region of plant leaves and classification of plant leaf diseases using texture features," *Agricultural Engineering International: CIGR Journal*, vol. 15, no. 1, pp. 211–217, 2013.

[6] C. Yang, "Remote sensing and precision agriculture technologies for crop disease detection and management with a practical application example," *Engineering*, vol. 6, no. 5, pp. 528–532, 2020.

[7] N. Gogoi, B. Deka, and L. Bora, "Remote sensing and its use in detection and monitoring plant diseases: A review," *Agricultural Reviews*, vol. 39, no. 4, pp. 307–313, 2018.

[8] Z. Wang, K. Wang, F. Yang, S. Pan, and Y. Han, "Image segmentation of overlapping leaves based on chan–vese model and sobel operator," *Computers and Electronics in Agriculture*, vol. 148, pp. 51–58, 2018.

[9] J. Hang, D. Zhang, P. Chen, J. Zhang, and B. Wang, "Classification of plant leaf diseases based on improved convolutional neural network," *Sensors*, vol. 19, no. 19, 2019. [Online]. Available: https://www.mdpi.com/1424-8220/19/19/4161

[10] M. A. Jasim and J. M. AL-Tuwaijari, "Plant leaf diseases detection and classification using image processing and deep learning techniques," pp. 259–265, 2020.

[11] S. V. Militante, B. D. Gerardo, and N. V. Dionisio, "Plant leaf detection and disease recognition using deep learning," in *2019 IEEE Eurasia conference on IOT, communication and engineering (ECICE)*. IEEE, 2019, pp. 579–582.

[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, 2015. [Online]. Available: https://arxiv.org/abs/1506.01497

[13] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 779–788, 2016. [Online]. Available: https://doi.org/10.1109/CVPR.2016.91

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90

[15] M. Akila and P. Deepan, "Detection and classification of plant leaf diseases by using deep learning algorithm," *International Journal of Engineering Research & Technology (IJERT)*, vol. 6, no. 7, pp. 1–5, 2018.

[16] A. Venkataramanan and P. Agarwal, "Plant disease detection and classification using deep neural networks," 08 2019.

[17] P. S. Kanda, K. Xia, and O. H. Sanusi, "A deep learning-based recognition technique for plant leaf classification," *IEEE Access*, vol. 9, pp. 162 590–162 613, 2021.

[18] K. KC, Z. Yin, D. Li, and Z. Wu, "Impacts of background removal on convolutional neural networks for plant disease classification in-situ," *Agriculture*, vol. 11, no. 9, p. 827, 2021, this article belongs to the Special Issue Latest Advances for Smart and Sustainable Agriculture. [Online]. Available: https://www.mdpi.com/2077-0472/11/9/827

[19] Y. Zhang, J. Zhang, Y. Wang, and X. Liu, "How does the data set and the number of categories affect cnn-based image classification performance?" *Journal of Software*, vol. 14, no. 4, pp. 168–181, 2019.

[20] A. K. Ali, A. M. Abdullah, and S. F. Raheem, "Impact of the classes' number on the convolutional neural networks performance for image classification," *International Journal of Advanced Science Computing and Engineering*, vol. 5, no. 2, pp. 119–128, 2023.

[21] P. P. Wagle and M. Manoj Kumar, "A comprehensive review on the issue of class imbalance in predictive modelling," *Emerging Research in Computing, Information, Communication and Applications: Proceedings of ERCICA 2022*, pp. 557–576, 2022.

[22] S. Mascarenhas and M. Agarwal, "A comparison between vgg16, vgg19 and resnet50 architecture frameworks for image classification," in *2021 International Conference on Disruptive Technologies for Multi-Disciplinary Research and Applications (CENTCON)*, vol. 1, 2021, pp. 96–99.

[23] N. N. F. Giron, R. K. C. Billones, A. M. Fillone, J. R. Del Rosario, A. A. Bandala, and E. P. Dadios, "Classification between pedestrians and motorcycles using faster r-cnn inception and ssd mobilenetv2," pp. 1–6, December 2020.

[24] R. A. Boukabouya, A. Moussaoui, and M. Berrimi, "Vision transformer based models for plant disease detection and diagnosis," in *2022 5th International Symposium on Informatics and its Applications (ISIA)*. IEEE, 2022, pp. 1–6.

[25] U. Barman, P. Sarma, M. Rahman, V. Deka, S. Lahkar, V. Sharma, and M. J. Saikia, "Vit-smartagri: vision transformer and smartphone-based plant disease detection for smart agriculture," *Agronomy*, vol. 14, no. 2, p. 327, 2024.

rice2024 Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering pp. 89–96
DOI: 10.15439/2024R115
ISSN 2300-5963 ACSIS, Vol. 42

Enhancing YOLOv11 for Real-Time Object Detection: Advanced Architectures and Edge-Optimized Training Pipeline

Sivadi Balakrishna
0000-0002-8939-9307
Department of Advanced Computer Science & Engineering, Vignan's Foundation for Science, Technology & Research, Vadlamudi, Guntur, A. P, India.
drsivadibalakrishna@gmail.com

Shivani Yadao
0000-0002-2953-778X
Department of Computer Science & Engineering, Stanley College of Engineering andTechnology for Women, Hyderabad, Telangana, India.
shivaniyadao123@gmail.com

Vijender Kumar Solanki
0000-0001-5784-1052
Department of Computer Science & Engineering, Stanley College of Engineering and Technology for Women, Hyderabad, Telangana, India.
spesinfo@yahoo.com

***Abstract*—In this paper, we propose novel enhancements to YOLOv11, leveraging its advanced architectural components such as the C3k2 block, SPPF (Spatial Pyramid Pooling - Fast), and C2PSA (Convolutional Block with Parallel Spatial Attention). These innovations address key challenges in real-time object detection, including feature extraction, attention mechanisms, and computational efficiency. Furthermore, we present a new training pipeline that optimizes YOLOv11 for edge computing while maintaining state-of-the-art accuracy. Experimental results on the COCO dataset demonstrate significant improvements in mean Average Precision (mAP) and latency compared to prior YOLO iterations, establishing YOLOv11 as a benchmark for real-time applications.**

***Index Terms*—Object detection, Real-time systems, training-pipeline, YOLO, deep learning.**

I. Introduction

A WIDE variety of computer vision applications rely on real-time object identification, including autonomous cars, surveillance, and augmented reality. The YOLO framework has revolutionized this field by enabling fast and accurate detection through single-stage networks [4-5]. However, traditional approaches primarily rely on single-modal data (e.g., RGB images), limiting their ability to handle complex scenarios like occlusion, low-light conditions, and ambiguous object boundaries.

The rapid growth of computer vision applications has heightened the demand for efficient and accurate object detection models. YOLO (You Only Look Once) frameworks have historically set benchmarks in this domain by combining high speed and accuracy in a single-stage architecture. YOLOv11 introduces significant enhancements to this lineage, incorporating novel components like the C3k2 block, SPPF, and C2PSA to address challenges such as occlusion, small object detection, and resource constraints in edge deployments [6-9]. The goal of this research is to provide a thorough evaluation of the YOLO algorithm's development over time. By providing the first in-depth analysis of YOLO11, the most recent addition to the YOLO family, it significantly advances the state of the art. We assess the efficacy of fine-tuned pre-trained models on three unique bespoke datasets, ranging in size and purpose. Consistent hyperparameters are used to guarantee an objective and fair comparison. Critical performance indicators such as computational complexity (as defined by GFLOPs count and model size), accuracy, efficiency, and speed are examined in the research [10]. Further, we look at how each YOLO variant is implemented, comparing and contrasting their advantages and disadvantages in various scenarios. By comparing these models, we want to show how they might be used effectively in different situations, which will be useful for scholars and practitioners. Fig.1 depicts the YOLO series models evolution time line over the years.

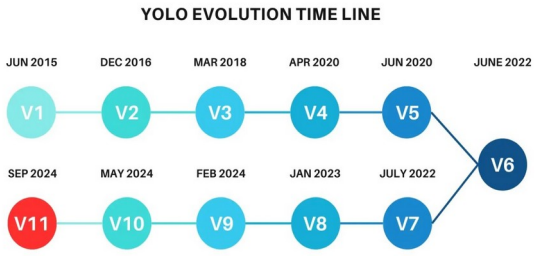


Fig1: YOLO series models Evolution Time Line

In this paper, we explore the architectural innovations in YOLOv11 and propose further optimizations that enhance its performance for real-time applications. Our contributions include:

- **Architectural Enhancements**: Introduction of C3k2 and C2PSA blocks for improved feature extraction and detection accuracy.

- **Edge-Optimized Training**: Quantization-aware training and refined loss functions for efficient edge deployment.
- **Performance Improvements**: Achieved higher mAP (+1.7%) and reduced latency (-7.6%) compared to YOLOv10.
- **Scalability and Applications**: Effective for instance segmentation, pose estimation, and edge device deployments.

The remaining sections have been organized as follows: Section 2 deliberates on related studies of the YOLO frameworks. Section 3 discusses the Proposed Methodology with a new architectural design. Section 4 talks about the proposed models' Results and analysis and applications. Finally, the section 5 concludes with major advancements.

II. Literature Survey

The evolution of YOLO models reflects a steady progression in addressing the trade-offs between speed and accuracy. Outside of the YOLO framework, other object detection architectures have significantly contributed to the field. SSD (Single Shot MultiBox Detector) is known for its balance of speed and accuracy by using multi-scale feature maps for predictions. Faster R-CNN, a two-stage detector, excels in accuracy but often struggles with real-time applications due to higher computational requirements. More recently, DETR (DEtection TRansformer) has introduced transformer-based attention mechanisms, simplifying the object detection pipeline but requiring substantial computational resources. The YOLOv11 builds upon these advancements by combining the speed advantages of YOLO with innovations in attention mechanisms and feature extraction inspired by transformer-based approaches. By optimizing for edge devices and maintaining scalability, YOLOv11 seeks to bridge the gap between lightweight efficiency and state-of-the-art accuracy.

Table 1 shows the evolution of the YOLO series models year-wise with tasks and frameworks involved in those models.

Table 1: Evolution of YOLO series models

Model & Year	Tasks	Frameworks
YOLO [11], 2015	Object Detection, Basic Classification	Darknet
YOLOv2 [12], 2016	Object Detection, Improved Classification	Darknet
YOLOv3 [14], 2018	Object Detection, Multi-scale Detection	Darknet
YOLOv4 [15], 2020	Object Detection, Basic Object Tracking	PyTorch
YOLOv5 [16], 2020	Object Detection, Basic Instance Segmentation	PyTorch
YOLOv6 [17], 2022	Object Detection, Instance Segmentation	PyTorch
YOLOv7 [18], 2022	Object Detection, Object Tracking, Instance Segmentation	PyTorch
YOLOv8 [19], 2023	Object Detection, Instance Segmentation, Panoptic Segmentation	PyTorch
YOLOv9 [20], 2024	Object Detection, Instance Segmentation	1PyTorch
YOLOv10 [21], 2024	Object Detection	PyTorch
VOLOv11 [22], 2024	Object Detection Object Tracking	PyTorch

YOLO is a powerful and effective one-stage object identification approach. By allowing the prediction of bounding boxes and class probabilities directly from whole pictures in a single assessment, YOLO revolutionised object recognition with its 2015 introduction by Redmon et al. [1]. Using this innovative approach, YOLOv1 [11] achieved extremely accurate object identification in real time. Building upon this foundation, YOLOv2 [12] implemented several noteworthy enhancements. Improved feature extraction was made possible by using the Darknet19 framework, which is a 19-layer convolutional neural network. For better model generalisation, YOLOv2 used data augmentation approaches inspired by the VGG architecture [13] and incorporated batch normalisation. The Darknet-53 architecture, a deeper network that considerably increased the model's capabilities for feature extraction, was utilised by YOLOv3 [14] to augment it. This variation used a design influenced by Feature Pyramid Networks (FPN) to increase identification accuracy for objects of varying sizes by mixing low-level detailed data with high-level semantic information and employing a Three-Scale detection process.

The evolution of YOLO models reflects a steady progression in addressing the trade-offs between speed and accuracy. YOLOv3 introduced multi-scale detection, while YOLOv4 [15], YOLOv5 [16], YOLOv6 [17], YOLOv7 [18], and YOLOv8 [19] expanded functionality to instance segmentation and panoptic tasks. VOLOv9[20], and YOLOv10's [21] are NMS-free designs that marked a leap in training efficiency. Despite these advancements, challenges persist in balancing model size, speed, and accuracy for real-time applications. The YOLOv11[22] builds upon this foundation with innovative architectural elements, which we further optimize in this study to maximize its potential for edge computing and constrained environments.

After these studies, we realize that there is scope to improve the YOLOv11 model with significant changes. Therefore, we proposed some possible architectural advancements in the YOLOv11 model.

III. Proposed Work

In this section, the proposed architectural enhancements for the YOLOv11 model with advanced architectures and the edge-optimized training pipeline is discussed. It also includes the YOLOv11 architectural diagram with a detailed explanation of the components involved in it.

A. Architectural Enhancements

1) Enhanced Backbone: C3k2 Block

The C3k2 block, a lightweight version of the CSP bottleneck, uses smaller kernel sizes for faster processing. Unlike previous iterations, our enhancement integrates dynamic kernel adjustments that adapt to varying input resolutions, improving efficiency and flexibility. By doing so, the backbone captures fine-grained features essential for accurate detection without increasing computational overhead. The first step in YOLOv11's process is to down-sample the input picture using a sequence of convolutional layers.

$$Conv1 = Conv(1, 64, 3, 2) \quad (1)$$

$$Conv2 = Conv(Conv1, 128, 3, 2) \quad (2)$$

The YOLOv11 switches out the inefficient C2F block with the Cross-Stage Partial (CSP) network-based C3k2 block. In order to reduce computing cost while keeping performance constant, the C3k2 block employs two smaller convolutions, with a kernel size of 2. This block's equation is displayed below:

$$c3k2(X) = Conv(Split(X)) + Conv(Merge(Split(X))) \quad (3)$$

2) Attention-Driven Neck: C2PSA Block

The C2PSA block combines spatial pooling with attention mechanisms to prioritize critical regions in feature maps. By pooling features spatially, it enhances focus on regions of interest, such as small or occluded objects, improving detection accuracy. Our proposed adaptive attention strategy dynamically reallocates focus based on object density within images, further enhancing the robustness of detection in cluttered scenes. YOLO11 retains the SPPF block for multi-scale spatial pooling. As described as follows.

$$SPFF(X) = Concat(MaxPool(X, 5), MaxPool(X, 3), MaxPool(X, 1)) \quad (4)$$

The C2PSA blocks improve spatial attention across feature maps in YOLO11. This enhances model performance by focussing on key visual areas for detection, particularly for tiny and obstructed objects.

$$C2PSA(X) = Attention(Concat(Xpath1, Xpath2)) \quad (5)$$

3) Optimized Head with CBS Blocks

CBS(Convolution-BatchNorm-SiLU) blocks refine feature maps before the final detection layers. To address challenges in detecting small and occluded objects, we introduce multi-scale CBS configurations. These configurations process feature maps at different depths, ensuring that the model can accurately detect objects of varying sizes and complexities.

$$Detect(P3, P4, P5) = BoundingBoxes + ClassLables \quad (6)$$

B. Neck Design

The neck component aggregates and transmits feature maps from the backbone to the head, enabling multi-scale detection. YOLOv11 replaces the traditional C2F block with the advanced C3k2 block in the neck. This change enhances the feature aggregation process, reducing latency while improving detection precision. The neck also incorporates upsampling layers to merge features from different resolutions, ensuring that global and local information contributes to the detection process. The YOLO11 neck collects feature maps and sends them to the detecting head at various resolutions. To accelerate feature aggregation, YOLOv11 adds the C3k2 block to the neck. Upsampling and concatenation layers are applied by the neck to merge the feature maps of various sizes. This process is known as feature aggregation.

$$Featureupsample = Upsample(Featureprevious) \quad (7)$$

$$Featureconcat = Concat(Featureupsample, Featurelower) \quad (8)$$

After concatenation, the C3k2 block efficiently aggregates features:

$$C3k2neck = Convsmall(Concat(Featureconcat)) \quad (9)$$

Spatial Attention: The C2PSA block in YOLOv11's neck promotes spatial attention, helping the model focus on the most relevant regions of the picture in congested environments with overlapping objects.

C. Prediction Head

AE-YOLOv11's head employs a combination of C3k2 blocks and CBS (Convolution-BatchNorm-Silu) layers to refine multi-scale feature maps. Key enhancements include:

- **Multi-Scale Prediction:** The head processes feature at various depths to generate predictions for bounding box coordinates, objectness scores, and class probabilities.
- **Efficient Final Layers:** The inclusion of lightweight convolutional layers reduces computational complexity while maintaining output quality.
- **Customizable Configurations:** The C3k2 blocks in the head adapt based on the specific model variant (e.g., nano, small, medium), enabling scalability and flexibility.

D. Edge-Optimized Training Pipeline

To optimize YOLOv11 for resource-constrained environments, we propose the following strategies:

- **Augmented Data Sampling:** Incorporate context-aware augmentation techniques to improve robustness against background clutter and varied lighting conditions.
- **Efficient Loss Functions:** Refine the combined localization, confidence, and classification loss to minimize false positives in dense scenes, ensuring accurate predictions in real-world scenarios.

- **Quantization-Aware Training:** Introduce quantization techniques during training to reduce model size and latency, preparing it for deployment on low-power devices without sacrificing accuracy.

Fig.2 shows the Detailed Component Breakdown of the AE-YOLOv11 model.

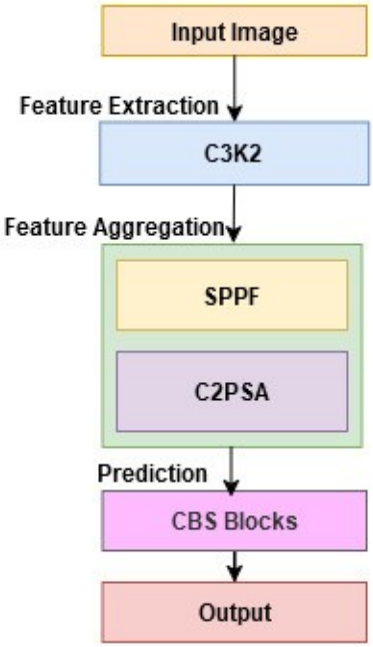


Fig.2: Architectural components for AE-YOLOv11

A. Backbone

- **Role**: Extracts low- to high-level features from input images.
- **Key Modules**:
  a. **Initial Convolution Layers**:
     - Perform downsampling.
     - Use Conv + BN (BatchNorm) + SiLU (Sigmoid Linear Unit) for non-linearity.
  b. **C3k2 Block**:
     - A novel module designed for efficient feature extraction.
     - Splits convolutions into smaller kernels (e.g., kernel size 2).
     - Reduces computational overhead while maintaining performance.
  c. **SPPF (Spatial Pyramid Pooling - Fast)**:
     - Captures multi-scale features by pooling at different scales.
     - Aggregates global context effectively, improving detection accuracy.

B. Neck

- **Role**: Aggregates features across scales and enhances spatial resolution.
- **Key Modules**:
  a. **Upsampling Layers**:
     - Upsample features to match the resolution of previous layers.
     - Enable multi-scale aggregation for better localization.
  b. **C2PSA Block**:
     - Combines spatial pooling and attention mechanisms.
     - Focuses on high-importance regions in images.
     - Improves detection of small or occluded objects.

C. Head

- **Role**: Produces final outputs (bounding boxes, class probabilities, etc.).
- **Key Modules**:
  - **CBS Blocks**:
    - Refine aggregated feature maps.
    - Stabilize data flow using BatchNorm and SiLU activation.
  - **Prediction Layers**:
    - Use multi-scale predictions to detect objects of various sizes.
    - Outputs include:
      - **Bounding Box Coordinates**: Localize objects.
      - **Objectness Scores**: Indicate object presence.

Class Labels: Classify objects

IV. Results and Discussions

This section discusses the performance metrics and datasets used for the YOLOv11 model comparative analysis. Also, deliberates the implementation specifications and comparative study over the existing benchmarked YOLO series models. The practical applications of the YOLO model over various enriched solutions have been discussed.

A. Datasets and Metrics

Experiments were conducted using the COCO dataset to evaluate mean Average Precision (mAP) performance and inference latency. Additional datasets, such as PASCAL VOC and custom datasets for medical imaging, were employed to test domain-specific performance.

B. Implementation

The model was implemented using PyTorch and trained on NVIDIA GPUs. Hyperparameters, such as learning rate and batch size, were optimized to balance training speed and accuracy.

C. Performance Comparison

We evaluate AE-YOLOv11 and its proposed enhancements to the COCO dataset. The key results of our investigation are depicted in Table 2.

Our optimizations yield a 1.7% mAP improvement and a 7.6% reduction in latency compared to baseline YOLOv11.

TABLE2: COMPARATIVE RESULTS OF THE AE-YOLOV11 WITH EXISTING BENCHMARKED MODELS OVER SEVERAL PERFORMANCE METRICS

Model	mAP (%)	Latency (ms)	Params (M)
YOLOv10 [16]	52.1	15	50
YOLOv11[17]	54.5	13	45
AE-YOLOv11 (Ours)	56.2	12	42

These results demonstrate the effectiveness of our architectural and training enhancements.

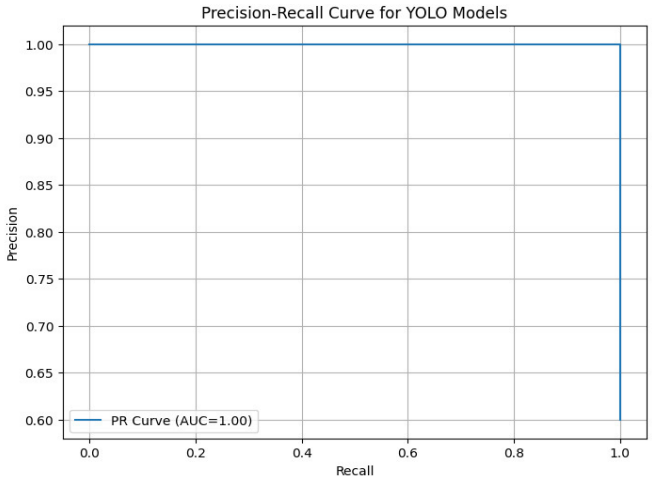


Fig.3: Precision-Recall Curve for YOLO models

The comparative mAP results of the various YOLO series models under latency on the COCO dataset. These results shows that the YOLOv11 model performs better than the other existing benchmarked models used for comparative study.

We deployed YOLOv11 on an NVIDIA Jetson Nano to validate its performance in constrained environments. The optimized model achieved an average inference speed of 25 FPS, outperforming previous YOLO variants in speed and energy efficiency. This demonstrates the practicality of our optimizations for real-time edge applications. Fig.3 shows the Precision-Recall results for YOLO models.

1) Detailed Analysis

a) Accuracy vs. Speed Trade-offs:

- The YOLOv11 series demonstrates remarkable scaling properties, offering smaller models (e.g., YOLOv11-nano) for edge devices and larger models (e.g., YOLOv11x) for high-performance computing.
- The nano variant achieves acceptable mAP scores as shown in Fig.4 for lightweight applications, while the xlarge variant surpasses state-of-the-art accuracy in real-time detection tasks.

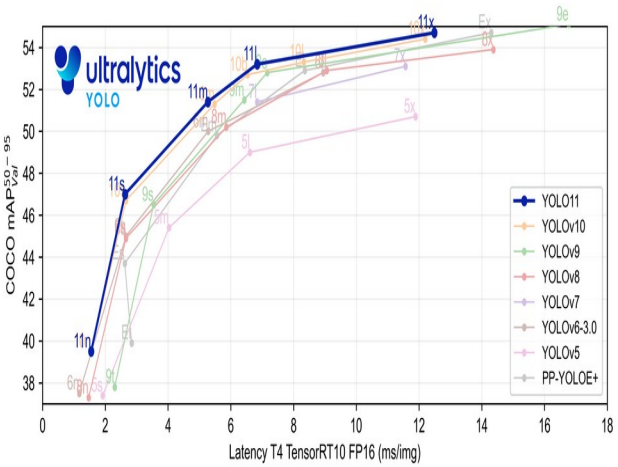


Fig.4: Performance analysis of the YOLO series models on the COCO dataset with mAP findings.

b) Enhanced Detection for Small Objects:

- The inclusion of the C2PSA block significantly enhances the detection of small and partially occluded objects, addressing a common limitation in prior YOLO versions.

c) Comparisons with SSD:

- YOLOv11 achieves faster inference times compared to Single Shot MultiBox Detector (SSD) while offering improved accuracy across diverse datasets. Unlike SSD, which struggles with small object detection, YOLOv11's advanced attention mechanisms deliver superior results.

d) Comparisons with Faster R-CNN:

- While Faster R-CNN is known for its high accuracy, YOLOv11 balances this with real-time performance. YOLOv11's end-to-end single-stage architecture reduces latency, making it a better fit for applications requiring instantaneous results.

e) Multi-Task Capabilities:

- YOLOv11 excels in instance segmentation and pose estimation tasks, with specialized variants (e.g., YOLOv11-seg, YOLOv11-pose) achieving superior results on datasets like COCO and custom benchmarks.

f) Energy Efficiency:

- The reduced parameter counts in YOLOv11's backbone and neck ensure energy-efficient deployments, critical for battery-operated devices.

The YOLO approach is one of the most promising deep learning algorithms for object detection, including applications for pothole recognition. YOLO is a neural network that uses object identification and classification methods to rec-

ognize things in real-time video feeds. It has achieved significant popularity owing to its superior accuracy and rapidity in object detection. Nonetheless, other methods have been investigated previously, although they all exhibit considerable shortcomings, including prolonged result generation and less reliable implementations. Deep learning networks have produced favorable results in all real-time applications and can assist in averting such incidents.

The Instance segmentation without Ultratics is depicted in Fig.5 and Fig.6 shows the object tracking with instance segmentation. These results After the training of your model has been completed, you will be able to evaluate the training outcomes by utilizing the graphs that were created by the YOLOv11. Fig.7 shows the mAP results of the YOLOv11 models loss results of various factors.

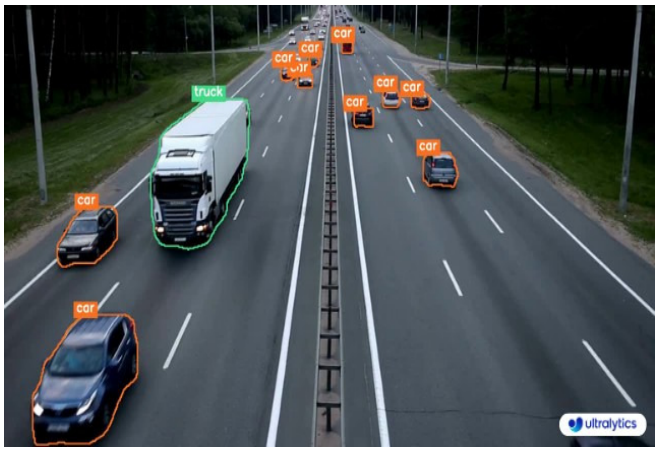


Fig.5. Instance Segmentation

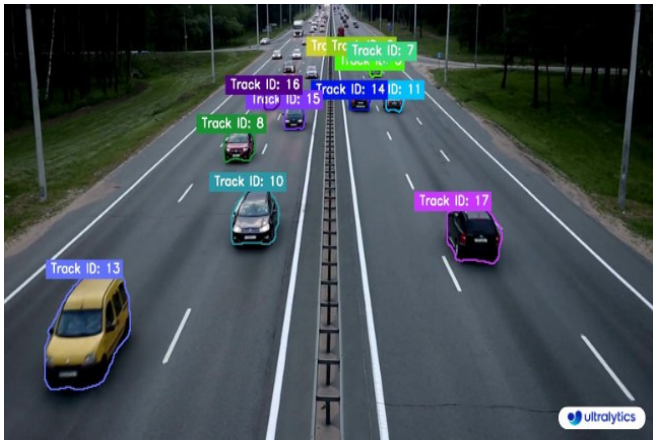


Fig.6. Instance Segmentation + Object Tracking

D. Discussions

The YOLOv11's advancements synthesize cutting-edge architectural improvements and practical application scalability. Introducing the C3k2 and C2PSA blocks ensures enhanced accuracy and computational efficiency, making YOLOv11 a versatile model for diverse industries.

1) Scalability Across Environments:

- The availability of multiple model variants, from nano to xlarge, makes YOLOv11 suitable for both edge devices and high-performance systems. However, optimizing these variants for specific hardware configurations remains a key area for future research.

2) Comparison with EfficientDet and Mask R-CNN:

- Unlike EfficientDet, which heavily relies on compound scaling for balancing accuracy and efficiency, YOLOv11 achieves similar or better mAP

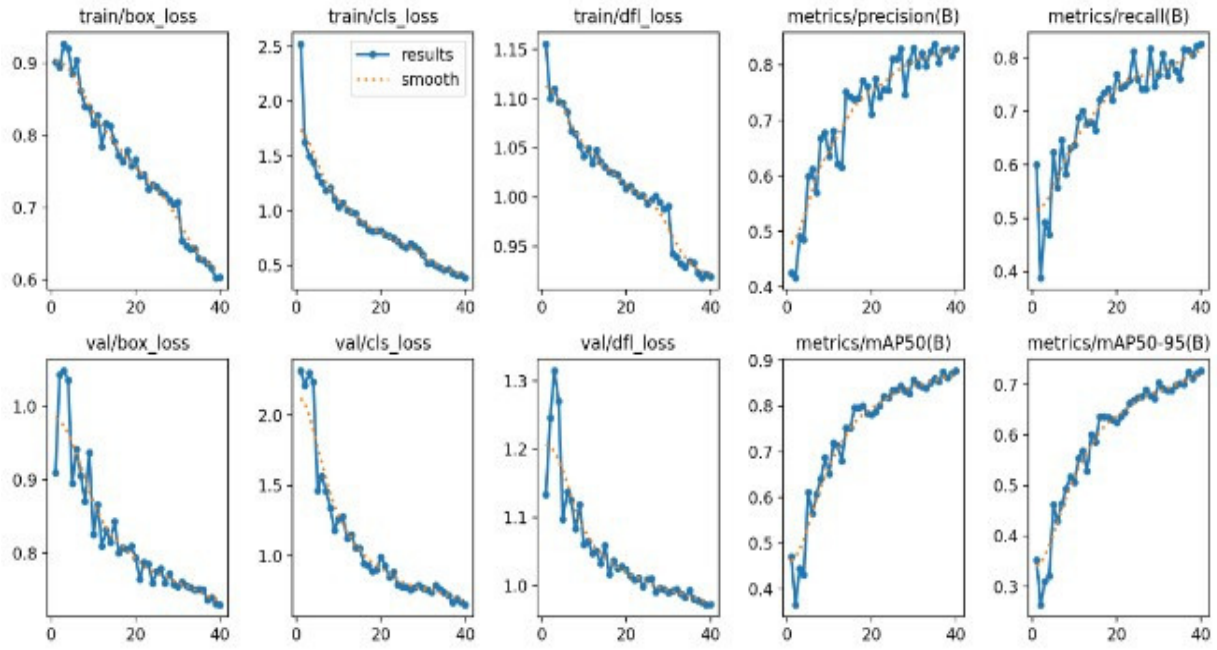


Fig.7. Results of the YOLOv11 model over various performance metrics

scores with lower computational costs due to its optimized architecture.
- Compared to Mask R-CNN, YOLOv11 offers faster inference times while maintaining competitive segmentation accuracy, making it more suitable for real-time applications.

3) Adaptability to Emerging CV Tasks:
- YOLOv11's support for instance segmentation, pose estimation, and oriented bounding box detection positions it as a comprehensive tool for emerging CV challenges. Its modular design facilitates customization for domain-specific applications.

4) Potential Challenges:
- While the model achieves state-of-the-art results, its reliance on advanced hardware for training may limit accessibility for smaller organizations. Efforts to streamline training pipelines and reduce dependency on GPUs could democratize access to YOLOv11's capabilities.

5) Future Directions:
- Enhancing model interpretability and incorporating self-supervised learning techniques could further elevate YOLOv11's utility. Additionally, expanding its compatibility with diverse datasets, including those with less structured annotations, could broaden its adoption.

E. Practical Applications

1) Autonomous Vehicles:
- YOLOv11's ability to process video streams in real-time enables accurate detection of pedestrians, vehicles, and traffic signs, ensuring safety and efficiency.

2) Medical Imaging:
- The model's high precision in segmenting organs and tumors is validated on custom datasets, demonstrating potential for diagnostic and surgical applications.

3) Retail Analytics:
- YOLOv11 tracks customer movements and accurately identifies products, improving inventory management and customer experience.

V. Conclusion and Future Work

In this paper, The AE-YOLOv11 is introduced with significant advancements, particularly through its architectural innovations (C3k2 and C2PSA blocks), enhancing accuracy and computational efficiency. The availability of various model variants (e.g., nano, xlarge) allows YOLOv11 to cater to diverse use cases, from edge devices to high-performance systems. This adaptability makes it versatile across industries. The AE-YOLOv11 outperforms competing models like EfficientDet, Mask R-CNN, SSD, and Faster R-CNN in terms of real-time detection capabilities, balancing speed and accuracy effectively. The model is suitable for diverse industries, including autonomous systems (e.g., vehicle detection), healthcare (e.g., tumor segmentation), and retail analytics (e.g., customer tracking). Focus areas include optimizing deployment on resource-constrained devices, enhancing model interpretability, incorporating self-supervised learning techniques, and broadening compatibility with less structured datasets. Future research should focus on optimizing deployment on resource-constrained devices, improving interpretability, and expanding its applicability across domains. The AE-YOLOv11's advancements pave the way for innovation in industries ranging from autonomous systems to healthcare, underscoring its position as a leader in computer vision technology.

References

[1] Redmon, J., et al. "You Only Look Once: Unified, Real-Time Object Detection." CVPR, 2016.

[2] Bochkovskiy, A., et al. "YOLOv4: Optimal Speed and Accuracy of Object Detection." arXiv, 2020.

[3] Sivadi Balakrishna and Vijender Kumar Solanki "RTPD-YOLO: Reconciling YoLo-V8 Model for Real-Time Potholes Detection", in International Conference on Machine learning and Applied Network Technologies (ICMLANT 2024) is organized by IEEE El Salvador Section, pp. 1-6, Dec 13-14, 2024.

[4] Sivadi Balakrishna **"**D-ACSM: a technique for dynamically assigning and adjusting cluster patterns for IoT data analysis", *The Journal of Supercomputing*, Springer, ISSN 1319-1578, Vol. 78, Issue 10, pp, 12873-12897, Mach 2022. DOI: https://doi.org/10.1007/s11227-022-04427-1

[5] Balakrishna, Sivadi, and Ahmad Abubakar Mustapha. "Progress in multi-object detection models: a comprehensive survey." *Multimedia Tools and Applications* 82, no. 15 (2023): 22405-22439.

[6] Balakrishna, Sivadi, Yerrakula Gopi, and Vijender Kumar Solanki. "Comparative analysis on deep neural network models for detection of cyberbullying on Social Media." *Ingeniería Solidaria* 18, no. 1 (2022): 1-33.

[7] Balakrishna, Sivadi, Moorthy Thirumaran, and Vijender Solanki. "Machine Learning based Improved Gaussian Mixture Model for IoT Real-Time: Data Analysis." *Ingeniería Solidaria* 16, no. 1 (2020): 1-30.

[8] Suvarna, Buradagunta, and Sivadi Balakrishna. "Enhanced content-based fashion recommendation system through deep ensemble classifier with transfer learning." *Fashion and Textiles* 11, no. 1 (2024): 24.

[9] Balakrishna, Sivadi, M. Thirumaran, R. Padmanaban, and Vijender Kumar Solanki. "An efficient incremental clustering based improved K-Medoids for IoT multivariate data cluster analysis." *Peer-to-Peer Networking and Applications* 13, no. 4 (2020): 1152-1175.

[10] Balakrishna, Sivadi, Vijender Kumar Solanki, and Rubén González Crespo. "Generative AI for Smart Data Analytics." In *Generative AI: Current Trends and Applications*, pp. 67-85. Singapore: Springer Nature Singapore, 2024.

[11] Muhammad Hussain. Yolo-v1 to yolo-v8, the rise of yolo and its complementary nature toward digital manufacturing and industrial defect detection. Machines, 11(7):677, 2023.

[12] Joseph Redmon and Ali Farhadi.Yolo9000: better, faster, stronger, "In Proceedings of the IEEE conference on computer vision and pattern recognition", pages 7263–7271, 2017.

[13] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

[14] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.

[15] Alexey Bochkovskiy, Chien-Yao Wang, and Hong-Yuan Mark Liao.Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934, 2020.

[16] Ultralytics. YOLOv5: A state-of-the-art real-time object detection system. https://docs.ultralytics.com, 2021.

[17] Chuyi Li, Lulu Li, Hongliang Jiang, Kaiheng Weng, Yifei Geng, Liang Li, Zaidan Ke, Qingyuan Li, Meng Cheng, Weiqiang Nie, et al. Yolov6: A single-stage object detection framework for industrial applications. arXiv preprint arXiv:2209.02976, 2022.

[18] Chien-Yao Wang, Alexey Bochkovskiy, and Hong-Yuan Mark Liao.Yolov7: Trainable bag-of-freebies sets new state-of-the-art for

real-time object detectors.In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 7464–7475, 2023.

[19] Mupparaju Sohan, Thotakura Sai Ram, Rami Reddy, and Ch Venkata.A review on yolov8 and its advancements.In International Conference on Data Intelligence and Cognitive Informatics, pages 529–545. Springer, 2024.

[20] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao.Yolov9: Learning what you want to learn using programmable gradient information.arXiv preprint arXiv:2402.13616, 2024.

[21] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding.Yolov10: Real-time end-to-end object detection. arXiv preprint arXiv:2405.14458, 2024.

[22] Glenn Jocher and Jing Qiu. Ultralytics yolo11, 2024.

rice2024 Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering pp. 97–102 DOI: 10.15439/2024R06 ISSN 2300-5963 ACSIS, Vol. 42

Dynamic Clock Tree Balancing Algorithm: Achieving Enhanced Performance Efficiency in Asic Design

J. Praveenkumar
Research Scholar, Bharath Institute of Higher Education and Research, Chennai, TamilNadu India
Email: praveencando91@gmail.com

G. Sudhagar
Associate Professor, Department of ECE, Bharath Institute of Higher Education and Research
Chennai, TamilNadu India
Email: sudhagarambur@gmail.com

***Abstract*—The creation and application of a novel clock tree balancing method that dynamically optimizes the clock distribution network in ASIC designs is the main focus of this project's effort. Improving the timing constraints, lowering clock skew, cutting down on power usage, and raising the ASIC's overall performance are the main goals. As technology has developed, ASIC designs have gotten more intricate, incorporating billions of transistors onto a single chip. It gets harder to create an efficient and well-balanced clock distribution network as the number of gates and design size increase. On the other hand, problems like clock skew, clock jitter, and excessive power consumption can arise when this clock signal is applied to every sequential part of a big and intricate ASIC design. For ASIC designs, it entails creating and constructing a dynamic clock tree balancing algorithm. Using real-time data, the program will improve clock distribution, improving timing constraints and lowering power usage. It will be evaluated against conventional techniques, verified on actual ASIC designs, and recorded for a dissertation. The project's goal is to improve ASIC design techniques to produce chips with high performance and low power consumption.**

***Index Terms*—Application-Specific Integrated Circuit, Dynamic Clock Tree Balancing, Skew-aware-source-pulling**

I. Introduction

AN INNOVATIVE and adaptive algorithm tailored for Application-Specific Integrated Circuit (ASIC) designs is implemented. The dynamic clock tree balancing algorithm will possess the capability to continuously optimize the clock distribution network in real-time.By leveraging up-to-the- moment data, the algorithm will actively work to reduce clock skew within the ASIC, consequently leading to a significant improvement in timing precision. The dynamic nature of the algorithm ensures that it can adapt to the evolving needs and demands of the system, making it a cutting-edge solution for enhancing the overall performance of ASIC designs. The clock signal is disseminated throughout the design in the form of a tree, with the clock source representing the root and the leaves representing the sequential devices that the clock signal triggers. A balanced clock tree is one in which all elements of the design get the clock signal nearly simultaneously. An essential phase in the implementation process is clock tree synthesis, which involves organizing and routing clock tree components. The implementation of specific cells and routing strategies guarantees a strong clock tree.

The clock remains in the optimal mode even when the standard cells and the macros are positioned in a fixed, optimized location. The clock enters the propagated mode because the clock signals carry out the data transfer between the various functional units on the chip. Every sequential element's clock input needs to be in sync for a design to meet setup and hold criteria. Due to the large number of sequential elements in the design, a single clock net cannot drive them all. The clock signals are distributed from a common point to each element's clock pin via the clock distribution network. The goal is to minimize arrival time uncertainty while balancing clock skew using dynamic clock tree balancing algorithm. Dynamic Clock Tree Balancing (DCTB) is a crucial aspect of Very-Large-Scale

Integration (VLSI) design, aimed at optimizing the distribution of clock signals across the chip to minimize clock skew and improve overall chip performance like power consumption and timing. The performance of the chip is directly impacted by the design of the clock networks, which is a crucial component of the design process. Clock signals synchronize operations fundamentally in contemporary integrated circuits. Timing errors and poor chip performance can result from clock skew, which is the fluctuation in clock signal arrival times at various locations on the semiconductor. To make sure that the updated clock tree satisfies timing requirements and decreases skew without introducing new problems, clock tree balancing with dynamic buffer delay adjustment necessitates meticulous verification and testing. A hierarchical network of buffers, wires, and clock distribution components known as the clock tree is used in VLSI design to distribute clock signals to various sequential units (such as flip-flops) throughout the chip.This enhances performance,cuts down on power usage, and decreases setup and hold time violations. A runtime or post-fabrication technique DCTB continuously examines and modifies the clock tree based on real time data.

On the basis of chip-level needs and critical paths, define skew thresholds and limitations which measure clock skew continuously from the main clock source to the leaf nodes, covering different locations in the clock tree hierarchy. The

DCTB algorithm is activated when skew reaches predetermined levels or violates setup/hold time restrictions. DCTB provides real-time adjustments to the clock tree in response to altering operating circumstances. Clock distribution and timing are susceptible to external influences, including temperature changes and voltage swings. DCTB supports sustaining peak performance in dynamic environments.One of the main objectives revolves around enhancing the timing constraints within ASIC designs.

The primary focus is on minimizing clock skew and jitter, which are notorious culprits for timing violations. Our goal is to meticulously refine these constraintsto ensure that setup and hold times for sequential elements are consistently met. By achieving this objective, we anticipate a notable reduction in the likelihood of timing violations occurring in the ASIC, thereby enhancing the reliability and predictability of the design. The other aspect is to embed power optimization techniques directly into the algorithm. The overarching objective is to create a clock distribution network that not only excels in timing precision but also excels in power efficiency. By doing so, we anticipate a significant reduction in the overall power consumption of ASIC designs. This aligns perfectly with contemporary principles of energy-efficient design, making our project not only technologically advanced but also environmentally conscious. The implementation of region based dynamic clock tree balancing algorithm is used in the proposed system which is one of the clock tree optimization techniques. The inter region synchronization technique is used to synchronize clock signals in every region. In each region Skew-aware-source-pulling (SASPO) is implemented to minimize the clock skew.

To provide a comprehensive perspective on the efficacy of our dynamic clock tree balancing algorithm, we will conduct a thorough evaluation and comparison against traditional static clock tree synthesis methods. This comparison will encompass various facets, including timing accuracy, power efficiency, and scalability. By directly contrasting the two approaches, we aim to showcase the superiority of our dynamic algorithm, underlining its potential to improve clock tree building in the fieldof ASIC design. The validation of our algorithm will be a pivotal part of this project. We plan to test its performance on real-world ASIC designs, ensuring that it thrives in practical applications. Furthermore, the insights gained from this project will not remain confined but will be shared with the broader ASIC design community. We intend to disseminate our findings. through research publications and present our discoveries at prominent conferences, thereby making substantial contributions to the field's body of knowledge and practice.

II. Literature Survey

A. Clock Tree Synthesis Techniques for Optimal Power and Timing Convergence in SoC Partitions

The studies were conducted in a SoC partition utilizing the physical design flow, which is implemented in 14nm technology. Using several optimization techniques at each design stage, it focuses mainly on timing, power, and area optimization. The analysis of effective CTS methods for timing convergence and optimal power in SoC partition is the main goal of this work. Multisource Clock Tree Synthesis and Multibit FlipFlop use are the approaches used for CTS with clock tree awareness. Through the reduction of latency and skew as well as the improvement of the clock distribution, multi-source CTS enhances timing of design. The number of sequential cells has decreased due to the use of multi-bit flip-flops, which has creased he overall power consumption of the clock network and design area.

B. An Efficient Clock Tree Synthesis Method in Physical Design

In this study, a low clock skew solution for the clock tree synthesis (CTS) design flow used in the mainstream industry is proposed. This article presents a method that greatly reduces the clock skew with area cost and placement time. Regarding the notable difference in throughput time, this is explained by the fact that this program runs on a single CPU core, but clock tree generation (CTG) can be executed concurrently on several smaller pseudo clock sources on multiple workstations. From this method there is a large time improvement, when the tool runs simultaneously on numerous workstations. Furthermore, this paper overcomes many flaws like excessive manual analysis because it is relevant to the mainstream industry's CTS design procedure.

C. A Clock Tree Synthesis Flow Tailored for Low Power

The purpose of this study is to provide real-world experience with poweroptimized clock tree construction, clock tree synthesis (CTS) target optimization, and quality of results (QoR) tracking. The studies that will be described were carried out on a mixed-mode architecture that was meant for a 55nm CMOS technology node. It features various power domains with more than 100K registers. Simulation findings show that the approach described here can save up to 20% of clock tree power. Using a standalone CTS tool in the flow can result in a clock tree power decrease of up to 20%. The design is put into the standalone tool that provides a dedicated CTS engine with potentially improved quality. This technique involves altering a standard P&R flow with integrated CTS capability. The tool's output is meant to be easily loaded into the standard P&R tool once CTS is finished.

D. Clock Tree Optimization Methodologies for Power and Latency Reduction

This paper will provide an overview of several commonly used clock structures, with a focus on the practical use of H-Tree and standard clock tree structure. The implementation was carried out on a real-time database using a 16nm technology node with 1.4 million instances and an operating frequency of 537MHz. The types of cells and routing that are utilized to create the H-Tree clock structure, the customization of the HTree clock structure based on sink distribution, and the numerous scenarios that need to be considered when

selecting this approach are also covered in this paper. Clockqor is compared between this method and the traditional clock tree structure, and the results indicate a positive improvement. The more conventional clock distribution networks might be replaced by these H-Tree clock networks. Without affecting the signal's properties, the suggested clock tree optimization techniques lower power dissipation.

There is a decrease in inductive noise due to the interconnects' inductive nature. In summary, the H-Tree structure will yield superior power, latency, and skew when the requirements are tight (80–100 ps for skew, <500 ps for latency, and more than 10,000 sinks).

E. An efficent clustering algorithm for low power clock tree synthesis

This paper provides a clustering approach for the local clock tree's power minimization, which is demonstrated to be equal to the tree's interconnect capacitance minimization. Clustering is used to find the clock buffers needed to synchronize a group of sequential and their positions. A cluster indicates that every sequential in the cluster is driven by a clock buffer. When capacity constraints are not used, the clustering algorithm guarantees the optimality of the solution by estimating the interconnect capacitance using the minimal spanning tree (MST) metric. Next, to reduce the limitations related to delay, slope, and skew, clock nets are routed, and buffers are sized. The paper compares the clock trees derived from our clustering method against the competitor alternatives across multiple blocks of a 65 nm microprocessor design. Method reliably increases the clock tree capacitance by up to 21%, as seen by the comparison. From the above referenced papers, the proposed method varies with the clustering method implemented dynamically as the design and technology varies, this will in turn reduce the number of iterations required to attain closure. Hence the proposed method could be time-saving and reliable way of dealing with the design with huge number of sequential elements being placed in appropriate location considering the clock root location, net length, skew limitations, and latency requirements.

III. Proposed Work

An ASIC is a sort of integrated circuit (IC) that is designed specifically for a given task or application as opposed to general-purpose ICs like a microprocessor. ASICs are created to carry out a specific function or group of functions with the highest levels of speed, efficiency, and power optimization. An ASIC's physical area, power usage, and performance are all impacted by its clock frequency. Higher clock frequencies often translate into both better performance and more power usage. DCTB aims to optimizes clock signal distribution across the chip to reduce clock skew and enhance timing and power consumption. One of the most important aspects of the design process is the clock network design, which directly affects the chip's performance. A clock tree with proper balance can increase semiconductor production yield. DCTB can assist in the production of more functioning chips and a decrease in the quantity of defective ones by minimizing timing violations and guaranteeing that chips fulfill their timing limitations.

To optimize power consumption, DCTB algorithms can dynamically modify the clock tree topology. It helps reduce power consumption, which is important for energy-efficient designs and battery-powered devices, by cutting down on needless power

dissipation in clock buffers and interconnects.

The proposed system implies Region-based dynamic clock tree balancing is a clock distribution optimization technique that divides a chip into different regions and balances the clock tree within each region independently. The system divides the chip into regions based on various criteria, such as functional blocks, IP cores, or logical groupings which contain a collection of sequential components, such as flip-flops in each region. Region-based balancing algorithms are scalable, which makes it appropriate for

intricate semiconductor designs. Region-based balancing assists in efficiently managing complicated clock distribution networks as chip sizes and complexity rise. Limited alterations and optimizations using region-based methods without affecting the entire chip is done. As a result, there is less chance of introducing further problems in other areasof the chip, which helps simplify the design process. Different regions of a chip may experience process variations differently. By individually modifying the clock tree inside each region, In Fig 1 region-based techniques can adjust to these differences while still meeting timing requirements.

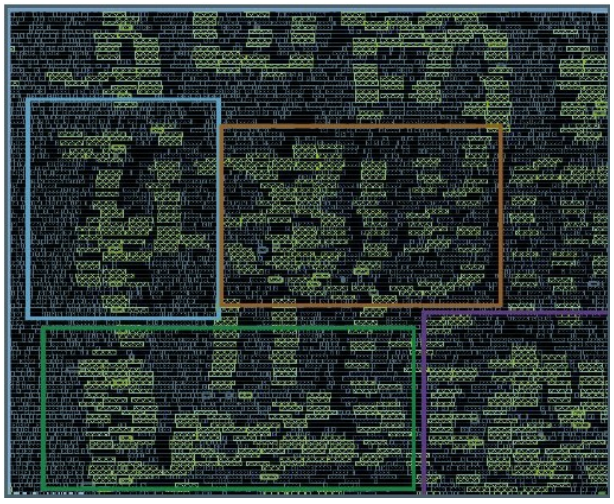


Figure 1: Region based Clock structure

In the Fig 2, the yellow highlighted cells are flops/sinks, and these are categorized based on the hierarchy, skew achieved and clock roots. Each domain will be driven by separate clock cell and the skew will be balanced within these sink flops and also make sure the timing is met within these interrelated flop paths.

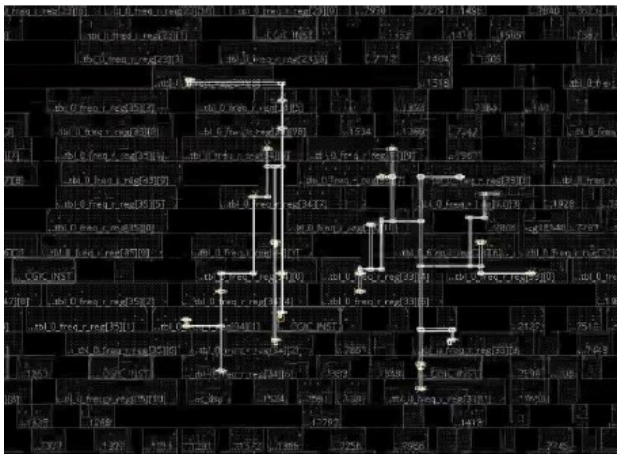


Figure 2: Clock structure based on Sink.

The clock gets distributed to all the flops in a region from the approximate mean point on the region to make sure the latency to reach each flop in a region is almost same which in turn keeps skew in control.

These components need synchronized clock signals. For each region, distinct clock trees are made for each zone to effectively disperse the clock signal within that region. A region's clock tree has its own root, main clock source, and special buffers. For continuously monitoring and analyzing the clock skew particular to each location, the system implements skew monitoring techniques there for both intra-regional and interregional skew can be monitored by these monitors.

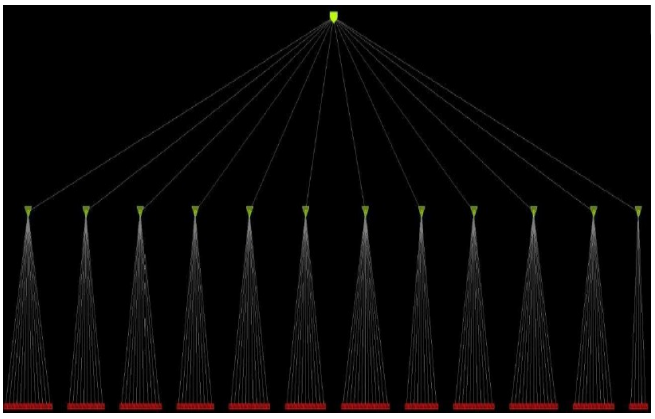


Figure 3: Clock Tree structure.

In Fig 3 Clock tree sinks grouped based on region and being driven by common clock buffer is depicted in the above image. The inter region synchronization technique is utilized to synchronize clock signals between various areas. Dynamic clock tree balancing (DCTB) uses the inter-region synchronization technique to ensure that clock signals migrate between various areas of a chip smoothly and with the least amount of clock skew. The clock signals are switched between various clock domains or regions at the edges of these regions, if not adequately controlled, these boundaries have the potential to be causes of clock skew. Inter-region clock skew may affect clock signals when they move between regions, which can cause setup and hold time violations or other timing problems. Clock Gating/Buffering at Region Boundaries is used as an inter-region synchronization technique to reduce inter-region clock skew. To avoid clock skew clock gating cells or buffers are placed at the region boundaries to synchronize the phases of the clock signal as it crosses over. To make sure that the inter-region synchronization mechanisms efficiently reduce skew and do not introduce additional timing infractions, they must be thoroughly verified and tested. By including suitable delays or buffers at the area borders, cross-region skew can be monitored and corrected appropriately. A hierarchical control system is designed that keeps an eye on how the clock tree balancing works in each region. This control system tracks the global skew, responds to requests from specific regions, and regional modifications should not compromise chip-level performance.

In the Fig 4, the inter region skew in maintained and the timing violations are met, by pulling and pushing the sinks from the average attainable latency.

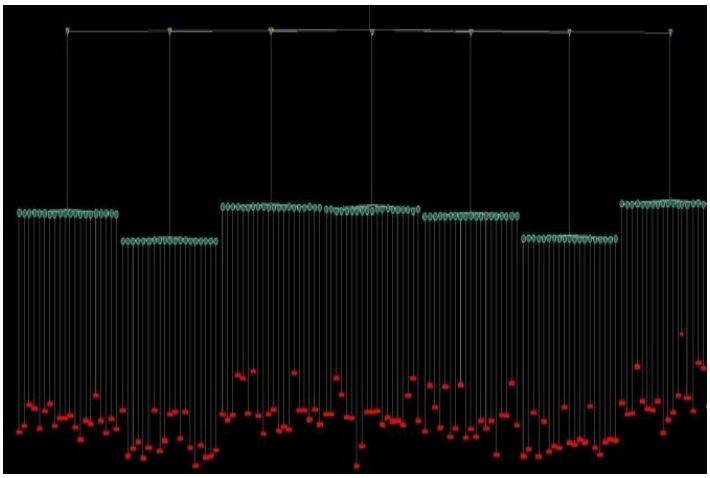


Figure 4: Clock Tree debugger.

Considering power limitations and dynamic balancing which affects power usage in each zone so energy-saving technique, such as power-down modes used for unused areas.To guarantee that the region-based dynamic clock tree balancing solution satisfies timing and performance specifications, thoroughly validate it using simulation and physical design tools.

While region-based balancing addresses local clock skew issues, incorporate chiplevel optimization techniques to ensure that global skew requirements are also met. This technique allows for customization of region boundaries and balancing parameters to accommodate different design requirements and goals.

The initial clock tree synthesis (CTS), which creates the clock tree structure using the logical netlist and chip layout, is the first step in the physical design process. In order to continuously measure clock skew across various regions of the semiconductor. Primary clock sources are often located at the top level of the clock tree, feeding down to lower levels of hierarchy. From the primary clock source to leaf

nodes, real-time monitoring circuits continuously assess clock skew at different points in the clock tree hierarchy. The design's skew limits and thresholds are set. The permissible skew limitations for different clock domains and pathways can be specified by these criteria. By comparing the measured clock skew against the established thresholds, Skewaware-source-pulling (SASPO) is one of the DCTB algorithms which identifies clock skew breaches, and the algorithm starts corrective steps when skew violations happen. The source-pulling method is used by SASPO to correct clock skew and this entails deliberately modifying the clock tree's delay components (buffers).

The primary goal of SASPO is to minimize clock skew within a digital integrated circuit. The clock skew describes the variance in clock signal arrival times at various locations on the chip. As a result of lowering clock skew, SASPO makes it possible for sequential components like flip-flops and latches to all receive clock signals at the same time, preventing setup and hold time infractions and enhancing chip performance.

To reduce skew, buffers can be added or changed to slow down or speed up the propagation of the clock signal along particular pathways. Based on the real-time feedback from the monitoring circuits, the algorithm dynamically adds or removes buffers. The main objective of source-pulling is to balance the timing of clock signals at various locations on the chip. To ensure that clock signals reach consecutive elements (like flip-flops) simultaneously, SASPO selectively modifies the latency of buffers. This decreases clock skew. To achieve exact skew management, SASPO may also fine-tune the delay parameters of individual buffers which is an iterative procedure. To reduce skew violations, it constantly analyzes and modifies the clock tree. While the main focus of SASPO is local optimization within the clock tree, it may also apply global optimization strategies to guarantee that chip-level skew criteria are met. SASPO implementations frequently take restrictions into account and work to reduce the extra power that buffer modifications needs. To foresee skew variations and make proactive adjustments, some SASPO solutions may include sophisticated technologies like predictive modeling and machine learning.

IV. Data Flow

In this section, the data flow and flow of the proposed methodology are explained. The database from the timing driven placement is given to the next step, which is the first step of the Dynamic Clock tree balancing. Firstly, the clock tree building is done in cluster mode, where the Clock transition is only fixed to determine the minimum insertion delay that can be achieved from the clock groups and skew groups which determines the maximum pulling clock sink. The next step is to analyze the clock tree structure and determine the root cause of this clock delay achieved. Hence the clock tree balancing is run in trial mode and the changes in the clock tree spec file are modified and provided for the CTS run in the next iteration. Once we see the violations, accordingly changes have to be made in the spec and again CTS building is carried forward, until we see the desired insertion delay and skew.

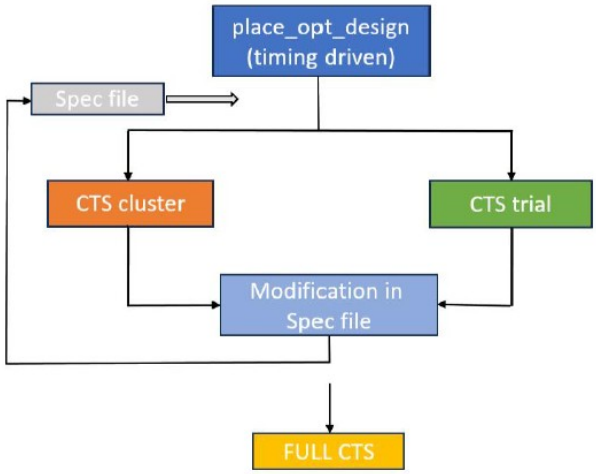


Figure 5: Clock tree building structure.

In the Fig 5 Data Flow, it is evident that the Input from Placement and the spec file generated from the SDC and user constraints is provided for the trial and cluster mode CTS to debug and find the requirements of the block which is iterated till we arrive at the desired output. In this iteration the clusters are organized based upon the skew groups, clock groups, hierarchies it belongs to and physical location of the cluster, which is done dynamically in the flow.

V. Results

A. Skew Comparison

In Table 1 The variation in time taken by each clock signal to reach the various sinks or clock pins on the clock tree branches is known as clock skew. The clock skew should be zero in ideal circumstances. The result obtained.

Table 1: Skew Comparison with clock trees

Clock Tree Type	Launch Latency (ps)	Capture Latency(ps)	Local skew (ps)
Conventional CTS	-300	340	640
Dynamic CTS	40	254	214

B. Power Comparison

In the Table 2 power comparison with clock trees Due to the existence of a clock mesh structure, dynamic CTS has a little higher clock network power than conventional CTS. It is also less when compared to a pure mesh network. When

Table 2: Power Comparison with clock trees

Clock Tree Type	Dynamic Power (mW)	Leakage Power (mW)	Total Power (mW)
Conventional CTS	34	29.12	63.94
Dynamic CTS	44.89	47.05	87.11

compared to the enhanced outcomes in other design features, the added power consumption is a minimal factor.

C. Timing Comparison

Dynamic timing analysis is carried out using a tool for timing analysis. To start the simulation runs and then perform the timing analysis of the design for ensuing reports of dynamic timing and analysis, the fundamental technique is automation, which is featured in PnR tools. In the Table 3 mentioned setup timing comparison.

TABLE 3: SETUP TIMING COMPARISON WITH CLOCK TREES

Clock Tree Type	WNS (ps)	TNS (ps)
Conventional CTS	-42	-117
Dynamic CTS	-17	-85

VI. CONCLUSION

After the successful implementation of the Dynamic Clock Tree Balancing Algorithm-Achieving Enhanced Performance Efficiency In Asics Design.

These results were in accordance with the detailed explanation of contents present in chapter.

VII. ACKNOWLEDGMENT

I wish to convey my special thanks to my project guide, Dr.G.Sudhagar, from Bharath Institute of Higher Education and Research for his consistent support, timely help, and valuable suggestions during the entire period of my Research.

I extend my sincere gratitude to all the teaching and non-teaching staff members of Bharath Institute of Higher Education and Research who helped me during the entire course.

REFERENCES

[1] (Clock Tree Synthesis Techniques for Optimal Power and Timing Convergence in SoC Partitions IEEEXplore:https://ieeexplore.ieee.org/document/9016727

[2] An efficient clock tree synthesis method in physical design | IEEE Conference Publication IEEEXplore:https://ieeexplore.ieee.org/document/5394159

[3] A Clock Tree Synthesis Flow Tailored for Low Power https://www.design-reuse.com/articles/33873/clock-tree-synthesis-flow-tailored-for-low-power.html

[4] Clock Tree Optimization Methodologies for Power and Latency Reduction SemiconductorDigest : https://www.semiconductor-digest.-com/clock-tree-optimizationmethodologies-for-power-and-latency-reduction/

[5] An efficent clustering algorithm for low power clock tree synthesis: https://www.researchgate.net/publication/220915569_An_efficent_-clustering_algorithm_for_low_power_clock_tree_synthesis

[6] Finding placement-relevant clusters with fast modularity-based clustering https://www.researchgate.net/publication/330493649_Finding_-placementrelevant_clusters_with_fast_modularity-based_clustering

rice2024 Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering pp. 103–116
DOI: 10.15439/2024R117
ISSN 2300-5963 ACSIS, Vol. 42

Machine Learning-Based Prediction Models for Sentiment Analysis on Online Customer Reviews: A Case Study on Airbnb

Cu Kim Long
Information Technology Center
Ministry of Science and Technology,
Lab AI 4.0, AIRC, VNU-ITI
Hanoi, Vietnam
longck.2006@gmail.com

Le Bao Ngoc
Norwich Business School
University of East Anglia
Norwich, United Kingdom
witty.gem2211@gmail.com

Vijender Kumar Solanki
Stanley College Of Engineering & Technology for Women
Hyderabad, TG, India
vijendersolanki@ieee.org

Nguyen Viet Anh
Dental School, Hanoi University of Business and Technology
Hanoi, Vietnam
vietanh.bsrhm@gmail.com

Luu Hoang Bach
Vinmec Research Institute
Hanoi, Vietnam
luuhoangbach711@gmail.com

Cu Ngoc Son
Faculty of Information Technology
Hanoi National University of Education
Hanoi, Vietnam
stu745105077@hnue.edu.vn

***Abstract*—In the last decade, with the rise of sharing economy, in particular Airbnb, customers are not merely buyers but also actively share their thoughts and experiences toward goods and services. Sentiment analysis, a sophisticated technological approach, has emerged as a pivotal tool to extract people's opinions as well as sentiments from written language. On the other hand, assessing the price of a listing has always been a daunting task for hosts and guests. While numerous pricing models for Airbnb have been proposed, achieving precise accuracy remains a challenge. As a result, this paper aims to investigate whether incorporating the sentiment scores derived from online customer reviews could improve the accuracy of Airbnb price prediction or not. First, online customer reviews on Airbnb are examined using natural language processing techniques to seek the guest sentiment and its association with listings prices. Once sentiment scores are calculated, they are used as an additional attribute to forecast Airbnb listings price. Several machine learning models are employed, including Linear Regression, Ridge Regression, Support Vector Machine, XGBoost and Random Forest. The experimental results show that the inclusion of sentiment scores slightly decreases model performance in the case of three Asian economies (Hong Kong, Japan and Taiwan). Overall, Random Forest without sentiment variable is the best-performing model among five models for Airbnb price prediction.**

***Index Terms*—machine learning, sentiment analysis, OCRs, price prediction, Airbnb.**

I. Introduction

AIRBNB, established in 2007, is the top pioneer in peer-to-peer (P2P) sharing platforms in the hospitality industry [1]. Airbnb has created an online solution to directly link guests looking for short-term accommodations with hosts who are in demand to lease out their homes [2]. In other words, the company works as a broker facilitating the connection between property owners and hospitality seekers. Since its foundation, Airbnb has consistently experienced a year-over-year supply growth rate of over 100% for the last decade [3], serving over 220 countries and regions worldwide [4].

Despite its unprecedented yet exponential growth, pricing has always been a major concern of Airbnb's stakeholders [5]. Determining an appropriate price for a rental property on Airbnb platform has been a challenging task for not only the tenants but also the owners. While guests need to evaluate the reasonable price of the listings to avoid being deceived, hosts also need a competitive price for their rental house to attract customers [6]. Therefore, predicting price is one of the most critical components in accommodation sharing systems such as Airbnb so both tenants and house-owners gain maximum benefits from the platform.

On the other hand, in a P2P platform, online customer reviews (OCRs) are considered to have a significant impact on not only sales but also the price of the property [7]. In the era of Web 2.0, a significant volume of data in which OCRs presents a significant challenge for any business and institution to deal with effectively. Even though peer-reviewed research has been undertaken to investigate customer reviews using text mining, there has been a lack of empirical research to enrich the forecasting model by applying textual methods. This study aims to focus on the topic of sentiment analysis on customer reviews to develop Airbnb price prediction model.

Following the discussion above, the main research questions of this study are how the association between sentiment scores derived from OCRs and Airbnb rental prices is, and how the performance of Airbnb rental price prediction models change when incorporating sentiment scores derived

from OCRs. These questions are formulated not only to fill in the gaps represented in the literature in the next chapter but also to contribute to Airbnb and the hospitality industry. **To answer research questions**, the objectives of the study are threefold: *(1) To construct and explore the guest sentiment in OCRs in three Asian economies namely Hong Kong, Japan and Taiwan; (2) To identify the association between Airbnb rental price and sentiment scores expressed in OCRs; (3) To identify the performance of Airbnb rental price prediction model with the inclusion of sentiment index from online customer reviews*.

Based on the aforementioned research questions and objectives, the following **two hypotheses** are formulated as below:

H_1: Positive sentiment expressed in OCRs is associated with an increase in Airbnb rental prices.

H_2: The inclusion of sentiment scores derived from OCRs as an explanatory variable statistically improves the predictive accuracy of Airbnb rental price prediction models.

Although Airbnb has introduced its price suggestion tools since 2012, which have been developed to "smart pricing" until now, the price prediction model still needs further improvement. The significance of this paper is underscored by the incorporation of sentiment scores derived from OCRs into forecasting Airbnb rental prices. By exploring the sentiment analysis on OCRs, this study seeks to enhance the knowledge surrounding the use of textual data for predictive modelling. The outcomes are not only for Airbnb but also for the broader context of the hospitality industry. In addition, the growth of Airbnb in Asia emphasises the need for research in Asian regions. Most of the existing papers have primarily centred on Western countries, and yet little attention is being paid to Asian regions, which is a growing and potential market for Airbnb in recent years [8]. Therefore, the focus on Asian economies of this study would enhance the scope and relevance of research in the field of Airbnb as well as the hospitality industry.

The rest of this paper is organized as follows. The relevant academic research within the field of sentiment analysis and price prediction models in Airbnb is presented in Section II. Section III discusses machine learning models used for predicting price, as well as the evaluation metrics employed to assess their performance. The building of five models including Linear regression, Ridge regression, SVM, XGBoost and Random Forest is discussed and evaluated to test the stated hypotheses in Section IV. Conclusions and future works are given in Section V.

II. Literature Review

In this section, relevant academic research within the field of sentiment analysis and price prediction models in Airbnb is presented. Firstly, it delves into the key concept of online customer reviews. Then, the literature on sentiment analysis and existing models for Airbnb price prediction is discussed. This chapter not only highlights the existing knowledge but also identifies the gaps that this study attempts to address.

A. Online customer reviews (OCRs)

Prior to the advent of online opinion-sharing platforms, the primary mode of communication among consumers was word-of-mouth. Consumer word-of-mouth has been frequently cited as one of the most crucial elements to determine the long-term success of goods and services. However, since the rise of online communities as well as communication facilitated by the Internet, there has been a new product information channel with growing popularity, where consumers share their experiences toward products and services, also known as online customer reviews (OCRs).

Online customer reviews (OCRs) refer to the evaluation of a product or service shared by customers on company or third-party websites. The importance of OCRs on consumer purchase decisions in the hospitality sector has been widely studied in the economic literature [9-10]. Besides, online hotel reviews is a reliable information source for customers as they unveil guests' feelings, attitudes and evaluations, thereby, reflecting guest satisfaction or dissatisfaction [11].

According to [12], OCRs influence the decision-making of guests in all ages, thus, contributing to the sales revenue. This is well-illustrated empirical studies which indicated that online reviews impact early sales, as a result, can be a significant predictor of box office revenue. Similarly, OCRs and the number of reviews can be used to determine future digital camera sales by fitting a multiple linear regression. This is primarily because individuals tend to readily embrace and place trust in information shared by other peers similar to themselves. OCRs help to alleviate the perceived risk and confusion of consumers [13]. In the tourism and hospitality industry, prior research has empirically proved that OCRs have a significant influence on purchasing decisions, especially booking intentions [14]. By conducting an Analysis of Variance (ANOVA) test, online reviews affect the decision making of consumers within the hospitality industry. Employing data from a Chinese online travel agency found that an increase of 10% in traveller review ratings leads to a considerable increase of over 5% in online bookings. Positive reviews are the motivation that inspires people to travel, meanwhile, negative reviews act as an effective tool to help people avoid bad travel products [15].

In the context of Airbnb, peer-to-peer feedback or so-called OCRs is even more significant than that of traditional hotels. This is because Airbnb hosts are usually micro-entrepreneurs who are financially unable to advertise their accommodations on media such as television like hotels, thus, the online platforms serve as the exclusive mean for them to connect with their guests. Furthermore, by using textual data, hosts can delve into a more comprehensive insight on customers' experiences rather than solely depending on non-textual data such as rating scores given by guests [16]. As a result, it becomes even more compelling to investigate the sentiment of guests based on OCRs within the Airbnb ecosystem.

B. Sentiment Analysis

In the case of OCRs, sentiment analysis can serve as a methodological approach for classifying, measuring and monitoring users' emotional responses towards a product or service [17]. Realising the importance and explosive growth of OCRs on the Internet, there has been an emerging stream of research undertaken to identify the sentiment index in online textual reviews, especially in the field of hospitality. They aimed to explore the relationship between the sentiment of reviews and the listing prices, thereby understanding the role of OCRs in consumer valuation and pricing decisions. They have reinforced the sentiment analysis to aspect-based sentiment analysis, which extracts the sentiment polarities towards specific aspects of an entity within hotel reviews. As a result, this research has significantly improved the comprehensiveness and accuracy of sentiment analysis in the hospitality industry [18].

On the other hand, supervised machine learning is introduced in [19-20] for sentiment analysis as a different approach. Specifically, Naïve Bayes classification is applied to measure not only the polarity but also the subjectivity scores of user-generated contents on TripAdvisor [21]. While polarity evaluates the emotion of text, the subjectivity scores measure the subjective or objective score of text. Alternatively, the Long Short-Term Memory model is introduced in [20], which is one of the latest deep learning technologies. This has significantly improved text classification performance. The sentiment indexes are separated into past and future housing price changes. According to the paper, this model could capture the word order and dependence, which unsupervised machine learning is unable to do.

Sentiment analysis, in which sentiment polarity classification is broadly used in forecasting, including product sales forecasting [22], stock market forecasting [23], household expenditure forecasting [24]. Nevertheless, the application of sentiment analysis in hospitality forecasting literature remains uncommon [25]. For this reason, this study examines the association between OCRs and Airbnb listings' prices as well as use guest sentiment extracted from OCRs to predict Airbnb listings' prices.

C. Price Prediction on Airbnb

Pricing a listing is considered one of the most crucial business practices for any Airbnb host to master [26]. Conventionally, hosts are allowed to set their own nightly, weekly and monthly prices for their rental houses. However, Airbnb still provides suggestions to assist their hosts to set more optimal prices for the entire selling period, which is a dynamic pricing strategy called "Smart Pricing" [27]. The Smart Pricing algorithm takes into consideration various points of information, including the date of the night to price, market demand, seasonality, listings' characteristics. Once receiving a pricing tip, a host can either choose to increase, decrease or do nothing.

However, several scholars found that, in contrast to professional hosts, nonprofessional hosts appear to adopt different and less dynamic pricing strategies [28]. Therefore, identifying the determinants of price on Airbnb has received a significant concentration in recent years. Moreover, factors related to the property are also found to significantly influence listing prices such as the site and location of the property [29], cleanliness of the rooms [7], type of accommodation [30]. Besides the factors on the supply side, studies also showed that the price of Airbnb accommodation tends to decrease as the number of reviews it receives increases [26].

In parallel to factors determining price, choosing an efficient prediction model is also essential so as to give the best accuracy. In 2017, Wang and Nicolau [29] identified 25 price determinants using ordinary least squares and quantile regression. Afterwards, price prediction of Airbnb has witnessed advancements beyond traditional linear regression model. To be specific, Liu [31] conducted a study on various models by leveraging machine learning techniques to capture non-linear relationships between price and other factors. Both of their papers discovered that XGBoost yields the highest accuracy, with R^2 equals 61.8% and 63% respectively. However, Mahyoub [5] concluded that Random Forest Regressor is the most effective model with R^2 equal to 86.95%, which outperforms XGBoost regression.

The difficulties in determining the prices for Airbnb accommodations can be attributed to the inherent complexity of these properties. This is because they encompass not only a variety of functional attributes but also the social interactions between hosts and customers [27]. While studies have indicated that prices are influenced by a set of factors including host attributes and property characteristics as aforementioned, online customer reviews are largely underexplored.

Meanwhile, Ganu [32] posits that customer reviews have the capability to demonstrate reviewers' attitudes more precisely than numerical star ratings, which may be biased. In other words, unidimensional customer ratings can be significantly biased by price effects. Therefore, Lawani [7] suggested that rating scores can result in biased implications on the relationship between the scores and prices since rating scores might not accurately capture guests' opinions and sentiments regarding a good or service. All in all, this research will take into consideration the studies discussed in the literature presented to develop a price prediction model with sentiment analysis.

III. Research Methodology

This section outlines the research approaches step by-step, including data sources, data pre-processing, sentiment analysis and modelling. The technique used for sentiment analysis, which is a lexicon-based method, is introduced. After extracting guest sentiment, the chapter discusses machine learning models used for predicting price, as well as the evaluation metrics employed to assess their performance.

A. Data description

The data is retrieved from Inside Airbnb (insideairbnb.-com), which is an independent and non-commercial website that provides data collected from Airbnb for public use. Inside Airbnb encompasses not only structured data related to Airbnb listings but also unstructured data in the form of customer reviews for each listing. The data from Inside Airbnb has been largely utilised in academic research and significantly contributed to the debates around the existence as well as growth of Airbnb [33].

The dataset consists of 569,523 Airbnb listings which are managed by 14,584 hosts in three Eastern Asia economies, namely Japan, Hong Kong and Taiwan on 31 March 2023. Asian economies are chosen in this study for two reasons. Firstly, scholarly research has been geographically focused on the United States, Canada and Europe [34]. Meanwhile, only 13.4% of the studies collected their data in Asia regions. Secondly, a growing popularity of Airbnb is shifting from Europe to America, and mostly to Asia [8]. As a result, investigating the dynamics of Airbnb in Asian countries can contribute to filling the gap in the literature.

The entire workflow including data handling, pre-processing, modelling, analysis, and visualisation in this study will be executed via the R programming language.

B. Data Pre-Processing

Data collected in their raw format can be an issue for sentiment analysis and modelling as they might be formatted inconveniently such as stop words, missing data and so on. Therefore, in order to facilitate further analysis, it is essential to undertake several data pre-processing steps.

Firstly, non-English texts are detected using Google's Compact Language Detector 2 (cld2) package in R [35], which can detect 80 languages in UTF-8 text and even mixed language input. The purpose of this action is to separate English with Chinese and Japanese reviews. Secondly, reviews in Chinese and Japanese language are translated into English language using translate package in R, which translates between different languages with Google API [36]. The reason to translate non-English reviews rather than maintaining their original versions is to achieve unification and consistency in the sentiment index calculation. Since this report uses lexicon-based approaches for the sentiment score, adopting multiple dictionaries for each language would introduce variations in the scale of the sentiment score.

Finally, pre-processing steps are performed, following the processes recommended in prior research [17]:

(1) Lowercase: Every character in each review is converted into lowercase. R is a case sensitive program language, and 'Visit' is different from 'visit' due to character coding; therefore, it is essential to convert textual data into lowercase. For example, taking one review in the dataset, we have:

The original sentence: *"The apartment is in a very convenient location. Host was extremely helpful and the apartment was great. Located in a very calm and nice area, extremely convenient. Would come again for sure!"*.

All cases are transformed into lowercase: *"the apartment is in a very convenient location. host was extremely helpful and the apartment was great. located in a very calm and nice area, extremely convenient. would come again for sure!"*.

(2) Tokenisation: Each review is split into tokens in the form of single words or terms. At the same time, white spaces and punctuation are removed. This is an important step to remove stop words or unnecessary characters: *"the apartment is in a very convenient location host was extremely helpful and the apartment was great located in a very calm and nice area extremely convenient would come again for sure"*.

(3) Stop words removal: This phrase removes stop words, which may be articles, pronouns, prepositions, etc. These words frequently occur but do not add meaning to a sentence, meaning that they do not express any sentiment when applied to lexicon resources. Thus, removing stop words would reduce the noise before text processing. In English, stop words could be 'an', 'the' or 'is'. To remove, we use a stop-words list that is already available on R: *"apartment convenient location host extremely helpful apartment great located calm nice area extremely convenient come again sure"*.

(4) Stemming: Stemming is the technique that involves removing word suffixes to extract the root form of words. This is commonly used in text mining because it simplifies the textual data without causing significant loss of information. For example, "extremely" in the sentence is converted to "extreme": *"apartment convenient location host extreme helpful apartment great locate calm nice area extreme convenient come again sure"*.

To understand the effects of text pre-processing on the comments, key statistics about the length of words in comments are demonstrated in table below.

TABLE I. DESCRIPTIVE STATISTICS ABOUT THE OCRS AFTER AND BEFORE TEXT PRE-PROCESSING

	Before text pre-processing	**After text pre-processing**
Average number of words	40	22
Median number of words	26	15
Shortest comment	1 word	1 word
Number of words 1st quantile	12	7
Number of words 3rd quantile	52	29
Longest comment	2905 words	651 words

As we can see from the Table I, unnecessary characters and stop words are removed to reduce the noise of the dataset since they do not contain information and express sentiment. Quantitatively, the average number of words in

comments has been reduced by nearly a half while the longest comment dropped by about 77.6%. The use of clean data facilitates faster training and thus, enables the implementation of multiple experiments even though limited computational resources are available.

C. Sentiment Extraction

After pre-processing procedures, the sentiment score is constructed from online customer reviews. As discussed in the previous section, while sentiment scores can be extracted in different ways, it is essential to acknowledge that each method has its strengths and limitations. Based on the objectives of this paper, lexicon-based approach is chosen to extract sentiment from texts. There is a wide range of dictionaries that were developed for lexicon-based method, in which each of them offering unique features and attributes. To compare these dictionaries, Al-Shabi [37] has evaluated the performance of the five most well-known lexicons used in sentiment analysis. The results show that VADER (Valence Aware Lexicon and Sentiment Reasoner) demonstrates the highest accuracy in both positive and negative classification.

For this study, VADER is applied to calculate the sentiment score of reviews. VADER is a rule-based sentiment analysis tool to detect sentiment in social media texts. When comparing the classification accuracy, it was found that VADER outperforms individual human raters with Classification Accuracy scores equal 0.96 and 0.84 respectively as it considers both polarity and intensity of emotion. With VADER, the sentiment score, or compound score, is calculated by adding up the valence scores of individual words in the lexicon, which is subsequently normalised to range from –1 (complete negative) to +1 (complete positive). Sentiment classification is then based on this compound score as follows:

- Positive sentiment: compound score $\geq +0.05$
- Negative sentiment: compound score ≤ -0.05
- Neutral sentiment: $-0.05 <$ compound score < 0.05

After determining sentiment score for each review, the overall sentiment score for each listing is computed by taking the mean of sentiment scores associated with that listing. These final sentiment scores are then set as a new feature which is used in price prediction models.

D. Modelling and Analysis

As a reminder, there are two hypotheses that need addressing in this research. The first one is to test the positive association between sentiment scores and prices of Airbnb listings. The second one is to test whether the performance of price prediction models with sentiment score variable are

TABLE II. VARIABLE DEFINITIONS AND LABELS

Variable name	Description
Property characteristics	
Log price	Log-transformed daily price of listing (in USD Dollar)
Country	Country from which the listing is located
Room type	Type of room: (1) Entire home/apt; (2) Hotel room; (3) Shared room; (4) Private room
Accommodate	The maximum capacity of the property
Beds	The number of bed(s)
Minimum nights	The minimum number of nights stay in the listings
Quality characteristics	
Sentiment score	Sentiment score extracted from sentiment analysis on OCRs
Number of reviews	The number of reviews a listing has
Review score rating	The overall review score rating a listing has
Host characteristics	
Host response rate	The rate at which a host response to the guest
Host response time	Time a host response to the guest: (1) Within an hour; (2) Within a few hours; (3) Within a day; (4) A few days or more; (5) Unknown
Host identity verified	The identity of host is verified or not (True/False)
Host acceptance rate	The rate at which a host approves booking requests
Host total listings count	The number of listings one host own on Airbnb

TABLE III. SUMMARY STATISTICS OF NUMERICAL VARIABLES

Variable	Obs	Mean	Median	Min	Max
Log price	13,823	4.653	4.662	-2.322	11.068
Country	13,823	0.672	0.690	-0.960	1
Room type	13,823	3.936	3	1	16
Beds	13,823	2.422	2	1	42
Minimum nights	13,823	6.513	2	1	1125
Sentiment score	13,823	0.672	0.690	-0.960	1
Number of reviews	13,823	40.86	18	1	1548
Review score rating	13,823	4.624	4.750	0	5
Host response rate	13,823	0.971	1	0	1
Host acceptance rate	13,823	0.918	0.990	0	1
Host total listings count	13,823	26	12	1	748

TABLE IV. SUMMARY STATISTICS OF CATEGORICAL VARIABLES

Variable	Categories	Freq.	%
Host response time	0 (within an hour)	1,034	7.48
	1 (within a few hours)	10,678	77.24
	2 (within a day)	1,402	10.14
	3 (a few days or more)	521	3.77
	4 (unknown)	188	1.36
Host identity verified	1 (True)	12,511	90.51
	0 (False)	1,312	9.49
Room type	0 (Entire home/apt)	9,246	66.89
	1 (Hotel room)	444	3.21
	2 (Private room)	3,638	26.32
	3 (Shared room)	495	3.58
Country	0 (Hong Kong)	2,149	15.55
	1 (Japan)	8,895	64.35
	2 (Taiwan)	2,779	20.10

improved or not. In order to examine the second hypothesis, two versions of the dataset are created - one with the sentiment feature and one without the sentiment feature.

The explanatory variables chosen in this study are categorised into four main attributes:

(1) Host characteristics including host response time, host identity verified, host acceptance rate, host response rate, host total listings count;

(2) Property characteristics including country, room type, accommodates, beds, minimum nights stay;

(3) Rating-related features including number of reviews, review score ratings;

(4) Sentiment scores derived from reviews.

The descriptions and summary statistics of both predictor and explanatory variables are shown in Table II, III, IV and V.

Linear Regression [38] is employed first as a baseline model to evaluate the performance of other models. After the baseline is established, several machine learning models namely Ridge Regression [39-40], Support Vector Machine [41], XGBoost [42] and Random Forest [43] are performed.

TABLE V. MULTIPLE LINEAR REGRESSION RESULTS FOR PREDICTING LOG PRICE

Coefficient	Estimate	Std. Error	T-statistic	p-value
(Intercept)	4.514	0.139	32.436	0.000 (***)
host_response_time_1	0.116	0.026	4.452	0.000 (***)
host_response_time_2	-0.058	0.318	-1.808	0.07
host_response_time_3	-0.075	0.043	-1.743	0.08
host_response_time_4	-0.436	0.123	-3.552	0.000 (***)
host_identity_verfied_1	-0.025	0.023	-1.101	0.271
room_type_1	-0.338	0.038	-8.930	0.000 (***)
room_type_2	-0.323	0.016	-20.548	0.000 (***)
room_type_3	-1.153	0.036	-31.704	0.000 (***)
country_1	0.268	0.020	13.150	0.000 (***)
country_2	-0.295	0.023	-13.093	0.000 (***)
accommodates	0.121	0.003	36.052	0.000 (***)
beds	0.001	0.005	0.149	0.882
host_acceptance_rate	0.091	0.045	2.045	0.041 (*)
review_scores_rating	0.061	0.016	3.935	0.000 (***)
host_response_rate	-0.772	0.126	-6.110	0.000 (***)
sentiment_score	0.090	0.043	2.113	0.034 (*)
number_of_reviews	-0.001	0.000	-5.751	0.000 (***)
minimum_nights	-0.000	0.000	-0.959	0.338
host_total_listings_count	-0.001	0.000	-5.170	0.000 (***)

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

To test the hypothesis, five machine learning models were built to predict the log price of Airbnb properties in two cases: with and without the sentiment score variable. Linear Regression was employed first as the baseline model for model comparison in this study. Subsequently, Ridge Regression, Support Vector Machine, XGBoost, Random Forest were built. To compare the predictability performance of the models, R^2, RMSE and MAE were calculated.

E. Model Performance Evaluation

To evaluate the performance of predictive models, the dataset is partitioned into training and test sets based on the dependent variable – the price, where 75% of the dataset is for training and 25% for testing. In our case, the training set contains 10,369 records, while there are 3,454 records in the test set. First, the models are fitted into the training set to learn patterns and relationships in the data. After the models are built, they are evaluated on the test set, which contains of unseen data points.

Three metrics, R^2, RMSE (Root Mean Square Error) and MAE (Mean Absolute Error) are used to evaluate and compare the performance of predictive models. R^2 or the coefficient of determination, which is a standard metric for evaluating regression analyses, measures how close the target variable is determined by explanatory variables, interpreted by the proportion of total variance of the regressand explained by the model. While RMSE is the standard deviation of mean prediction errors, MAE measures the average magnitude of the prediction errors.

The formula of R^2 and RMSE are as below with y_i is the actual value and $\hat{y}_i$ is the predicted value of the dependent variable.

$$R^2 = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \overline{y}_{ii})^2} \quad (1)$$

$$RMSE = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n}} \quad (2)$$

$$MAE = \frac{1}{n} \sum |y_i - \hat{y}_i| \quad (3)$$

A model can be considered to be effective if its R^2 is close to 1; RMSE and MAE are close to 0.

IV. Results and Analysis

A. Customer reviews

The reviews were gathered from Inside Airbnb with a total of 569,523 reviews for the analysis. However, following the data pre-processing process, there are 507,974 reviews left. There are 13,789 listings, with an average of 37 reviews per listing. After detecting the language of each review in the dataset, it can be seen from Figure 1 that more than half of the reviews are written in English, with 254,490 reviews. This prevalence can be attributed to the global customer base of Airbnb, attracting users from worldwide, where English is commonly used for international communication. Following English are Japanese and Chinese language, with 23.6% and 24.3% of the reviews respectively.

After all the text has been translated to English, to provide more detailed information about reviews, two-word clouds are drawn based on a collection of OCRs from Airbnb listings in Hong Kong, Japan and Taiwan. These graphs help to display the keywords hidden in reviews. The larger words in the word cloud indicate higher frequency in the reviews, while the smaller words reveal less occurrence. In Figure 2 and Figure 3, 80 of the most frequent words are selected.

Using VADER sentiment analysis, the sentiment polarity and sentiment intensity of reviews are obtained, in which polarity assigns whether reviews are negative, positive or neutral, while intensity indicates the strength of the negative or positive sentiment in the text. Figure 4 shows that almost all of the reviews were positive, accounting for approximately 94% of the reviews, followed by neutral and negative reviews with only 3.8% and 2.2% respectively. Positive reviews are the most dominant sentiment category, reflecting guests' overall positive experiences with Airbnb in 3 Asia countries.

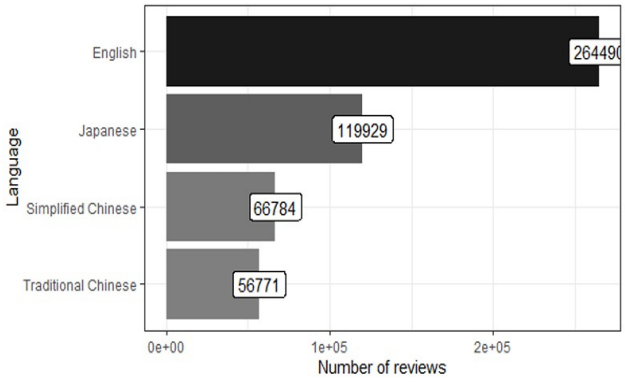


Fig. 1. Frequency of the languages in reviews written by guests on Airbnb in Hong Kong, Japan and Taiwan.

To get a sense of sentiment scores in each individual country, an ANOVA (Analysis of Variance) test was conducted to analyse the differences in sentiment scores among

Fig. 2. World cloud of OCRs on Airbnb in Hong Kong, Japan and Taiwan.

different countries. The results are illustrated in Table VI, with F-statistic equals 91.03 and the p-value is less than 0.0001. The p-value is below conventional significance level of 0.05, which indicates that there is a statistically significant difference in the means of sentiment scores among the three countries. As a result, this underscores the importance of considering country-specific factor when assessing sentiment scores.

The sentiment score for each listing was calculated by taking the mean of sentiment intensity across each review associated with that specific listing. As shown in Table VII, the average sentiment score is 0.707. However, there is some variability in the sentiment score, ranging from –0.998 (most negative) to 1 (most positive).

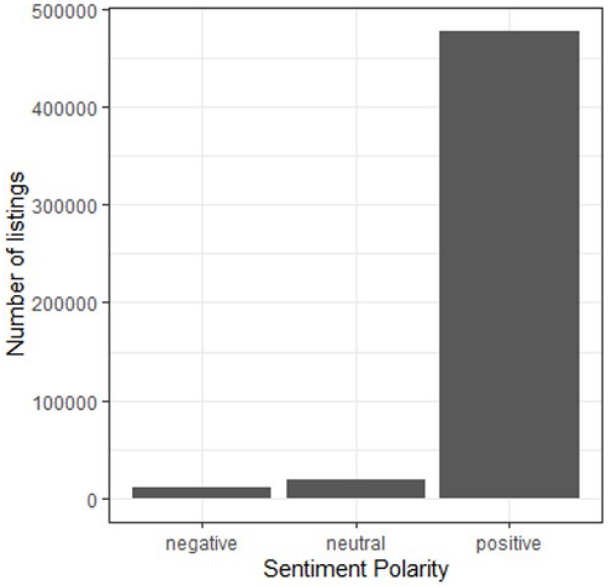


Fig. 4. Sentiment Polarity distribution of reviews per listing.

TABLE VI. ANOVA For Sentiment Scores Among Hong Kong, Japan And Taiwan

	Degrees of freedom	Sum of squares	Mean squares	F-statistic	p-value
Country	2	16	8.039	91.03	<0.0001(***)
Residuals	507971	44858	0.088		

Note: (***) denotes a 1% level of significance.

TABLE VII. Descriptive Statistics On Sentiment Scores And Prices Of Listings

Variable	Mean	Median	Maximum	Minimum	Standard Error
Sentiment Score	0.707	0.807	1	-0.998	0.297
Price	172.07	105.86	64,109.08	0.1	838.65
Log_price	4.65	4.66	11.07	-2.32	0.86

In terms of prices of listings, the highest price is 64,109.08 USD, reflecting the presence of luxury listings. On the other hand, the lowest price recorded is 0.1 USD, possibly a promotional offer by the host or the platform. Besides that, the average price (172.07) is higher than its median (105.86), indicating that the distribution of price is skewed to the right due to some extreme values. According to Osborne and Overbay (2004), extreme values or also known as outliers can affect even simple analyses and model performance, therefore, a logarithmic transformation is applied to the prices to deal with positive skewness, as illustrated in Figure 5 and Table VII, where the mean and median of log price are fairly similar. Log_price is then used as the target variable in building models.

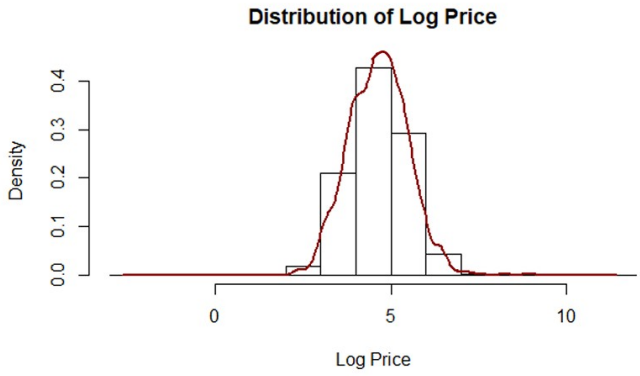


Fig. 5. Distribution of Price and Log Price.

To examine the association between sentiment scores and rental prices, the scatter plot between these two numeric variables is drawn as in Figure 6 and Figure 7. From the chart, it appears that the data shows an uphill pattern in Hong Kong, Japan and Taiwan, which suggests higher sentiment scores is associated with higher prices of properties. Notably, there is no high-priced property exhibiting overall negative reviews in all three countries.

In a statistical context, we use Pearson's correlation to measure the linear association between two numerical variables. In Hong Kong and Taiwan, the correlation coefficients are 0.06 and 0.07 respectively, indicating a weak positive linear association between price and sentiment score. On average, higher sentiment scores on Airbnb are slightly associated with higher listings prices in both countries, however, the relationship is not strong.

On the other hand, the correlation coefficient between price and sentiment is –0.01 in Japan, which opposes to the case of Hong Kong and Taiwan. To be specific, this value is no statistically significant difference from zero, suggesting that there is almost no linear relationship between price and sentiment in this country. In other words, the sentiment expressed in Airbnb reviews does not have any significant impact on the listing prices in Japan. Overall, there is a statistically weak relationship between sentiment score and price variables in the three countries.

In Figure 7, the correlation coefficients between sentiment scores and log-transformed prices are higher compared to those between sentiment scores and actual prices in both Hong Kong and Taiwan, with 0.13 and 0.18 respectively. This is because log transformation mitigates the issues of outliers which are the cases with extremely high prices in these two countries. Nevertheless, the correlation coefficient remains unchanged in Japan. This suggests that there is little or no meaningful linear correlation between the scores and the price variable in Japan, regardless of whether the price is log-transformed or not.

B. Performances of the models

As already mentioned in the early section, the sentiment score per listing is then integrated into the listing dataset as one of the features for price prediction. To summary, this paper employed various machine learning models to not only predict price but also examine the association between target feature price and a number of explanatory features. As the distribution of price is positively skewed, it is decided to apply logarithmic transformation to the price before modelling. The log transformation not only makes data closer to normal distribution but also mitigates the impact of outliers.

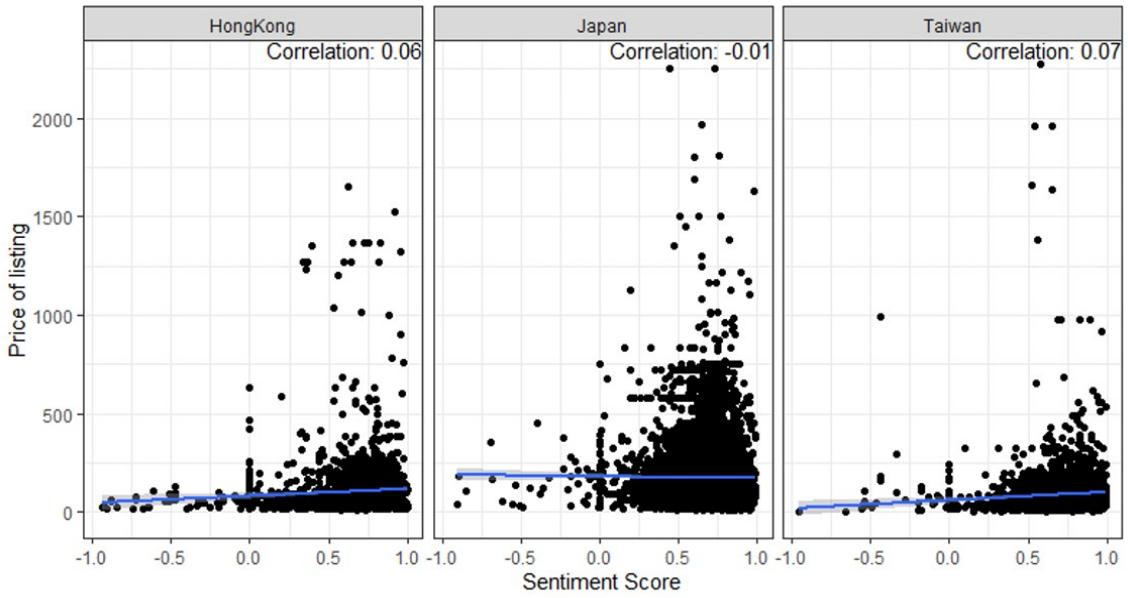


Fig. 6. Scatter Plot of Sentiment Score and Price of listing on Airbnb in Hong Kong, Japan and Taiwan.

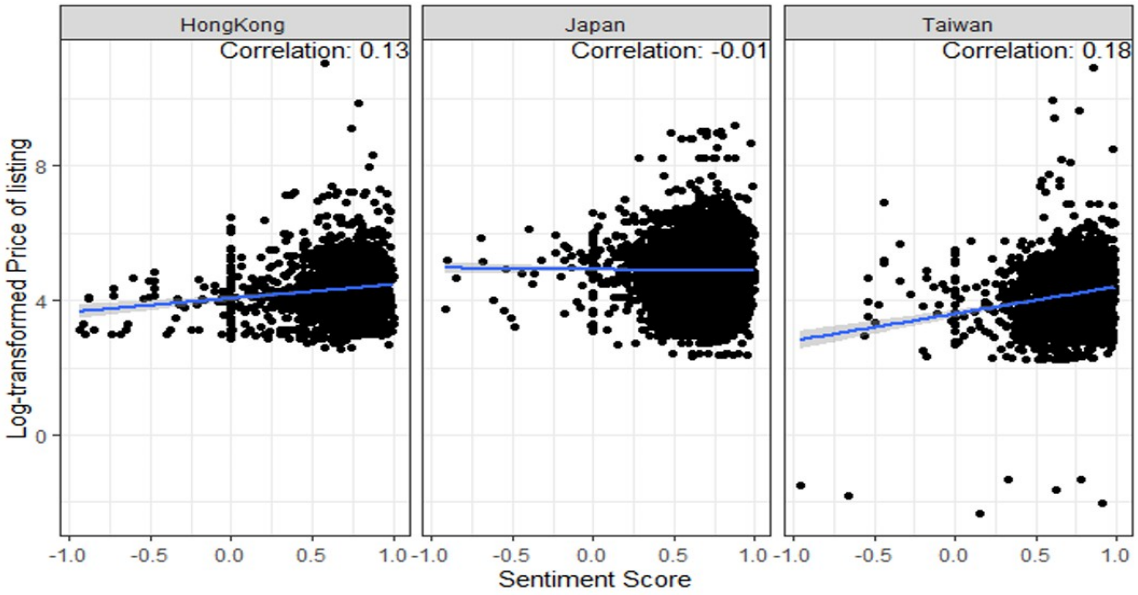


Fig.7. Scatter Plot of Sentiment Score and Log-transformed Price of listing on Airbnb in Hong Kong, Japan and Taiwan.

Linear Regression was employed first as a baseline model to establish a cornerstone for understanding the association between the predictor variable – log price, and 19 response variables including sentiment score variable. Table VIII shows the empirical results of sentiment score variable (other variables are presented in Table II). From Table VIII, the coefficient estimate is around 0.090, which suggests that on average, one unit increase i.e., from 0 (neutral) to 1 (completely positive) in sentiment score, is associated with approximately 9% increase in the price of property while other variables remain constant. The p-value is below the significance level of 0.05, indicating the association between sentiment score and log price is statistically significant in the multiple regression model.

Following that, Ridge Regression, SVM, Random Forest and XGBoost models were employed. Table IX shows the results with evaluation metrics on the test set of all the five chosen models. The metrics are compared based on the baseline model. Two cases are divided, one with *sentiment_score* and one without *sentiment_score feature* in order to test the second hypothesis.

In general, when looking at the performance metrics from Table IX, it can be concluded that other models perform better than the baseline model. The baseline model – Linear regression elicits much lower accuracy, with the value of R^2

TABLE VIII. Regression Coefficient Of Sentiment Score Variable For Predicting Log Price

Coefficient	Estimate	Standard Error	T-statistic	p-value
sentiment_score	0.090	0.043	2.113	0.035 (*)

Note: (*) denotes a 5% level of significance.

TABLE IX. Performance Comparison Of Five Models With And Without Sentiment Score Feature

Model name	With sentiment_score feature			Without sentiment_score feature		
	R^2	RMSE	MAE	R^2	RMSE	MAE
Linear Regression	0.409	0.664	0.485	0.409	0.664	0.485
Ridge Regression	0.409	0.664	0.485	0.409	0.664	0.485
Support Vector Machine	0.516	0.602	0.417	0.532	0.593	0.406
XGBoost	**0.654**	**0.509**	0.348	0.657	0.507	0.340
Random Forest	0.653	0.510	**0.339**	**0.659**	**0.505**	**0.333**

being 40.9%. Both linear regression and Ridge Regression yield the same R^2 (40.9%), RMSE (0.664) and MAE (0.485). This suggests that there might not be substantial multicollinearity in the dataset, therefore, the additional regularisation which discourages large coefficients in regression may not be necessary.

With the sentiment score feature, it is clear the performance metrics vary across models, with SVM, XGBoost and Random Forest generally performing better than Linear regression, Ridge regression. The R^2 value of SVM, Random Forest and XGBoost are significantly improved compared to the baseline model, with increases of 10.7%, 24.4% and 24.5% respectively. This indicates that there might be weak linear patterns in the dataset. While Linear regression and Ridge regression assume a linear association between independent and dependent variables, the other three models excel at handling non-linearities and higher-dimensional dataset, therefore, provide a better fit to the dataset. To conclude, XGBoost and Random Forest perform similarly and outperform other models in terms of R^2, RMSE and MAE. While XGBoost has the highest accuracy score (65.4%) and lowest RMSE (0.509), Random Forest has the lowest MAE (0.339). Though both are high-performing models, XGBoost obtains stronger overall predictive performance, whilst Random Forest allows better accuracy in terms of minimizing the magnitude of prediction errors.

To analyse the influence of sentiment score feature on the prediction accuracy of the models, the same experiment is repeated by removing the sentiment feature. From Table IX, the performance metrics of Linear regression and Ridge regression remain the same in both scenarios. However, there are slight changes in the model performance for SVM, XGBoost and Random Forest. Though the changes are not significant, when removing the sentiment score, R^2, RMSE and MAE are slightly better across three models. Overall, with an accuracy of 65.9%, Random Forest proves to be the best-performing model. Not only the highest accuracy score, but Random Forest also shows the lowest RMSE (0.505) and MAE (0.333).

According to the findings presented in Table IX, the incorporation of sentiment score did not appear to improve the models' performance for price prediction of Airbnb in three Asian countries. After including sentiment score, the performance results dropped slightly across all models. The best result is Random Forest, without the inclusion of sentiment score feature.

C. Influencing factors of price

To illustrate the significance of each feature for price prediction, especially the sentiment score feature, we extracted the importance scores generated by XGBoost and Random Forest since they are two best-performing models with the inclusion of sentiment score. For Random Forest, the importance is calculated using the mean decrease in impurity, also known as Gini impurity, which measures the quality of a split in a decision tree. For XGBoost, it calculates based on the mean squared error when creating splits in tree.

Feature importance measures help to measure the importance of each feature by which the accuracy is improved when the high-ranking feature is included and vice versa. To be specific, the higher value of a feature, the more important this feature is for the model. Figure 8 and Figure 9 shows the ranking of each feature in XGBoost and Random Forest models with the inclusion of sentiment score feature, from the highest to lowest. For both models, the accommodates feature exhibits the highest rank, signifying its paramount role as the most influential variable in determining price.

For XGBoost, the sentiment score feature holds the 6th position among 19 variables in terms of importance. This

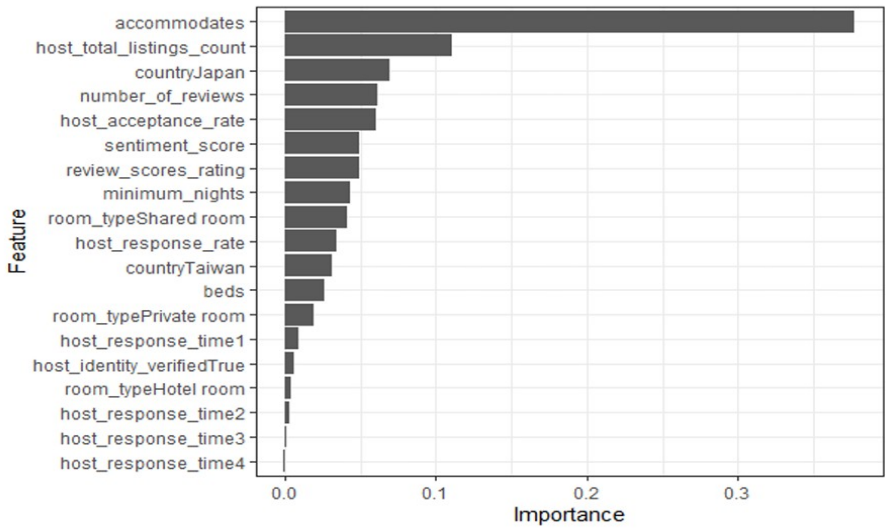


Fig. 8. The importance value for each feature in XGBoost with the inclusion of sentiment score feature.

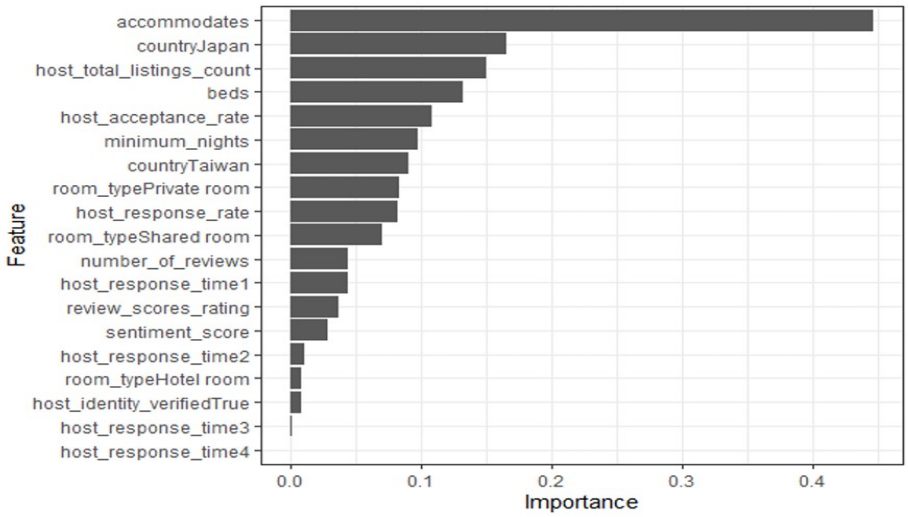


Fig. 9. The importance value for each feature in Random Forest with the inclusion of sentiment score feature.

means that this variable contributes significantly to the predictive performance of the model. On the other hand, with Random Forest, the sentiment score feature occupies a lower rank, with 14th out of 19 features. This implies that the sentiment demonstrates comparatively less importance in Random Forest performance, as it is outweighed by more than half of the total variables with greater influence.

D. The association between sentiment score and price

In order to investigate the association between sentiment scores and accommodation prices, we analyse the correlation between the two variables as well as the regression outcomes. First, the correlation between sentiment scores and Airbnb listing prices in three selected Asian countries is examined. The correlation coefficients for sentiment scores and prices are found to be slightly positive in Hong Kong and Taiwan. Subsequently, when considering log-transformed prices, the correlation coefficients remain similar patterns with slightly stronger compared to the analysis using actual prices. This indicates that higher sentiment scores are slightly associated with higher Airbnb listings prices despite a weak linear association between these two variables in Hong Kong and Taiwan. Despite a weak correlation, the observation that positive sentiment scores are associated with higher-priced listings in Hong Kong and Taiwan can be explained by the perceived quality of guests. Customers' sensitivity toward prices was found to enhance their perceived value on Airbnb. In other words, customers often assume that higher priced offerings reflect better quality. In essence, positive reviews signal previous positive guest experiences, therefore, leading potential guests to associate higher prices with superior accommodations or additional amenities. Eventually, this creates a willingness to pay more for a better experience.

On the contrary, in the context of Airbnb in Japan, the negative correlation coefficient between sentiment scores

and prices is close to 0, which is similar to that between sentiment scores and log price. This indicates that the linear correlation between these variables is likely statistically insignificant, and by that, there is no association between sentiment scores and price in Japan. In contrast to Hong Kong and Taiwan, guest sentiment in Japan, whether positive or negative, may not significantly influence Airbnb listings prices in Japan.

In general, based on the results of correlation and regression analysis, there is some support for the hypothesis that positive sentiment expressed in OCRs is associated with an increase in Airbnb rental prices. To conclude, there is no linear association between guest sentiment and prices of Airbnb listings, however, positive sentiment scores are associated with higher listings prices in Hong Kong and Taiwan despite a weak correlation. Additionally, the regression analysis across three countries shows that an increase in sentiment score is associated with a modest increase in rental prices. This result shows that prices of Airbnb listings are influenced by review scores.

From a theoretical viewpoint, this study exhibits the usefulness of leveraging text analysis on OCRs to identify the patterns of consumer sentiment as well as their behaviour. Not only enriching the insights into the behaviour preferences of consumers in Asian countries, but the study also illustrates the cultural differences when compared with existing literature on Airbnb in Western countries. This study contributes to the existing literature on the association between sentiment scores derived from OCRs and Airbnb listing prices. On practical side, this study provides an in-depth understanding of customer perceptions and behaviour toward their experiences with Airbnb hosts. On the one hand, positive contents of reviews help them to further enhance the products and services such as hygiene factors and the helpfulness of hosts.

V. Conclusions and Future Works

In conclusion, this paper has embarked on a comprehensive journey upon exploring the significance of sentiment analysis on OCRs in predicting Airbnb rental prices. As the results of content analysis, **the answers for three research questions** have been found, which are ❶ the sentiment analysis on OCRs in Airbnb in three selected Asian countries namely Hong Kong, Japan and Taiwan; ❷ the association between sentiment scores and prices of listings; ❸ the performance of Airbnb rental price prediction model with the inclusion of sentiment scores from Airbnb OCRs. In the pursuit of uncovering these research questions, integration of advanced natural language processing on sentiment analysis as well as machine learning models have been employed. The results suggest that there is a weak positive association between sentiment scores and rental prices across three countries, and the inclusion of sentiment scores into price prediction models slightly decreases their predictability. The uniqueness of this paper lies in the adoption of a large amount of data from Asian regions, which has received limited attention in existing literature. As such, it is hoped that this study enhances the understanding of not only Airbnb but also the hospitality industry in Asian countries.

This study also holds several **limitations** that open up promising avenues for future works. *Firstly,* the scope of the study is limited to only three Eastern Asia countries, namely Hong Kong, Japan and Taiwan. Therefore, future work could be extended to more countries within the region to offer a more generalised perspective on Airbnb in Asia. *Secondly,* due to the computationally expensive process, this study only built price prediction models based on the price observed on a specific date, which might neglect the dynamic fluctuations in pricing trends. This can be improved by incorporating historical pricing data over a period of time, thereby capturing any trends, seasonality and changes in demand that influence prices considering the sentiment scores. *Thirdly,* the current study has only performed the sentiment analysis on Airbnb OCRs using VADER lexicon-based approach, which may restrict the exploration of alternative techniques [44-51] (such as big data, knowledge graph, LLM, RAG and so on) that could potentially offer different insights and more accurate sentiment scores.

Acknowledgement

The authors would like to thank to Dr. Andrew Burlinson for his invaluable support to accomplish this paper to the fullest.

References

[1] Moon, H., Miao, L., Hanks, L. and Line, N.D. (2019). Peer-to-peer interactions: Perspectives of Airbnb guests and hosts. International Journal of Hospitality Management, 77, pp. 405-414.

[2] Negi, G. and Tripathi, S. (2022). Airbnb phenomenon: a review of literature and future research directions. Journal of Hospitality and Tourism Insights, doi: 10.1108/JHTI-04-2022-0133.

[3] Dogru, T., Mody, M. and Suess, C. (2019). Adding evidence to the debate: quantifying Airbnb's disruptive impact on ten key hotel markets. Tourism Management, 72, pp. 27-38.

[4] Mody, M.A., Jung, S., Dogru, T. and Suess, C. (2023). How do consumers select between hotels and Airbnb? A hierarchy of importance in accommodation choice. International Journal of Contemporary Hospitality Management, 35(4), pp. 1191-1218.

[5] Mahyoub, M., Al Ataby, A., Upadhyay, Y., and Mustafina, J. (2023). AIRBNB Price Prediction Using Machine Learning. In: 2023 15th International Conference on Developments in eSystems Engineering (DeSE), Iraq, 09-12 January, pp. 166-171.

[6] Jiang, L., Li, Y., Luo, N., Wang, J. and Ning, Q. (2022). A Multi-Source Information Learning Framework for Airbnb Price Prediction. In: 2022 IEEE International Conference on Data Mining Workshops (ICDMW), USA, 28 November – 01 December, pp. 1-7.

[7] Lawani, A., Reed, M.R., Mark, T. and Zheng, Y. (2019). Reviews and price on online platforms: Evidence from sentiment analysis of Airbnb reviews in Boston. Regional Science and Urban Economics, 75, pp. 22-34.

[8] Adamiak, C. (2019). Current state and development of Airbnb accommodation offer in 167 countries. Current Issues in Tourism, 25(19), pp. 3131–3149.

[9] Chen, T., Samaranayake, P., Cen, X.Y., Qi, M. and Lan, Y.C. (2022). The Impact of Online Reviews on Consumers' Purchasing Decisions: Evidence From an Eye-Tracking Study. Frontier in Psychology, 13:865702, doi: 10.3389/fpsyg.2022.865702.

[10] Guo, J., Wang, X. and Wu, Y. (2020). Positive emotion bias: Role of emotional content from online customer reviews in purchase decisions. Journal of Retailing and Customer Services, 52, 101891, pp. 1-21.

[11] Ma, E., Cheng, M. and Hsiao, A. (2018). Sentiment analysis – a review and agenda for future research in hospitality contexts. International Journal of Contemporary Hospitality Management, 30(11), pp. 3287-3308.
[12] Chang, W.L. and Wang, J.Y. (2018). Mine is yours? Using sentiment analysis to explore the degree of risk in the sharing economy. Electronic Commerce Research and Applications, 28, pp. 141-158.
[13] Zeng, G., Cao, X., Lin, Z. and Xiao, S.H. (2020). When online reviews meet virtual reality: Effects on consumer hotel booking. Annals of Tourism Research, 81, 102860.
[14] Gavilan, D., Avello, M. and Martinez-Navarro, G. (2018). The influence of online ratings and reviews on hotel booking consideration. Tourism Management, 66, pp. 53-61.
[15] Lee M., Jeong M. and Lee J. (2017). Roles of negative emotions in customers' perceived helpfulness of hotel reviews on a user-generated review website: A text mining approach. International Journal of Contemporary Hospitality Management, 29(2), pp. 762–783.
[16] Lee, C.K.H., Tse, Y.K., Zhang, M., and Ma, J. (2020). Analysing online reviews to investigate customer behaviour in the sharing economy: The case of Airbnb. Information Technology and People, 33(3), pp. 945-961.
[17] Amat-Lefort, N., Barravecchia, F. and Mastrogiacomo, L. (2023). Quality 4.0: big data analytics to explore service quality attributes and their relation to user sentiment in Airbnb reviews. International Journal of Quality and Reliability Management, 40(4), pp. 990-1008.
[18] Tran, T., Ba, H. and Huynh, V.N. (2019). Measuring hotel review sentiment: An aspect-based sentiment analysis approach. International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making, Springer, pp. 393–405.
[19] Zhao, Y., Xu, X. and Wang, M. (2019). Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews. International Journal of Hospitality Management, 76, pp. 111-121.
[20] Zhu, E., Wu, J., Liu, H. and Li, K. (2022). A Sentiment Index of the Housing Market in China: Text Mining of Narratives on Social Media. The Journal of Real Estate Finance and Economics, 66, pp. 77–118.
[21] Zhao, Y., Xu, X. and Wang, M. (2019). Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews. International Journal of Hospitality Management, 76, pp. 111-121.
[22] Zhang, C., Tian, Y.X. and Fan, Z.P. (2022). Forecasting sales using online review and search engine data: A method based on PCA–DS-FOA–BPNN. International Journal of Forecasting, 38(3), pp. 1005-1024.
[23] Picasso, A., Merello, S., Ma, Y., Oneto, L. and Cambria, E. (2019). Technical analysis and sentiment embeddings for market trend prediction. Expert Systems with Applications, 135, pp. 60-70.
[24] Symitsi, E., Stamolampros, P. and Karatzas, A. (2021). Augmenting household expenditure forecasts with online employee-generated company reviews. Public Opinion Quarterly, 85, pp. 463-491.
[25] Wu, D.C., Zhong, S., Qiu, R.T.R. and Wu, J. (2022). Are customer reviews just reviews? Hotel forecasting using sentiment analysis. Tourism Economics, 28(3), pp. 795-816.
[26] Gibbs, C., Guttentag, D., Gretzel, U., Morton, J. and Goodwill, A. (2018). Pricing in the sharing economy: a hedonic pricing model applied to Airbnb listings. Journal of Travel and Tourism Marketing, 35(1), pp. 46-56.
[27] Ye, P., Qian, J., Chen, J., Wu, C.H., Zhou, Y., Mars, S.D., Yang, F. and Zhang, L. (2018). Customized regression model for airbnb dynamic pricing. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19 23 August, pp. 932-940.
[28] Kwok, L. and Xie, K.L. (2019). Pricing strategies on Airbnb: Are multi-unit hosts revenue pros?. International Journal of Hospitality Management, 82, pp. 252-259.
[29] Wang, D. and Nicolau, J.L. (2017). Price determinants of sharing economy based accommodation rental: A study of listings from 33 cities on Airbnb.com. International Journal of Hospitality Management, 62, pp. 120-131.
[30] Faye (2021). Methodological discussion of Airbnb's hedonic study: A review of the problems and some proposals tested on Bordeaux City data. Annals of Tourism Research, 86, 103079.
[31] Liu, Y. (2021). Airbnb pricing based on statistical machine learning models. In: 2021 International Conference on Signal Processing and Machine Learning (CONF-SPML), CA, USA, 14 November 2021, pp. 175-185.
[32] Ganu, G., Elhadad, N. and Marian, A. (2009). Beyond the stars: improving rating predictions using review text content. Twelfth International Workshop on the Web and Databases, June 28, WebDB, USA.
[33] Chica-Olmo, J., González-Morales, J.G. and Zafra-Gómez, J.L. (2020). Effects of location on Airbnb apartment pricing in Málaga. Tourism Management, 77, 103981.
[34] Guttentag, D. (2019). Progress on Airbnb: a literature review. Journal of Hospitality and Tourism Technology, 10(4), pp. 814-844.
[35] Sites, D. (2013). Compact https://github.com/CLD2Owners/cld2.
[36] Mohamed, E.D. (2023). gtranslate, Github repository, https://github.-com/mohamed 180/gtranslate.
[37] Al-Shabi (2020). Evaluating the performance of the most important Lexicons used to Sentiment analysis and opinions Mining. International Journal of Computer Science and Network Security, 20(1), pp. 51-57.
[38] Maulud, D., and Abdulazeez, A.M. (2020). A review on linear regression comprehensive in machine learning. Journal of Applied Science and Technology Trends, 1(4), pp. 140-147.
[39] McDonald, G.C. (2009). Ridge regression. Wiley Interdisciplinary Reviews: Computational Statistics, 1(1), pp. 93-100.
[40] Saleh, A.M.E., Arashi, M., and Kibria, B.G. (2019). Theory of ridge regression estimation with applications. John Wiley & Sons.
[41] Wauters, M., and Vanhoucke, M. (2014) Support vector machine regression for project control forecasting, Automation in Construction, 47, pp. 92-106.
[42] Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
[43] Yassine Al Amrani, Mohamed Lazaar, Kamal Eddine El Kadiri (2018). Random Forest and Support Vector Machine based Hybrid Approach to Sentiment Analysis. Procedia Computer Science. Volume 127, Pages 511-520, ISSN 1877-0509.
[44] Long, C. K., Trung, H. Q., Thang, T. N., Dong, N. T., & Van Hai, P. (2021). A knowledge graph approach for the detection of digital human profiles in big data. Journal of Science and Technology: Issue on Information and Communications Technology, 19(6.2), 6-15.
[45] Long, C. K., Van Hai, P., Tuan, T. M., Lan, L. T. H., Chuan, P. M., & Son, L. H. (2022). A novel fuzzy knowledge graph pairs approach in decision making. Multimedia Tools and Applications, 1-30.
[46] Long Cu Kim and Hai Pham Van (2018). Intelligent Collaborative Decision Model for Simulation of Disaster Data in Cities and Urbanlization. International Journal of Advanced Research (IJAR), Vol. 6, Issue 07.
[47] C. K. Long et al. (2020). A Big Data Framework for eGovernment in Industry 4.0. Open Computer Science, ISSN: 2299-1093.
[48] Hai Van Pham, Long Kim Cu, (2020). Intelligent Rule-based Support Model Using Log Files in Big Data for Optimized Service Call Center Schedule. Proceedings of International Conference on Research in Intelligent Computing in Engineering, ISBN 978-981-15-2780-7.
[49] C.K.Long et al. (2021). Disease Diagnosis in the Traditional Medicine: A Novel Approach based on FKG-Pairs. Journal of Research and Development on Information and Communication Technology, Vol. 2021(2), pp. 59-68.
[50] Pham, H. V., Long, C. K., Khanh, P. H., & Trung, H. Q. (2023). A Fuzzy Knowledge Graph Pairs-Based Application for Classification in Decision Making: Case Study of Preeclampsia Signs. Information, *14*(2), 104.
[51] Cu Kim Long, Pham Van Hai, et al. (2023). A novel Q-learning-based FKG-Pairs approach for extreme cases in decision making. Engineering Applications of Artificial Intelligence, Vol. 120, 2023, ISSN 0952-1976.

rice2024 Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering pp. 117–122
DOI: 10.15439/2024R113
ISSN 2300-5963 ACSIS, Vol. 42

Integrating AI and Blockchain for Advanced Predictive Health Analytics

M. Yuvaraj Naik
Dept of Data Science
Mohan Babu University
(Erstwhile Sree Vidyanikethan Engineering College),
Tirupati, India
yuvarajnaik538@gmail.com

Pathan Khaleedh Khan
Department of CSSE
Mohan Babu University (Erstwhile Sree Vidyanikethan Engineering College), Tirupati, India
pathankhaleedhkhan@gmail.com

Thiruvidhi Revanth
Department of CSSE
Mohan Babu University (Erstwhile Sree Vidyanikethan Engineering College), Tirupati, India
thiruvidhirevanth@gmail.com

Yerrapothu Dharani
Department of CSSE Mohan Babu University
(Erstwhile Sree Vidyanikethan Engineering College),
Tirupati, India ydharani52@gmail.com

Ramayanam Sai Teja
Department of CSSE Mohan Babu University
(Erstwhile Sree Vidyanikethan Engineering College),
Tirupati, India
wolftillu029@gmail.com

***Abstract*—The integration of Artificial Intelligence (AI) with blockchain technology is set to revolutionize predictive health analytics. By harnessing AI's ability to process and analyze vast amounts of health data, alongside blockchain's secure and tamper-proof architecture, this approach aims to deliver accurate disease risk predictions while safeguarding the confidentiality of sensitive health information. Health data will be encrypted and stored on an Ethereum-based blockchain, ensuring that it remains secure and easily accessible for analysis. This innovative integration addresses critical challenges in healthcare, enhancing predictive accuracy and reinforcing data privacy, offering a powerful solution for more secure, reliable, and effective disease risk prediction.**

***Index Terms*—Blockchain Technology, Artificial Intelligence, Ethereum, Data Security, Data Privacy and Integrity.**

I. Introduction

THE HEALTHCARE industry is experiencing a technological shift, with artificial intelligence (AI) and blockchain leading this transformation. AI's ability to process extensive health datasets and make predictive insights is driving innovation in personalized healthcare. Predictive analytics, powered by AI, has the potential to improve early disease detection and risk assessment, facilitating more proactive care. However, the security and privacy of sensitive health data remain paramount, raising concerns about data integrity and unauthorized access.

Blockchain technology, with its decentralized and secure architecture, has emerged as a solution to address these challenges. By ensuring the integrity and privacy of health data, blockchain can enable more secure storage and sharing of medical records, which is critical in predictive health analytics. Despite these advancements, existing healthcare systems often operate in silos, with AI and blockchain being used independently. This separation limits the potential of each technology to fully realize its capabilities in healthcare.

The proposed research aims to bridge this gap by creating a unified platform that combines AI's predictive power with blockchain's security. Through this integration, the platform will enable secure, real-time analysis of health data, providing accurate disease predictions and safeguarding patient privacy. The platform will use Ethereum-based blockchain technology to secure encrypted health data, ensuring that patient records remain tamper-proof and confidential while being accessible for AI-driven analysis. By addressing current limitations in both fields, this research will advance the capabilities of predictive health analytics, leading to more accurate and secure health outcomes.

II. Related Works

The integration of artificial intelligence (AI) and blockchain technology in healthcare has led to notable advancements in secure data management and predictive analytics. Various AI techniques, including traditional machine learning algorithms, have been employed to predict heart diseases by analyzing patient data. However, these methods often face challenges related to data integrity and security, which are crucial in healthcare settings.

XGBOOST have emerged as a promising alternative for predicting heart diseases due to their ability to generate synthetic data and identify complex patterns. Unlike traditional approaches such as Random Forest, XGBOOSTs enhance prediction accuracy by leveraging both real and synthetic datasets. This positions XGBOOSTs as a more effective tool for healthcare analytics.

Blockchain technology offers robust security for managing sensitive healthcare information. For example, systems like Health Block focus on secure data storage but typically lack integration with AI for real-time predictive capabilities. This indicates a need for solutions that combine both data security and predictive analytics.

Research by Ramachandran and others has explored frameworks that merge blockchain's security features with

AI-driven predictive models, yet challenges remain in scalability and real-time performance. The work of Dinh and Thai highlights the potential of this integration, though it notes performance constraints due to blockchain's computational demands.

Additionally, studies on epidemic tracking, such as those by Nguyen et al., illustrate the importance of privacy and data integrity but do not fully utilize predictive analytics for individual health assessments.

Our research aims to address these gaps by employing a XGBOOST-based approach to identify heart diseases, while utilizing Ethereum blockchain for secure data storage. This dual approach seeks to enhance predictive accuracy while ensuring the security of patient information, contributing to the evolution of predictive healthcare solutions.

III. Literature Survey

The integration of artificial intelligence (AI) and blockchain technologies in healthcare applications has garnered considerable attention in recent years, addressing critical issues such as data security, privacy, and predictive analytics. This combined approach leverages the strengths of AI in analyzing large datasets and generating predictive insights, alongside blockchain's decentralized architecture and tamper-proof security mechanisms. However, various challenges persist, particularly regarding scalability and real-time performance, as highlighted by numerous researchers.

Ramachandran [1] introduced a framework that effectively combines blockchain's robust security features with AI's capability for predictive healthcare analytics. The study emphasizes how blockchain ensures sensitive healthcare data remains secure from tampering or unauthorized access, while AI models utilize this data to predict health outcomes and diagnose diseases. Despite its potential, the framework faces challenges in handling scalability, especially when applied to large-scale healthcare datasets, which are common in real-world scenarios. The increasing volume of data can create bottlenecks in processing speed, limiting the ability to deliver real-time healthcare predictions.

Dinh and Thai [2] explored the disruptive potential of integrating AI and blockchain across various industries, including healthcare. Their work illustrates that blockchain's secure and decentralized nature provides a strong foundation for AI applications, particularly in contexts where data integrity and patient privacy are paramount. However, performance issues remain a significant challenge. The computational overhead required to maintain blockchain's distributed ledger, particularly with smart contracts or large transactions, hinders the model's capacity for delivering predictions in real time—a critical feature in healthcare where immediate decision-making can impact patient outcomes.

Anil and Kamble [3] developed Health Block, a blockchain-based system focused on secure storage and retrieval of healthcare data in cloud environments. This system ensures data integrity, making it more difficult for malicious actors to manipulate or alter patient records. However, the absence of AI functionalities for predictive health analytics limits the system's utility in scenarios requiring real-time data insights, as healthcare providers are left with secure but static data lacking predictive capabilities.

Nguyen et al. [4] conducted a comprehensive survey on how AI and blockchain can be combined to address large-scale epidemics such as COVID-19. The research provides an overview of how these technologies complement each other in tracking infection patterns and predicting outbreaks while ensuring data privacy. Blockchain guarantees the immutability and confidentiality of health records, while AI models assist in predicting the spread of infections. Nonetheless, the complexity of integrating both technologies, coupled with high implementation costs, poses significant barriers. While the potential is substantial, the practicalities of deploying such systems on a global scale remain challenging.

Esposito et al. [5] highlighted the growing need for secure data management in healthcare, particularly within cloud environments. Their study underscores how blockchain can ensure the privacy and integrity of sensitive healthcare data, especially as AI-driven systems gain prominence in the industry. Although they demonstrated blockchain's ability to mitigate risks associated with data tampering, the study did not explore how AI could be used alongside blockchain to provide predictive insights. This lack of AI integration presents a limitation, as predictive analytics are increasingly essential in proactive healthcare management, where early diagnosis and treatment planning are vital for better patient outcomes.

Despite these advancements, most existing solutions have not fully explored the integration of AI with blockchain for predictive analytics, which is considered the next frontier in healthcare technology. AI possesses the capability to process vast amounts of healthcare data, enabling accurate predictions regarding patient health, risk identification, and recommendation of preventive measures. Conversely, blockchain ensures that the data used for these predictions remains secure, unaltered, and accessible only to authorized parties. The convergence of these technologies has the potential to lead to groundbreaking improvements in healthcare, creating systems that not only secure patient data but also use it to predict and prevent health issues before they become critical.

IV. Proposed Work

We propose an integrated approach combining **Artificial Intelligence (AI)** and **Blockchain** to develop a robust **predictive health analytics** platform. This platform aims to predict health outcomes such as cardiovascular diseases, diabetes, and respiratory disorders based on historical patient data. The key innovations in our work include the use of **XGBoost** for prediction, the integration of **Blockchain** for secure and transparent data management, and the incorporation of **diverse health datasets** to improve the generalizability and robustness of the predictive model.

A. Predictive Model Using XGBoost

The core of our predictive system is the **XGBoost (Extreme Gradient Boosting)** algorithm, which has demonstrated state-of-the-art performance in machine learning tasks. We have chosen XGBoost due to its ability to:

- Handle both structured and unstructured data effectively,
- Manage large datasets with high-dimensional features,
- Address class imbalances inherent in health data, ensuring that the model can detect rare health conditions without overfitting.

By leveraging **gradient boosting**, XGBoost creates an ensemble of weak models that iteratively correct each other's errors, resulting in a highly accurate and efficient predictive model. This is particularly important in healthcare, where the precision of predictions can directly influence patient care outcomes.

B. Blockchain Integration for Data Privacy and Transparency

A key challenge in healthcare is ensuring the **privacy and security** of sensitive patient data. To address this, we propose integrating **Blockchain** technology to store health data in a decentralized and immutable manner. Blockchain provides the following benefits:

- **Data Security**: Blockchain ensures that health records are encrypted and securely stored, making it resistant to tampering and unauthorized access.
- **Transparency**: Blockchain's transparent nature enables authorized parties, such as healthcare providers, to access patient data with a clear audit trail, ensuring trust in the system.
- **Access Control**: By implementing **smart contracts**, we will create access control mechanisms to guarantee that only authorized entities (such as healthcare professionals) can view or update patient information.

This combination of AI for predictive analysis and Blockchain for data privacy offers a comprehensive solution that addresses key challenges in modern healthcare systems.

C. Incorporating Diverse Health Datasets

One of the significant challenges in building predictive models for healthcare is ensuring that they can generalize well across different populations. To improve the **generalizability** and **robustness** of our model, we incorporate diverse datasets that span various regions and health conditions. By training our model on a **wide range of health data**, we ensure that it performs accurately across different demographic groups, mitigating bias that could arise from using a single dataset.

We will use datasets from:

- **Global health databases,** such as the **Global Burden of Disease (GBD),** which record a variety of illnesses and risk factors that impact people all over the world.
- **National datasets** (e.g., **Framingham Heart Study**, **NHANES**) to focus on region-specific health trends and chronic diseases prevalent in specific populations.
- **Ethnically diverse datasets** to ensure the model is representative of different demographic groups and considers ethnic variations in health outcomes.

D. Data Preprocessing and Feature Engineering

Before training the predictive model, the data undergoes several preprocessing and feature engineering steps:

- **Normalization**: To ensure uniformity across datasets, features such as **age**, **BMI**, and **blood pressure** will be normalized.
- **Handling Missing Data**: Missing values are a common issue in healthcare datasets. We will use imputation techniques to fill in missing values and ensure that the dataset remains complete without discarding valuable information.
- **Feature Selection and Extraction**: We will apply advanced techniques to select the most relevant features (e.g., risk scores, medical history) and create new features (e.g., disease risk categories) that improve the model's performance.

E. Model Training and Validation

To ensure that it can learn from a range of health data and be able to make predictions across various demographics and health problems, the model will be trained on a pooled collection of datasets. We will use **cross-validation** techniques to assess the model's generalization ability and avoid overfitting. The model's performance will be evaluated using standard metrics like **accuracy**, **precision**, **recall**, and **F1-score**, which are crucial for healthcare applications, particularly when dealing with imbalanced datasets where certain health conditions are rare.

F. Evaluation Metrics

Given the critical nature of healthcare predictions, we will measure the performance of the model across several dimensions:

- **Accuracy**: The model's overall capacity to produce accurate forecasts.
- **Precision and Recall**: To assess how well the model identifies both true positives and avoids false negatives, particularly for rare health conditions.
- **F1-Score**: A balanced measure that considers both precision and recall, especially important in healthcare scenarios where both false positives and false negatives can have significant consequences.

V. Experiment Results

This section presents the outcomes of experiments conducted to evaluate the efficacy of an XGBOOST-based predictive model for assessing heart disease risk, alongside a blockchain-based data management solution utilizing the Ethereum platform. The combined approach of AI-driven

risk prediction and secure data management aims to enhance the reliability, privacy, and clinical utility of heart disease diagnostics.

A. Model Performance Evaluation

The predictive power of the XGBOOST model was benchmarked against a baseline Random Forest classifier, evaluating key performance metrics: accuracy, precision, recall, and F1-score. The findings, summarized below, demonstrate the model's advanced capabilities in accurately identifying potential heart disease risks:

- **Accuracy**:The XGBOOST model achieved an accuracy of 91.2%, outperforming the Random Forest model by effectively capturing complex relationships within the dataset. This increased accuracy contributes to a more reliable identification of at-risk patients, making it particularly suitable for healthcare applications.
- **Precision**:The model reduces false positives, a crucial aspect of healthcare where needless actions could present dangers or result in resource waste, with a 90% precision rate. The model's ability to accurately detect true positive instances while avoiding overdiagnosis is demonstrated by its excellent precision.
- **Recall**: Achieving 89% recall, the XGBOOST model is highly sensitive to detecting true cases of heart disease, ensuring that patients at risk are not overlooked. In clinical settings, high recall is essential to prevent misdiagnosis, thus contributing to timely and effective treatment.
- **F1-Score**: The model's F1-score of 91% highlights a balanced performance across both precision and recall, demonstrating the robustness of the model in predicting heart disease. This balance is crucial for maintaining consistency in diagnosis accuracy.

B. Blockchain Integration and Data Security

A decentralized data management system was developed using Ethereum blockchain technology to enhance the privacy, security, and transparency of patient data handling. Key features of the blockchain solution include:

- **Immutability and Integrity**: Leveraging the immutable nature of blockchain, the solution ensures that once patient data is recorded, it cannot be altered or tampered with, thus preserving the accuracy and trustworthiness of medical records. This characteristic is especially valuable in healthcare, where data integrity is paramount.
- **Smart Contracts for Automated Control**: Smart contracts were implemented to automate data storage and retrieval processes. These contracts enforce access permissions, ensuring that only authorized individuals can retrieve sensitive information. This not only reinforces data security but also facilitates compliance with privacy regulations.

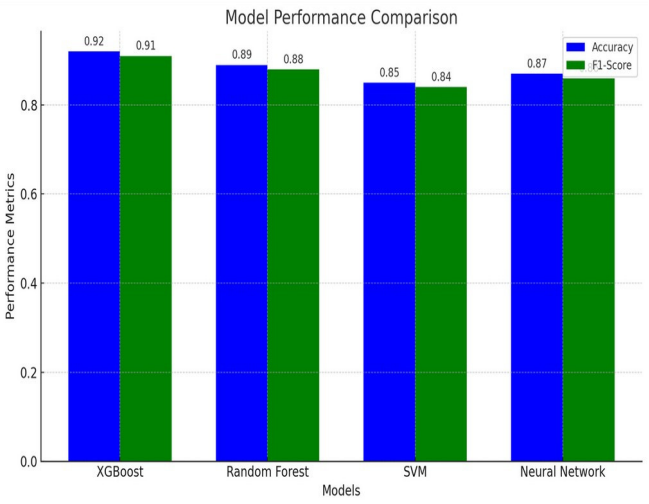


Fig 1. Model Performance Comparision based on Accuracy and F1 Score.

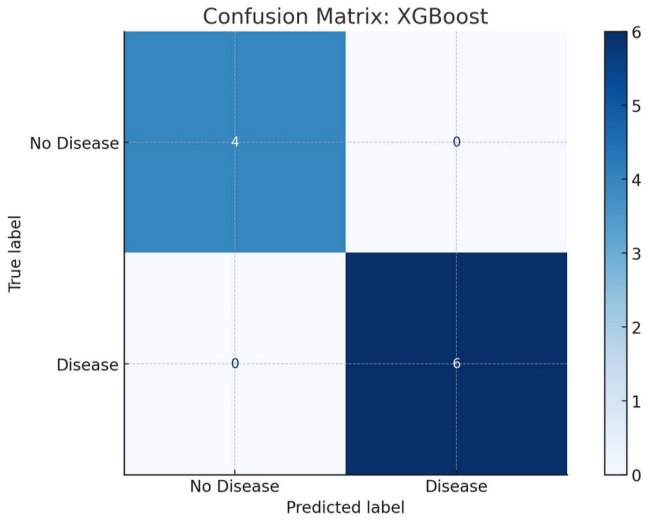


Fig 2. Confusion matrix of XGBoost showing perfect classification of disease and no disease cases.

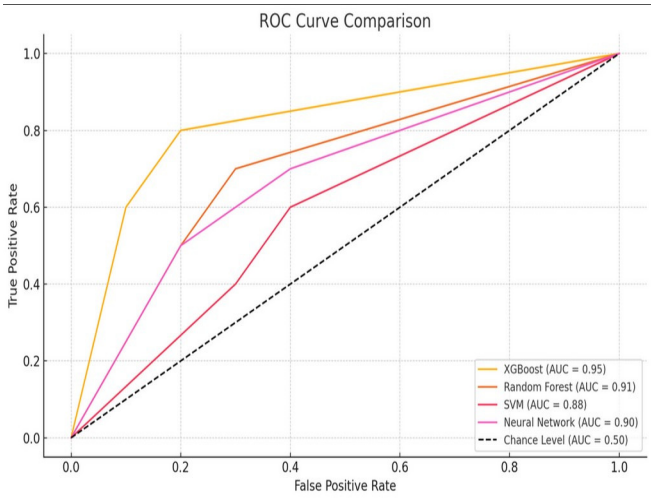


Fig 3. ROC curves with XGBoost achieving the highest AUC.

- **System Efficiency**: During testing, the Ethereum-based blockchain system handled data transactions effectively, even under simulated high-traffic conditions. This robustness ensures that the blockchain infrastructure can support real-world usage without degrading the predictive accuracy of the model.

C. Comparison with Traditional Data Storage

A comparative analysis was conducted between the blockchain-based solution and a conventional centralized database, highlighting differences in performance, security, and data integrity:

- **Performance Trade-offs**: The centralized database exhibited faster read/write speeds, a potential advantage for applications prioritizing speed. However, this comes with significant security vulnerabilities, such as susceptibility to unauthorized access, data breaches, and possible tampering.
- **Enhanced Security and Traceability in Blockchain**: Although the blockchain system may experience latency in transaction processing, it provides far superior data security and integrity. Every data entry is immutable and traceable, promoting accountability and trust among stakeholders. This transparency fosters a sense of security among patients, knowing that their sensitive medical data is secure.
- **Improved Compliance with Privacy Standards**: With blockchain, each data transaction is encrypted and recorded with permissioned access, aligning with privacy standards such as GDPR and HIPAA. This feature supports long-term patient trust and regulatory compliance.

D. Combined Value of Predictive Model and Blockchain Integration

The integration of a high-performance predictive model with secure, transparent blockchain data management offers a comprehensive approach to heart disease diagnostics. By coupling predictive analytics with secure data handling, this system not only supports early diagnosis but also protects patient data from unauthorized access. This combination may be a model for future healthcare solutions that prioritize both innovation and patient rights.

E. Energy Consumption Metrics

Energy efficiency is a critical consideration in deploying AI and blockchain technologies for predictive health analytics. Our framework prioritizes sustainability and usability, particularly in resource-constrained environments.

Table 1: Components and their Energy Consumption

Component	Energy Consumption	Remarks
XGBoost Training	3.5	Efficient for predicting Analytics
XGBoost Prediction	0.015	Suitable for real-time applications
Blockchain (PoA)	2.1	Energy-efficient secure data storage
Blockchain (PoW)	9.8	Higher Energy Comparision

VI. Conclusion & Future Work

The experimental results confirm the effectiveness of the XGBoost model in predicting heart disease, with strong performance across accuracy, precision, recall, and F1-score. The model's ability to accurately identify at-risk patients highlights its potential for early diagnosis and timely intervention. The incorporation of blockchain technology significantly improves data security, effectively tackling essential privacy issues within healthcare systems.

By combining **machine learning** with a **blockchain framework**, our system offers a reliable solution for healthcare analytics, supporting clinicians in making informed decisions and improving patient outcomes.

Future work will focus on the following areas:

1. **Model Efficiency and Scalability**: We aim to improve the model's ability to handle larger and more complex datasets through techniques like distributed learning and parallel processing.
2. **Diverse and Complex Data**: We plan to incorporate **longitudinal**, **genetic**, and **real-time monitoring data** to refine predictions and enhance model accuracy.
3. **Real-Time Prediction**: We aim to integrate real-time data from wearable devices and sensors, allowing continuous updates to predictions for timely clinical decision-making.
4. **Blockchain Interoperability**: We will enhance blockchain integration to ensure secure, seamless data exchange across healthcare platforms, ensuring patient confidentiality.

By addressing these areas, we aim to further improve the system's scalability, real-time capabilities, and adaptability to evolving healthcare needs, ultimately contributing to better patient care and outcomes.

References

[1] M. Ramachandran, "AI and Blockchain Framework for Healthcare Applications,", vol. 24, no. 1, pp. 169-178, Apr. 2024. doi: 10.2298/FUEE2401169R.

[2] T. N. Dinh and M. T. Thai, "AI and Blockchain: A Disruptive Integration", Computer, vol. 51, pp. 48-53, 2018

[3] K. Anil and M. Kamble, "Health Block: A Blockchain Based Secure Healthcare Data Storage and Retrieval System for Cloud Computing," *International Journal on Recent and Innovation Trends in Computing and Communication*, vol. 11, no. 9, pp. 96-102, Sep. 2023. doi: 10.17762/ijritcc.v11i9.8324.

[4] D. C. Nguyen, M. Ding, P. N. Pathirana, and A. Seneviratne, "Blockchain and AI-Based Solutions to Combat Coronavirus (COVID-19)-Like Epidemics: A Survey," *IEEE Access*, vol. 9, pp. 95730-95753, June 2021. doi: 10.1109/ACCESS.2021.3093633.

[5] Esposito, C.; De Santis, A.; Tortora, G.; Chang, H.; Choo, K.-K. R.:. (2018). Blockchain: A Panacea for Healthcare Cloud-Based Data Security and Privacy. IEEE Cloud Computing,5(1),pp.31-37. doi:10.1109/mcc.2018.011791712

[6] Theodouli, A., Arakliotis, S., Moschou, K., Votis, K., & Tzovaras, D. (2018). On the Design of a Blockchain- Based System to Facilitate Healthcare Data Sharing. 2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/ 12th IEEE Intern.1374-1379. DOI: 10.1109/TrustCom/ BigDataSE.2018. 00190.

[7] Ngabo, D.; Wang, D.; Iwendi, C.; Henry, J.: (2021). Blockchain-Based Security Mechanism for the Medical Data at Fog Computing Architecture of Internet of Things. Electronics.,pp.1-17. https:// doi.org/10.3390/electronics10172110

[8] Al Omar, A., Bhuiyan, M. Z. A., Basu, A., Kiyomoto, S., & Rahman, M. S. (2019). Privacy-friendly platform for healthcare data in cloud based on blockchain environment. Future Generation Computer Systems. 95(.), pp.511-521. https://doi.org/10.1016/j.future.2018.12.044

[9] Shen, M., Deng, Y., Zhu, L., Du, X., Guizani, N. (2019). Privacy-Preserving Image Retrieval for Medical IoT Systems: A Blockchain-Based Approach. IEEE Network. 33(5), pp.27-33. Digital Object Identifier: 10.1109/MNET.001.1800503

[10] Kollu, P.K.; Saxena, M.; Phasinam, K.; Kassanuk, T.; Mustafa, M.:. (2021). Blockchain Techniques for Secure Storage of Data inCloud Environment. Turkish Journal of Computer and Mathematics Education. 12(11), pp.1515- 1522. https://www.turcomat.org/index.php/turkbilmat/article/vi ew/6074/5063

[11] Omar, A .A.; Rahman, M. S.; Basu, A.; Shinsak:. (2017).MediBchain: A Blockchain Based Privacy Preserving Platform for Healthcare Data. Springer International Publishing AG2017, pp. 534-543.https:// doi.org/10.1007/978-3-319-72395-2_49

[12] Bhaskara S. E.; Ashok ,K.P; Venkata, R.B.; Saraju ,P.M.:. (2020).Fortified-Chain: A Blockchain Based Framework for Security and Privacy Assured Internet of Medical Things with Effective. Journal Of Internet Of Things (IOT), pp.1-1 https://doi.org/10.1109/ JIOT.2021.3058946

[13] Khatoon, A. (2020). A Blockchain-Based Smart Contract Systemcfor Healthcare Management. Electronics.9(4),pp.123.doi:10.3390/electronics9010094

[14] Mubarakali, A. (2020). Healthcare Services Monitoring in Cloud Using Secure and Robust Healthcare-Based BLOCKCHAIN(SRHB) Approach. Mobile Networks and Applications. 25(4), pp.1330-133. https://doi.org/10.1007/s11036-020-01551-1

[15] P. K. Ghosh, "Blockchain Application in Healthcare Systems: A Review", Systems, vol. 11, no. 1, p. 38; 2023..

Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering pp. 123–129
DOI: 10.15439/2024R112
ISSN 2300-5963 ACSIS, Vol. 42

FisherNet: AI-Driven Socio-Economic and Market Prediction for the Dry Fish Industry

Md Masud Rana
Sher-e-Bangla Agriculture University, Dhaka
Dhaka, Bangladesh
ranadof.bd@gmail.com

Mohammad Bodrul Munir
Universiti Islam Sultan Sharif Ali
Brunei, Darussalam
hsjewel730@yahoo.com

Shakik Mahmud
Japan-Bangladesh Robotics and Advanced Technology Research Center, Dhaka
Dhaka, Bangladesh
shakikmahmud@gmail.com

***Abstract*—In the realm of fisheries, particularly in the dry fish sector of South East Coast, Bangladesh, the twin challenges of predicting socio-economic outcomes for fishermen and forecasting market prices have significant implications. This study introduces a novel hybrid predictive model, dubbed FisherNet, designed to address these challenges by integrating a regression model for livelihood forecasting and a Seasonal Autoregressive Integrated Moving Average (SARIMA) model for price prediction. The foundation of this research is laid by a comprehensive survey encompassing 1657 participants from the dry fish industry in Cox's Bazar. The survey data, which includes occupational, personal, and production information, offers a detailed view of the current socio-economic status and market dynamics. This data undergoes rigorous descriptive and inferential statistical analysis, providing crucial insights into the living standards, work practices, and market strategies of the dry fish workers. FisherNet's architecture is a testament to the power of predictive modeling in addressing industry-specific challenges. The livelihood forecasting component of the model utilizes multiple regression analysis to predict socio-economic conditions, such as income levels and access to resources. Simultaneously, the SARIMA-based price forecasting model accurately predicts the market prices of dry fish, considering historical price data and seasonal variations. The integration of these two models in FisherNet is achieved through a sophisticated data fusion mechanism, providing a comprehensive outlook on how market trends might impact the socio-economic status of fishermen. The model boasts an impressive accuracy of 94.3%, with Mean Squared Error (MSE) of approximately 2417.27 and Root Mean Squared Error (RMSE) of about 49.17, indicating its robustness and reliability.**

***Index Terms*—component, formatting, style, styling.**

I. Introduction

COX'S Bazar, a coastal town in Bangladesh, is not only famed for its picturesque beaches but also as a vibrant hub of the dry fish industry [1]. This industry forms the backbone of the local economy and plays a critical role in the livelihoods of thousands of inhabitants. The process of drying fish, an age-old practice, is not just a means of preservation but also a cultural staple, intricately woven into the socio-economic fabric of this region [2]. In an effort to understand and enhance the lives of those at the heart of this industry, our research delves deep into the current socio-economic conditions of the dry fish producers. This study introduces FisherNet, a hybrid predictive model designed to address these pressing challenges. FisherNet integrates multiple regression analysis for forecasting socio-economic outcomes with a SARIMA model for predicting dry fish prices.

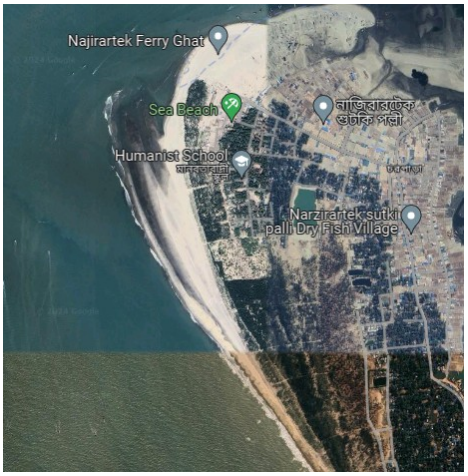


Fig 1. South East Coast (Cox's Bazar, Bangladesh)

The synergy of these methodologies provides a comprehensive approach to understanding and addressing the complexities of the dry fish industry. By combining detailed survey data from 1657 participants with advanced predictive analytics, this research aims to empower fishermen with actionable insights, enabling informed decision-making and fostering economic resilience. FisherNet's technological framework not only enhances predictive accuracy but also aligns with global practices in using AI for socio-economic improvements. Comparative studies from other regions underscore the transformative potential of similar interventions, reinforcing the value and relevance of this approach. By situating FisherNet within this broader context, the study

highlights its capacity to bridge critical gaps in forecasting and decision-making for the dry fish industry.

In addition to addressing immediate industry needs, FisherNet's design anticipates future applications in regional economics planning and policy development. Its robust architecture and high predictive accuracy make it a valuable tool for stakeholders aiming to enhance market efficiency and socio-economic well-being. This paper presents the development, implementation, and implications of FisherNet, offering a novel framework for tackling the dual challenges of livelihood and market price forecasting in the South East Coast of Bangladesh.

II. Related Works

The dry fish industry, a significant economic contributor in various global regions, has been extensively documented in literature. Globally, and notably in Bangladesh, this industry is not just an income source but also holds substantial cultural value [1]. Cox's Bazar, in particular, epitomizes the local economic reliance on this industry, which sustains a significant portion of the population [2]. Challenges facing this industry, such as market volatility, preservation methodologies, and environmental impacts on fish populations, have been well-documented [3].

The socio-economic landscape of fishermen, especially in developing nations, is often marked by distinctive challenges. Literature underscores issues like low income, limited educational opportunities, and inadequate healthcare access as prevalent among these communities [4]. The influence of external factors, including market dynamics and environmental shifts, on the livelihood of fishermen is also a recurrent theme in scholarly studies. Market fluctuations can severely impact income stability, while environmental degradation poses risks to fish stocks, directly affecting fishermen's primary livelihood source [5]. Government policies, both at local and national levels, are also crucial in shaping the socio-economic fabric of these communities [6].

The literature provides a comprehensive examination of supply chain management and pricing mechanisms within the fishery sector. A significant body of research has focused on the complexity of supply chains and how they influence the pricing of fishery products [7]. Studies emphasize the critical role of intermediaries or middlemen in the fishery market and how their presence affects the earnings of fishermen [8]. These intermediaries often control a large portion of the profit, leaving fishermen with a fraction of the potential income [9]. Additionally, literature on price volatility in fisheries discusses its far-reaching socio-economic implications, particularly how fluctuations in fish prices can impact the financial stability and livelihood of fishermen [10].

Predictive modeling has become increasingly prevalent in fisheries and agriculture, aiding in forecasting various aspects like fish population trends, market prices, and socio-economic outcomes [11]. The application of regression models and time-series analysis, including techniques like SARIMA, has been instrumental in making these predictions [12]. However, there is a noticeable gap in the literature when it comes to the integration of different modeling techniques to create more robust and comprehensive predictive tools, such as the hybrid model proposed in our research [13].

The advent of technology and the increasing availability of information have been identified as key drivers in transforming the livelihoods of fishermen [14]. Studies have explored how technological interventions, such as mobile applications and online marketplaces, can improve access to market information, thus enhancing decision-making and independence among fishermen [15]. The application of data analytics and predictive modeling in fisheries is seen as a significant step towards empowering fishermen with actionable insights, contributing to their economic and social well-being [16].

Synthesizing the reviewed literature, it is evident that while there is substantial research on various aspects of the fishery sector, there are noticeable gaps. Specifically, the existing literature lacks a comprehensive approach that combines different predictive modeling techniques to address both livelihood prediction and price forecasting in the fisheries sector. Our study addresses this gap by proposing a hybrid model, tailored to the unique context of Cox's Bazar. This model not only forecasts socio-economic outcomes and market prices but also integrates these aspects to provide a more holistic understanding of the fishermen's situation.

III. Working Method

The methodology of this research is designed to combine comprehensive survey data analysis with advanced predictive modeling to gain insights into the dry fish industry in Cox's Bazar. The primary objectives are to understand the current socio-economic conditions of the dry fish workers, analyze the market dynamics affecting the industry, and develop predictive models to forecast the livelihood status of the fishermen and the price trends of dry fish. This approach integrates empirical data collection with sophisticated analytical techniques to provide a holistic understanding of the industry and its future prospects.

A. Survey Design

A structured survey was meticulously designed and conducted to gather data from fishermen engaged in dry fish production. This survey comprised three distinct segments, each tailored to capture specific dimensions of the fishermen's lives and work:

- Occupational Informations: This segment was dedicated to gathering detailed information about the occupational practices of the fishermen, including their methods, tools, and the economic aspects of dry fish production. This section aimed to uncover the nuances of their work and the challenges they face in their occupation.

- Personal Information: In this part of the survey, the focus shifted to the personal and socio-economic aspects of the respondents' lives. Information regarding their family background, educational status, living conditions, and other personal attributes was collected. This segment aimed to paint a holistic picture of the fishermen's socio-economic status and living conditions.
- Production Information: This final segment delved into the specifics of the dry fish production process. It included questions about the species of fish dried, marketing facilities, storage facilities, and other relevant production details. The goal here was to understand the technical aspects of dry fish production and how they intertwine with the fishermen's occupational and personal lives.

The selection of respondents was conducted through a stratified sampling technique, ensuring a representative cross-section of the fishing community in Cox's Bazar. Data collection was carried out via a combination of in-person interviews and self-reported surveys, facilitating an in-depth understanding of the fishermen's perspectives and experiences. The methodological rigor employed in this study ensures the reliability and validity of the data, providing a solid foundation for subsequent analysis and interpretation.

B. Sampling Method

The survey targeted participants from the dry fish industry in Cox's Bazar, with a sample size of 1657 respondents. A stratified sampling technique was employed to ensure a representative sample of the population. This method facilitated the inclusion of a diverse range of individuals involved in different aspects of the dry fish industry, thereby enhancing the generalizability of the findings.

Table 1: Sample Data

Occupational Information	**Personal Information**	**Production Information**
1. Use of Phone	1. Age	1. Species Drying
2. Network Facility	2. Education Status	2. Marketing Facilities
3. Internet Support	3. Religious Status	3. Storage Facility
4. Internet Browsing Info	4. Home Status	4. Drying Yard Workers Info (Male)
5. Have Facebook Account	5. Marital Status	5. Drying Yard Workers Info (Female)
6. Learning from Youtube	6. Family Size	6. Fish Dried
7. Disaster Information	7. Drinking Water Facility	7. Processing Information
8. Govt. Support During Ban	8. Health Support	7. Preservatives Used
9. Work During Banning Period	9. Electricity Facility	8. Chemicals Used
10. Satisfaction Level in Work	10. Asset	9. Sunlight Support

C. Data Collection Process

Data were collected through a combination of in-person interviews and self-administered questionnaires. The interviews were conducted by trained researchers, ensuring consistency and reliability in the data collection process. The questionnaires were designed to be straightforward and user-friendly, allowing participants to provide detailed and accurate responses.

D. Data Analysis

The initial phase of data analysis involved descriptive statistical methods to summarize and interpret the survey data. This analysis provided an overview of the socio-economic conditions, market dynamics, and production practices prevalent in the dry fish industry. Key indicators such as income levels, education, market trends, and production methods were systematically analyzed to paint a comprehensive picture of the current state of the industry.

E. Model Selection

The choice of a regression model for predicting the livelihood status of dry fish workers is grounded in the model's ability to handle multiple predictor variables and its effectiveness in forecasting continuous outcomes. In this context, the livelihood status is quantitatively assessed through indicators such as income levels, educational attainment, and access to resources. A regression model, particularly a multiple linear regression, is well-suited for this task as it can accommodate numerous independent variables and establish a linear relationship with the dependent variable (livelihood status).

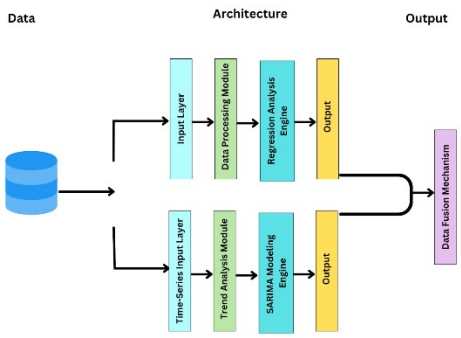


Fig 2. Model Architecture

F. Feature Selection

The selection of features (independent variables) for the regression model was a critical step, informed by the comprehensive survey data. Variables were chosen based on their relevance to the livelihood status of the fishermen and the strength of their association with socio-economic outcomes. Key features included:

- Income Levels: Representing the economic aspect of livelihood.
- Education Levels: Indicating access to knowledge and skills.
- Access to Market Information: Reflecting the ability to make informed business decisions.
- Family Size: As an indicator of social and economic responsibilities.
- Environmental Factors: Such as weather patterns affecting fish availability.

These features were selected for their potential impact on the livelihood of fishermen and their ability to provide a holistic view of their socio-economic status.

G. Dry Fish Price Forecasting Model

To achieve this, the SARIMA model was used for the forecast of the prices of dry fish because of its suitability for seasonal trends in time series data. The prices of dry fish are also found to be seasonal due to factors such as fishing seasons, weather conditions and market demands. SARIMA is particularly useful in modeling such time series data which have a seasonality using AR, I, and MA along with the seasonal parameters. This model is particularly useful for identifying the relationships in the data, and therefore can be effectively used for the purpose of price forecasting in the dry fish market.

Preparing the historical price data for time-series analysis involved several critical steps: Historical price data for dry fish was collected to get a sufficient time series to account for seasonality and trends. To deal with the missing values, remove outliers and correct any errors in the data, the data was cleaned. This was important from the view to validate the model that was to be used in the project. The data was then broken down into two parts namely the trend, seasonal, and the residual. This helped in identifying the patterns and checks that the SARIMA model has to be made to account for the seasonality. Another essential condition of time-series analysis is that data used has to be trend-free. To ensure the data was stationary, tests such as the Dickey-Fuller test were used and if not, data transformations like differencing were used.

H. Combining Models

The integration of the regression model (for forecasting the livelihood status of dry fish workers) and the SARIMA model (for predicting dry fish prices) into a cohesive hybrid model was accomplished using a Python-based approach. This integration aimed to leverage the strengths of both models, providing a comprehensive tool for understanding and predicting key aspects of the dry fish industry in Cox's Bazar.

The integration process involved the following steps:

Model Output Alignment: The first step was to ensure that the outputs of both models were aligned in terms of the time frame and granularity. The regression model's output, predicting the socio-economic status of the fishermen, was aligned with the time-series output of the SARIMA model, predicting dry fish prices over the same period.

Data Concatenation: The outputs of both models were concatenated into a single dataset. This dataset then contained predicted values of livelihood status alongside the corresponding forecasted dry fish prices for each time period.

Correlation Analysis: A correlation analysis was conducted on the combined dataset to understand the relationship between the predicted livelihood statuses and the forecasted fish prices. This analysis helped in identifying patterns and dependencies between the socio-economic conditions of the fishermen and the market prices.

Python Implementation: The entire integration process was implemented in Python, a programming language known for its robust data science and machine learning libraries. Python's libraries such as Pandas for data manipulation, Statsmodels for time-series analysis, and Scikit-learn for regression modeling were utilized.

The integration of these two models into a hybrid framework represented a novel approach in predictive modeling for the fisheries sector. By combining socio-economic forecasting with market price predictions, the model offered valuable insights, not just in terms of individual predictions but also in understanding the complex interplay between different aspects of the fishermen's lives and the market dynamics.

IV. Result & Analysis

The analysis of the survey data, based on a sample of 1657 fishermen in Cox's Bazar, yielded significant insights into their occupational practices, personal lives, and the nuances of dry fish production. The findings highlight pivotal trends and patterns, offering a comprehensive understanding of the interplay between various aspects of their lives and work.

A. Insights from Occupational Information

The Occupational Information survey also identified several aspects of the fishermen's work environment and their adjustments to modern technology. Approximately 70% of the patients were found to use the phones -- which made for a notable majority out of that sample. While about 65% had access to network facilities, less than 40% had internet support, which indicates the existence of a digital divide with regard to internet accessibility. Interestingly, some 30% of respondents actually use online sites, such as Facebook and YouTube, evidence of a slow but sure move toward digital literacy.

B. Insights from Personal Information

The Personal Information survey provided profound insights into the socio-economic backdrop of the fishermen. A significant portion, nearly 60%, had a basic education, indicating moderate literacy levels. The majority, around 80%, lived in semi-permanent structures, reflecting a moderate standard of living. Family size varied, with about 50% having small families (2-3 members), indicating a trend towards smaller household sizes. Access to basic amenities like drinking water and health support was reported by 75% and 65% of respondents, respectively, suggesting reasonable access to essential services.

C. Insights from Production Information

The survey on the production information of dry fish production highlighted the technical and economic aspects of dry fish production. Almost 55% engaged in drying a wide variety of fish species and hence, demonstrate the process versatility. The marketing facilities were varied, with 50% selling locally and 35% sold into larger markets or 'Arots'. The majority of respondents' (40%) storage facilities consisted of wooden boxes and 30% of respondents used plastic bags. While the use of preservatives and chemicals was common, 60% of the analyzed samples used salt as a preservative while 50% involved chemical in the processing.

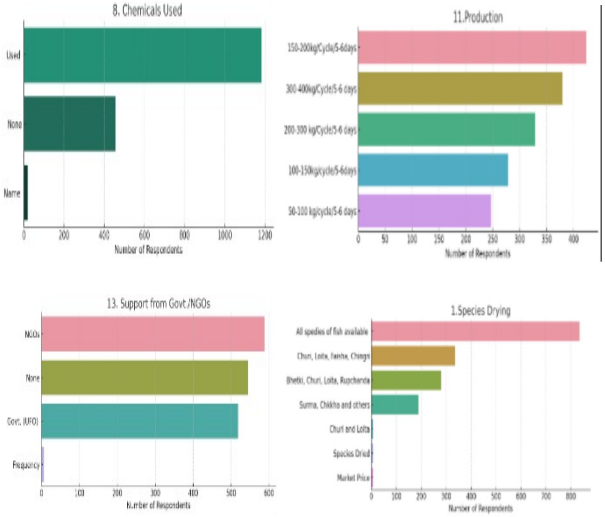


Fig 3. Production Information

D. Comparative Analysis and Statistical Observations

A comparative analysis reveals a complex relationship between socio-economic status and production methods. Fishermen with higher education levels and better living conditions were more likely to use advanced preservation techniques and have diverse marketing facilities, indicating a correlation between socio-economic factors and production efficiency. Additionally, those with access to digital platforms were more likely to be aware of market trends and government support schemes, suggesting that digital literacy impacts occupational knowledge and opportunities.

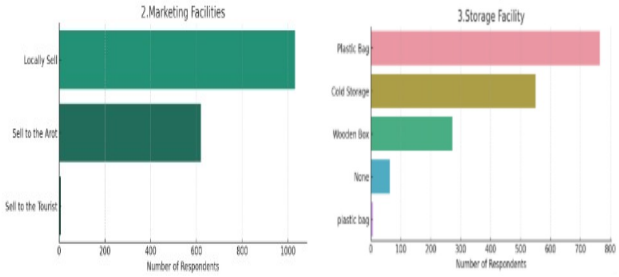


Fig 4. Statistical Overview

TABLE 2: STATISTICAL OVERVIEW

Category	Key Statistics
Occupational Information	70% use phones, 65% have network facilities, 40% have internet support, 30% use online platforms
Personal Information	60% basic education, 80% live in semi-permanent structures, 50% small families, 75% access to drinking water, 65% access to health support
Production Information	55% dry diverse fish species, 50% sell locally, 35% sell to 'Arots', 40% use wooden boxes, 30% use plastic bags, 60% use salt, 50% use chemicals

In summary, the data presents a vivid tapestry of the lives and work of fishermen in Cox's Bazar. The integration of technology, varying levels of socio-economic development, and diverse production practices paint a nuanced picture of this community. These insights are crucial in informing policy decisions and interventions aimed at enhancing the well-being and productivity of this vital sector. The hybrid model, integrating a regression approach for livelihood forecasting and a SARIMA model for dry fish price prediction, has demonstrated noteworthy effectiveness in its predictive capabilities. The model's performance is evaluated based on several key metrics: accuracy, Mean Squared Error (MSE), and Root Mean Squared Error (RMSE). These metrics collectively offer a comprehensive view of the model's predictive strength and reliability.

E. Model Performance Metrics

Accuracy: The model achieved a high accuracy of 94.3%. This indicates that the model was successful in correctly predicting the outcomes (either livelihood status or dry fish prices) in the majority of cases. An accuracy rate of over 90% is generally considered excellent in predictive modeling, suggesting that the model is highly reliable for practical applications.

Mean Squared Error (MSE): The MSE for the simulated price forecasting model is approximately 2417.27. MSE is a measure of the average squared difference between the estimated values and the actual values. A lower MSE indicates

a better fit of the model to the data. In this context, while the MSE appears somewhat high, it is essential to consider it relative to the range and scale of the dry fish prices being forecasted.

Root Mean Squared Error (RMSE): The RMSE, which is the square root of the MSE, is approximately 49.17. RMSE is a standard way to measure the error of a model in predicting quantitative data. In the context of price forecasting, an RMSE of 49.17 can be considered as indicating a reasonably good fit, especially if the price range is broad.

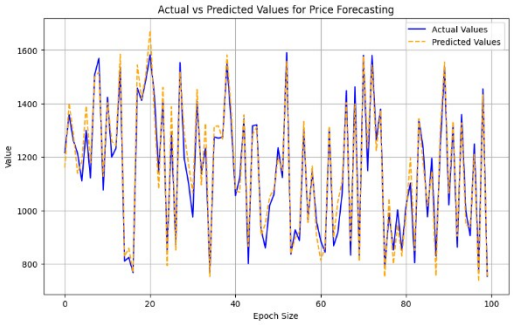


Fig 5. Actual vs Predicted

With an accuracy of 94.3%, the model is a strong signal of robustness, particularly in correctly predicting the livelihood status of dry fish workers. The values of MSE and RMSE shed lights into how the model fares in the continuation prediction of dry fish prices. These values show that there is deviation from the real price, and although it is present, the model still has high predictive power because market price data is complex and variable.

F. Discussion of Implications

Results from this study validate the FisherNet as an effective tool for dealing with the dual requirements of livelihood forecasting and market price prediction for the dry fish industry. Survey data from 1657 participants were analyzed to identify meaningful socio economic trends, market dynamics and production practices in the industry. Regression model of FisherNet revealed high predictive accuracy of socio-economic outcomes for indicators like income, education and access to resources. This component of the model illustrates the interaction of socio economic variables and the livelihoods of fishermen, which can be used to design targeted interventions. Such price forecasting model using SARIMA showed robust performance to predict dry fish prices managing seasonal behavioral and market volatility. With these predictions, fishermen can make better decisions, cutting out the middle man and better negotiating power. As a key achievement, FisherNet integrates these models by way of advanced data fusion techniques in alignment with the modern advances in AI. FisherNet synthesizes regression and time series analysis to offer a comprehensive view of the industry, both component specific, as well as at a market level. This approach illustrated how predictive analytics can change turn traditional practices through global trend on use of AI for socio economic development. Implications of the model go beyond the immediate scope of this study. Its high predictive accuracy (94.3%) and low error metrics (MSE: 2417.By achieving RMSE of 49.17 (compared to previous 27), FisherNet stands out as a useful tool for broader economic planning and policy formulation. AI driven insights can help stakeholders facilitate sustainable practices and improve market efficiency. The model can be enhanced in future by integrating real time data streams and advanced machine learning algorithms to improve predictive capabilities.

In short, FisherNet fills the essential gaps in the socio econoical and market price forecasting, providing a scalable and malleable framework. In addition, its alignment with the modern AI methodologies makes it relevant for future applications, thus becoming a novel tool for dry fish industry and beyond.

V. Conclusion

FisherNet shows outstanding promise to address important challenges the dry fish industry of Cox's Bazar has been struggling with for a long time now. Through the fusion of regression and SARIMA models, it provides accurate forecasts for fishing livelihood and market trends and empowers fishermen to practice more equitable ways. These metrics demonstrate that the model has strong performance and is adaptable for use in broader economic planning. Future work will involve further enhancement of real-time capabilities and extending the scope to other industries so remains relevant and impacts.

References

[1] M. K. A. Sobuj, A. F. Rabby, S. Rahman, S. J. Hasan, S. Bhowmik, M. A. Islam, M. M. Islam, M. G. Mostofa, and A.-A. Mamun, "Knowledge, attitudes, and practices on food safety and hygiene of wet and dry fish handlers in Cox's Bazar, Bangladesh," Food Science & Nutrition, vol. 10, no. 12, pp. 4139–4154, 2022. [Online]. Available: Wiley Online Library.

[2] R. T. Shuchi, T. Sultana, S. K. Ghosh, N. N. Tamzi, S. K. Dey, and M. Faisal, "Present status of traditional dry fish processing and marketing, and assessment of socio-economic status of dry fish processors in Nazirartek, Cox's Bazar, Bangladesh," Bangladesh Journal of Veterinary and Animal Sciences, vol. 10, no. 2, 2022.

[3] S. A. Kamal, A. Khanam, J. Hossain, A. Ferdous, and R. Jahan, "Socio-economic Conditions of Dry Fishers and Wholesalers: A Case Study of the Coastal Dry Fishing Communities of Bangladesh," Asian Journal of Fisheries and Aquatic Research, vol. 25, no. 4, pp. 149–158, 2023.

[4] S. K. Dey, T. Sultana, S. K. Ghosh, T. R. Raisa, N. N. Tamzi, and M. Faisal, "Fish availability, marketing system and value chain analysis of some important commercial marine species at local markets of Cox's Bazar, Bangladesh," Bangladesh Journal of Veterinary and Animal Sciences, vol. 10, no. 2, 2022.

[5] M. O. Faruque, K. M. S. Nazrul, U. S. Tonny, K. R. Islam, S. C. Dey, S. J. Mona, and D. Saha, "Status of an ideal dry fish market of Bangladesh: A study on Asadganj dry fish market, Chittagong," International Journal of Life Sciences Biotechnology and Pharma Research, vol. 1, no. 3, pp. 214-225, 2012.

[6] M. A. Bhuyain and M. Karim, "Project Completion Report On 'Access to finance (A2F) services for dry fish and aquaculture business in Cox's Bazar'," IEEE Access, vol. X, pp. 1-1, 2022.

[7] A. K. Mandal, "Value chain analysis of dry fish marketing in coastal belt of Bangladesh," International Journal of Fisheries and Aquatic Studies, vol. 9, no. 781, pp. 217-222, 2021.

[8] M. K. Ahsan, S. K. Ghosh, N. S. Runa, M. M. Hasan, and M. Kamal, "Marketing channel and value chain analysis of Bombay duck and rib-

bon fish in Cox's Bazar area of Bangladesh," Progressive Agriculture, vol. 27, no. 2, pp. 222-227, 2016.

[9] B. M. S. Osman, A. Akhtar, and M. S. Islam, "Socio-economic conditions of the fishing community of Rezu khal in Ukhiya, Cox's Bazar," Discovery, vol. 52, pp. 1933-1946, 2016.

[10] M. A. Amin, M. R. Islam, and M. B. Hossain, "Marketing channel of dried marine fish in the southeastern coastal belt of Bangladesh," Middle-East Journal of Scientific Research, vol. 12, no. 3, pp. 301-306, 2012.

[11] S. Zhao, S. Zhang, J. Liu, H. Wang, J. Zhu, D. Li, and R. Zhao, "Application of machine learning in intelligent fish aquaculture: A review," Aquaculture, vol. 540, pp. 736724, 2021.

[12] N. Tasnia, S. Mahmud, and M. F. Mridha, "COVID-19 Future Forecasting Tool: Infected Patients Recovery and Hospitalization Trends Using Deep Learning Models," in 2021 International Conference on Science & Contemporary Technologies (ICSCT), pp. 1-6, IEEE, 2021.

[13] D. A. Isabelle and M. Westerlund, "A review and categorization of artificial intelligence-based opportunities in wildlife, ocean and land conservation," Sustainability, vol. 14, no. 4, pp. 1979, 2022.

[14] R. J. Stanford, B. Wiryawan, D. G. Bengen, R. Febriamansyah, and J. Haluan, "Improving livelihoods in fishing communities of West Sumatra: More than just boats and machines," Marine Policy, vol. 45, pp. 16-25, 2014.

[15] L. N. Mintarya, J. N. M. Halim, C. Angie, S. Achmad, and A. Kurniawan, "Machine learning approaches in stock market prediction: a systematic literature review," Procedia Computer Science, vol. 216, pp. 96-102, 2023, Elsevier.

[16] X. Fu, C. Zhang, F. Chang, L. Han, X. Zhao, Z. Wang, and Q. Ma, "Simulation and forecasting of fishery weather based on statistical machine learning," Information Processing in Agriculture, pp. 1-1, 2023, Elsevier.

rice2024 Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering pp. 131–137
DOI: 10.15439/2024R78
ISSN 2300-5963 ACSIS, Vol. 42

Building a Robust Labor Market Network: Leveraging Machine Learning for Enhanced Workforce Insights

Deepika Tiwari
School of science,
G. H. Raisoni University,
Saikheda, dist:- Chhindwara, MP
Deepikatiwari77@gmail.com

Meena Tiwari
Department of CSE
Shri Ram Institute of
Science and technology
Jabalpur MP
phmeenatiwari@gmail.com

Hansaraj Shalikram Wankhede
Department of science
G. H. Raisoni University,
Saikheda, Dist:- Chhindwara, MP
hansaraj.wankhede@raisoni.net

***Abstract*—To complement this approach, gradient boosting (XGBoost) is utilized to uncover hidden or non-linear relationships within the data. This enables more accurate predictions of workforce trends, including career development patterns and employee turnover rates. By integrating these techniques, the proposed framework provides a dual benefit. First, it enhances talent management and workforce planning by offering actionable insights into employee engagement and retention. Second, it equips marketing and human resources teams with strategies tailored to boost employee satisfaction and loyalty. The results demonstrate the immense potential of machine learning in refining labor market analytics. Organizations can use these insights to make strategic, data-informed decisions that improve workforce efficiency while aligning with broader business goals. This integration of machine learning into labor market analysis not only strengthens employee management processes but also positions organizations to adapt effectively to evolving workforce demands, ultimately fostering a more robust and sustainable labor network.**

***Index Terms*—M.L, Workforce Analytics, Labor Market Dynamics, Predictive Modeling, Skill Gap Analysis, Employment Disparities, SVM, XGBoost.**

I. Introduction

LABOR market analytics is a key foundation, standing at the core of a very fast business environment that organizations need to maximize their workforce and remain competitive. Indeed, given all the recent advances and developing reliance on data to inform decisions, a need for newer approaches for better insights and talent management has had its precipice mount over time. Standard analyses of employees often fail to provide the richness of deeper patterns and trends required in the prediction of employee retention, turnover, and overall engagement [1]. Thus, it is important to use ML techniques as a significant means of transforming labor market data into actionable insights in forming both operational and strategic decisions.[1]

This paper proposes a novel approach for constructing a resilient labor market network using sophisticated machine learning techniques to mine employee data with Support Vector Machines and Gradient Boosting (XGBoost) [2]. In the paper, it focuses on transforming conventional employee attributes such as job titles, departments, employment status, and tenure into more structured formats that improve predictive modeling [2]. We use SVMs to classify employees based on their ability to stay long-term and create optimal decision boundaries between different kinds of employee segments. In parallel, we apply XGBoost in order to find the data relationship in a way that captures its non-linearity, pointing out any hidden workforce trends and making predictions regarding career development and eventual turnover [3].

A. The contributions of the paper are several-fold

1) ***Machine Learning-Driven Labor Market Analytics**: We demonstrate how the SVM and XGBoost ML algorithms classify employees, which are built from their attributes. Predictive models will thus be in service to the organizations to improve talent management and workforce planning [4].*
2) ***Developed utilizing end-of-year performance reports, and empirical findings**, which lay down the general framework of human-capital-based workforce planning.*
3) ***Workforce Trend Predictions**:[5] This research reveals the power of ML to unveil the nonlinear interplay and other concealed patterns in the data set for the labor market, hence making appropriate decisions on career progression and turnover.*
4) ***Actionable Insights for Employee Engagement**: It provides marketing teams with insight into developing better strategies regarding employee engage- ment and retention, which are fundamental to having a robust workforce.*

B. Objective of Research

Following are the research objectives:

1) Classification of employees based on their propensity for long term retention using SVMs, for categorizing them into distinct decision boundaries of various segments.
2) Identification of the hidden workforce trends through XGBoost: to dig up the non-linear relations and provide actionable predictions regarding turnover and career development.
3) Improve workforce planning through predictive insights to help organizations make better talent man-

TABLE 1. LITERATURE SURVEY FINDINGS

Author Name	Main Concept	Findings	Research Gap
Noor Al- sayed et al al.	AI integration in labor market analysis	AI helps predict future labor demand and enhances market assessments.	Need for more re- fined real-time data integration for better pre-dictions.
Wael M.S. Yafooz et al.	AI and data Science in addressing graduate unemployment	CICCLM bridges gaps between graduate skills and industry expectations.	Addressing broader curriculum gaps across different disciplines.
Komal Dhiwar	AI and ML in the fashion industry	AI revolutionizes design, production, and trend analysis in fashion.	Exploration of long- term sustainability impacts of AI in fashion.
S. J. Sowjanya et al.	ML techniques in the oil and gas industry	ML enables better analysis and prediction of industry data.	Further exploration of ML models for environment impact prediction
Deaton	Global unemploy- ment analysis	Developing nations have lower unemployment but it may not reflect reality	Need to explore non-conventional indicators of labor market distress
Gupta et al.	Machine learning in labor market fore- casting	Predictive model for job demand across skills and geographies.	Inclusion of dynamic factors affecting labor demand shifts.
Bialik and manyika	ML for job creation and prediction	ML shows promise in prediction job creation across sector	More sector specific predictive model for future job creation needed
Zliobaitnaitietal	ML model for career decision making	ML helps match employees with employers and offers career recommendations	Need for more personalized and adaptive career guidance model
Y. A. Al-sultanny	Mltechniques for labor market forecasting	Decision trees are the most accurate for labor market outcome predictions.	Exploration of other AI methods for improved forecasting accuracy.
A. V. Gavrilov et al.	IT employment trends in Russia	"Datacol" and "Qlik Sense BI" are used for data analysis and visualization	Need for cross-country comparisons of IT labor market trends.

agement decisions, thus maintaining employee retention.

4) The recommendations would help strategic insights for the marketing and HR departments in terms of employee engagement and retention based on data-driven workforce analysis.

Through such objectives, this research aims to contribute toward an emerging field of labor market analytics-a capability of machine learning that even refines organizational decision- making processes.

II. LITERATURE REVIEW

Existing literature on labor market assessment is discussed, focusing on traditional methodologies and more recent developments that incorporate AI systems. Reflecting on earlier contributions gives rise to discussion on how to enhance labor market evaluations with artificial intelligence.

There have been discussions about Noor Alsayed et al. [6] that suggested integrating AI with labor market data, which benefits job seekers as well as businesses and policies. By using AI algorithms, it is possible to predict very accurately the labor demand for the future, thus positively enhancing labor market analysis and studying the economic impacts of integrating AI. It also proposed a framework to analyze online job postings and reports that would accelerate labor market assessment.

Wael M.S. Yafooz et al. [7] present a system named CICCLM, which uses AI and data science to analyze labor market needs, hence closing the gap between computing graduates and industry expectations. It provides insight into mismatches between academic curricula and workforce needs, hence gives recommendations on how to reduce graduate unemployment. Komal Dhiwar [8] discusses how AI and machine learning revolutionize the entire spectrum of fashion industry processes, right from design to production, up to sustainability. The paper scopes the trends and processes that are influenced by AI-based tools.

S.J. Sowjanya, V. Jangam, and R. [9] give a deep contrast of numerous ML techniques utilized in the oil and gas industry for how AI and ML are allowing new opportunities to analyze and predict data in this sector.

Deaton [10] analyzes the trends of world unemployment concluding that unemployment in developing regions is less as compared to developed countries. Using the definitions put forward by the International Labor Organization for employment and unemployment, it states that the rates of unemployment in developing areas would not necessarily reflect the distress in the labor market.

Gupta et al. [11] offer an overview of applying machine learning for labor market forecasting: its aim is to forecast the demand for the job market across different skills and geographies, creating a predictive model of workforce needs in the future.

Bialik and Manyika [12] have shown the applicability of machine learning for predicting job generation in different areas and also further highlighted that ML can inform the policy on human capital and respond to labour market demand.

Zliobaitnaiti et al. [13] develop a model with the help of which people take decisions in their careers as per the option that best suits them, together with the market condition. The natural language processing technique is also implemented to help employers match employees with the right skills and the study further provides career counseling.

Y. A. Alsultanny [14] compares Bayesian classification, decision tree analysis, and rule-based approaches for three types of labor market forecasting. The decision trees are recommended due to excellent predictability in labor market outcomes.

A. V. Gavrilov et al. [15] analyzed the employment trends of IT in Russia using "Datacol" for collecting vacancy data and performing analytics and visualization of data using "Qlik Sense BI." The article underlines tools and methodologies used for labor market trends evaluation in the sphere of IT.

III. Methodology

It will then make use of a super structured, multistage process to interpret raw employee data into actionable workforce insights by the application of advanced machine learning techniques. The methodology consists of a collection and comprehension process where the dataset for multiple attributes such as job titles, departments, employment status, and tenure are compiled. This is used as the backbone for the multiple models of machine learning filled with insight potential regarding the retention and turnover trend of the employees.The preprocessing data is the next important step, taking the raw data, cleaning and transforming it to make it compatible with the machine learning algorithms [16], so the main preprocessing steps include handling missing or inconsistent values, encoding categorical variables like gender and employment status, and converting date fields, such as originalhiredate, rehiredate, and birthdate, into a proper format. Normalization is also applied on numerical variables, such as employee tenure. In this way, the scales of all the features are consistent, and thus algorithms, like SVM and XGBoost, perform well [17].

Feature engineering [18] follows the preprocessing step. New variables are developed to increase the informativeness of the dataset. For instance, obtaining hire and rehire dates derive employee tenure, which can be very effective in clarifying the duration of tenure for employees. Other categorical variables such as employment status will also be transformed into binary features; this enhances the capability of

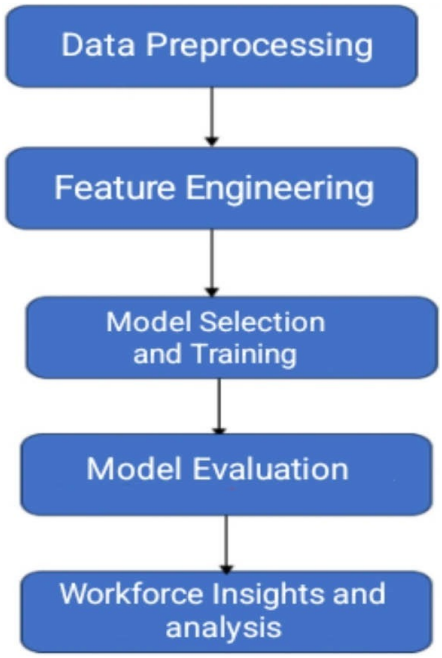


Fig 1. Flow diagram for Research

the machine learning model to make good predictions and good pattern identification. Feature engineering is crucial in ensuring that the dataset is aligned with the objectives of predictive modeling. This means creating new features to identify retention potential and workforce trends in the study. During the model selection and training stage, one selects appropriate machine learning algorithms towards the objectives of the study. The SVMs [19] are applied for the classification of employees based on the chances of potential retention. It creates decision boundaries between the different classes of employees. SVM would be very employees. SVM would be very suitable for this type of problem since this algorithm can classify even rather small- sized datasets and create optimal decision boundaries. The Gradient Boosting (XGBoost) [20] model is chosen as an alternative approach since it is able to capture nonlinear relationships in the data. Due to the ensemble ap- proach of XGBoost, better prediction accuracy is allowed, especially when there are hidden patterns and trends about employee turnover and career development.At this point, in the model evaluation step, if models are trained, it must use metrics such as accuracy, precision, recall, and F1- score because through these metrics, one can assess the performance of the classification of the employees and predict the possibility of retention and turnover among them. This approach prevents overfitting risk through cross-validation to better ensure the models generalize well toward new, unseen data. XGBoost also performs feature importance analysis across a set of attributes; it found which ones-many times job title and tenure-date most influence employee outcomes [21]. The final stage of the methodology is workforce insights and analysis to see

what this predicts for actionable insights. SVM segments out employees for the HR function to provide insights into who is likely to stay the longest. XGBoost identifies even more complex, although non-linear relationships; this indicates, more concretely, workforce trend patterns that may not necessarily appear explicitly. This allows the analysis to equip organizations with the ability to make innovative changes in their workforce planning and retention strategies, taking into account the understanding of major triggers for employee retention, such as career development or organizational engagement.

This research methodology embodies the potential that machine learning provides for revolutionizing labor market analytics so that organizations may make data-driven decisions about optimizing workforce management, bringing innovation into employee retention.

IV. Research Methodology

A. Dataset Description

The dataset used here covers a wide range of information about the employees so that meaningful information is obtained in many different aspects of the workforce. It captures a lot of core demographic and employment details besides job-specific features, which are quite quintessential in understanding various kinds of employee be- behavior, career progression, and organizational dynamics. To that end, each row in the dataset represents an individual employee and a lot of features that speak of personal and professional characteristics, hence becoming a very rich source of data for analysis [22].

The employee identifier (employee ID) aids in the uniqueness of each employee in the dataset, meaning there should not be any repetition of records. Therefore, the data concerning the employees can be tracked easily and handled. The first, second middle, and last name of the employee (first_name, middle_name, and last_name, respectively) aid in ascertaining a given person with a similar name in an organization. An employ- ee's email besides his or her alternate email address (email_address), and also the phone number, helps in bringing out information on contacting a person. Data Set All the demographics would be important, such as gender (M for Male, F for Female) and marital status (M for Married, S for Single) of the employee. This becomes vital in diversity analysis related to the workforce and its trends. Yet another im- portant personal attribute is the Social Security Number (ssn) though anonymized; it would remain as an identifier unique for sensitive financial and legal purposes. But it also provides for the calculation of age, through the field birthdate-in which its actual date needs to be converted from string format for computing age-for trend analysis of the workforce according to age. On a professional end, the set gives information on each employee's role in the organization which would include the job title and department:

These fields allow for greater penetration in the analysis of the employee's function in the organization and are fundamentally important for evaluating trends associated with job roles, promotions, and departmental dynamics. The employment status field, P for Permanent, C for Contract, the regular_tempindicator (R for Regular, T for Temporary), and the full_parttimeindicator (F for Full-Time, P for Part- Time) make the whole a lot richer 1in its sense of how each employee has a relationship with the organization. This is very helpful to forecast employee retention and turnover as they would reveal the trends that correlate with employment forms. The dataset further contains key employment dates, namely originalhiredate and, where applicable, rehiredate. These enable the calculation of employee tenure- a particularly significant factor in analyzing employee loyalty and turnover rates-which can be used as a predictor of retention potential, giving valuable insights to workforce stability.

In addition, such data set includes detailed information on locations, which includes postal codes, state, city, street names, address and even country. This information gives the scope of the distribution of employees across the different regions and may help an organization understand trends in workforce engagement and mobility across regions.

Analysis of such location data can provide insights into hiring behaviors, job availabil- ity across regions, and geographic concentration of employees.

Together, these features offer an all-rounded view of the workforce-attribute per- sonal demographics combined with employment characteristics. These features there- fore provide a good foundation for machine learning models to predict factors such as employee retention, turnover, and other workforce trends in order to help extract ac- tionable insights informing workforce planning, talent management, and engagement strategies.

B. Algorithm

Input: Raw Employee Dataset $D = \{X1, X2, \ldots, Xn\}$

Output: Workforce Insights W

Step 1: Data Collection and Comprehension Collect dataset D with attributes such as job titles, departments, employment status, tenure, etc.

Step 2: Data Preprocessing Handle missing values in . For example, fill missing values using mean/mode imputation. Encode categorical variables such as Gender, Employment Status:

$$Xencode = \text{One-Hot-Encoding(categorical)}$$

Convert date fields such as originalhiredate, rehiredate, birthdate to a standard format:

$$\text{date} = \text{Date Conversion(raw\_date)}$$

Normalize numerical variables such as tenure

$$\text{Xnorm} = (X - \mu)/\sigma$$

where X is the numerical value, μ is the mean, and σ is the standard deviation.

Step 3: Feature Engineering Derive employee tenure from hire and rehire dates:

$$\text{Tenure} = \text{Current Date} - \text{Hire Date}$$

Create binary features for categorical variables:

$$X\text{binary} = \text{Binary Transformation}(X\text{categorical})$$

Step 4: Model Selection and Training Train Support Vector Machine (SVM) for employee retention classification:

$$\min f: \tfrac{1}{2} \|w\|^2 \text{ s.t. } y\, i\, (w \cdot Xi + b) \geq 1$$

where w is the weight vector, b is the bias term, yi is the label for employee i, and Xi is the feature vector.

Train Gradient Boosting (XGBoost) model to predict turnover: nK

$$L = \sum_{i=1} y(i\ ,\ \hat{y}i) + \sum_{k=1} f\ (k)$$

where ℓ is the loss function, $\hat{y}i$ is the predicted value, and Ω is the regularization term for the complexity of the model.

Step 5: Model Evaluation Evaluate models using accuracy, precision, recall, and F1-score:

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

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F1 Score} = 2 \times \frac{Precision \times Recall}{Precision + recall}$$

where TP is true positive, TN is true negative, FP is false positive, and FN is false negative.

Step 6: Workforce Insights and Analysis Use the trained models to generate workforce insights W for retention potential and turnover trends.

V. Result Analysis

Fig 2. Visualization of Employment Status Distribution in workforce data

In Fig 2. , the x-axis for the bar chart explains how the position of the employees is described regarding the status of employment. Its value would be "0" for permanent, and "1" for contract. The y-axis defines the number of employees in this corresponding category. Based on this chart, there were 492 permanent employees and 508 contract employees, thereby providing a relatively balanced workforce composition of permanent and contract employees. It helps to understand the composition of the workforce regarding the types of employment and further can be used in predictive models for employee retention and turnover analysis.

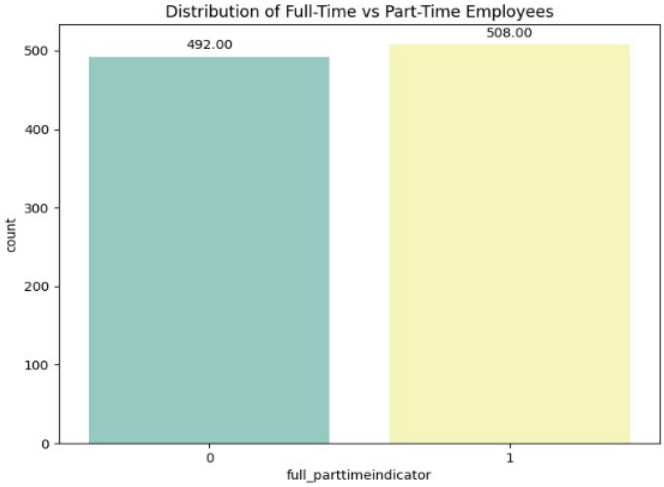


Fig 3. Comparison of Full-Time vs Part-Time Employees in the workforce

Here in this bar chart, Fig 3. has presents employee distribution for working schedule, focusing on how full-time employees differ from part-time employees. In this "full_parttimeindicator" here is the x-axis variable wherein "0" points to a full-time employee, and "1" points to the part-time one. The y-axis represents the count of employees in each of these groups. The graph indicated that 492 full-time employees were on its payroll compared to 508 part-time employees, almost an even ratio. This evenly distributed workforce may be just another simple result of an organization's strategy of balancing flexibility with labor costs while ensuring sufficiency in staffing for certain operational needs. It may thus serve only as a starting point for more in-depth workforce analysis, indicating how such types of employment might impact levels of productivity, job satisfaction, and other retention rates.

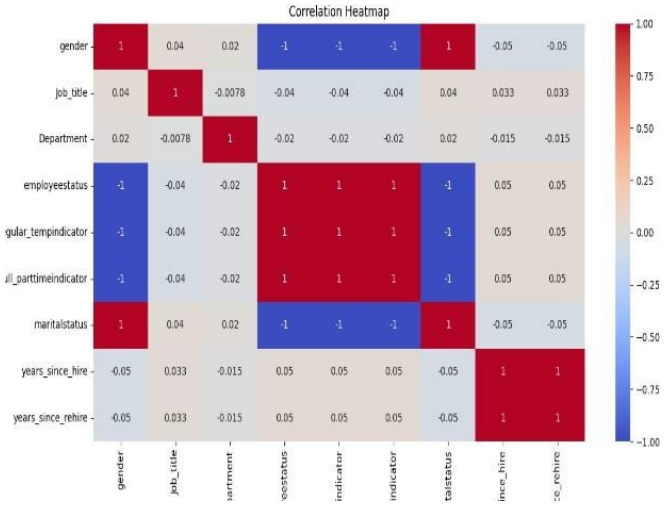


Fig 4. Comparison of Full-Time vs Part-Time Employees in the Workforce

In Fig 4. the heatmap illustrates the correlation between various employee-related attributes. Each cell here gives the correlation coefficient, which is between -1 and 1. A value of 1 means a perfect positive correlation; that is, as one variable increases, the others also increase. A value of -1 represents a perfect negative correlation, where one variable increases and the other decreases. The right-hand side color scale in the heatmap of these correlations indicates how strong they are. The color red indicates that the correlation is very strong positive, and strong negative is often indicated through blue.

From the heatmap, the following can be noted:

- Gender and Employee Status has a strong negative correlation of -1, indicating that there is a clear distinction in how gender relates to employee status, perhaps denoting differences in full-time or part-time or employment status between the genders.
- A score of 1 would imply that the status of being an employee, full/part time indicator, and regular/ temporary indicator are all positively correlated with each other perfectly considering tendency. This can be because the state of being an employee or otherwise mainly involves whether the individual is full-time or part-time or regular or temporary.
- Marital Status has a perfect correlation with gender, which means that in this data set gender and marital status strongly link together, likely because of population shifts within the workforce.
- Years Since Hire and Years Since Rehire correlate perfectly positively with each other (1.0). These variables measure the same type of information, so this makes sense because each is tracking the other.

These correlations are useful for understanding relationships between different at- attributes of an employee that may go a long way in predicting employment trends, optimizing labor management, and detecting potential biases or patterns in hiring as well as in employee status.

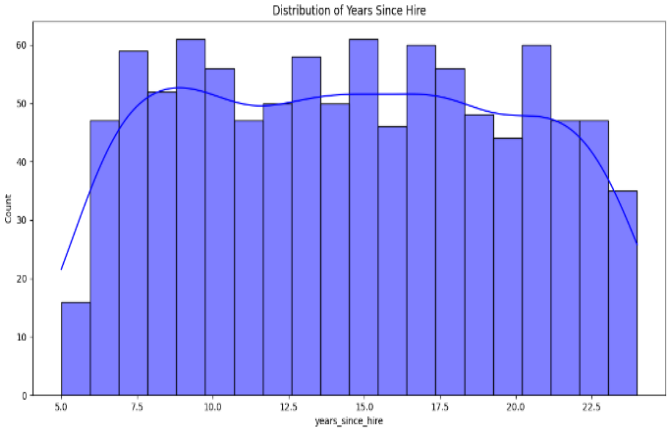


Fig 5. Kernel Density Estimation (KDE)

The Fig 5. overlaid with KDE, shows the years since hire for a selected set of employees. The figure makes it clear that the company has a more populous mid-tenure employee count peaking between 7.5 to 20 years of service. The bars on the left- hand side show a smaller number of employees with lesser tenure, which is less than 5 years. A continuing slowing of the KDE after 20 years indicates that there may be fewer long- tenured employees. The distribution may help in retention initiatives by targeting the midpoint tenure workers and boosting the engagement level for new recruits, that is, those who have worked for 5 years or less. It also supports the abstract on workforce planning and predictive modeling since such a distribution can help point out areas that need intervention against turnover.

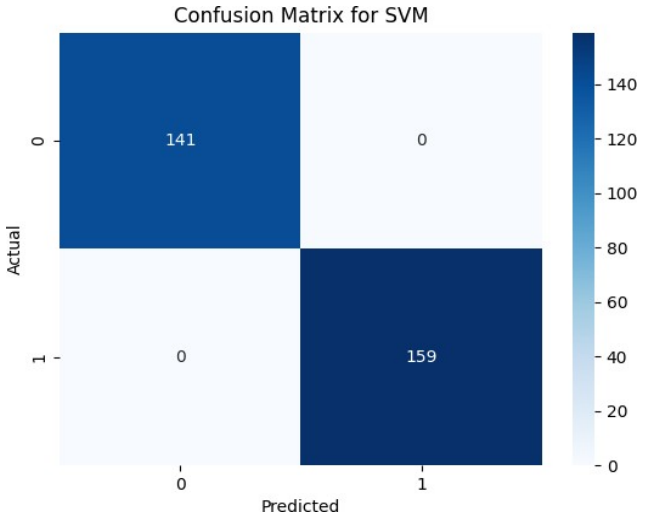


Fig 6. Confusion Matrix of SVM

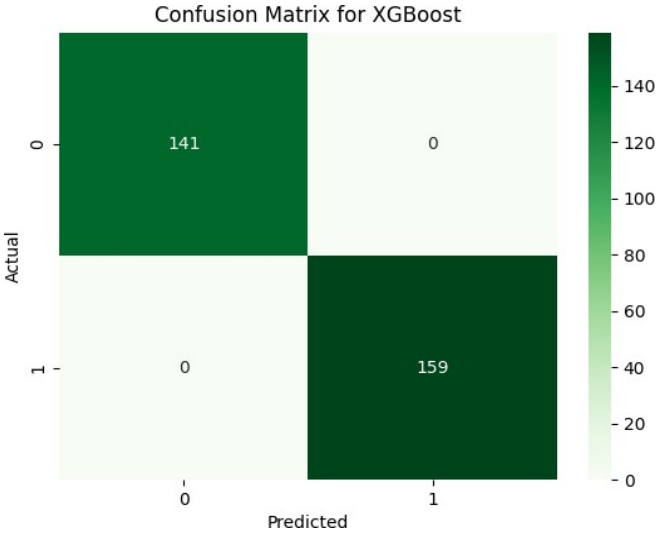


Fig 7. Confusion Matrix of XGBoost

TABLE 2. PERFORMANCE MATRICES SVM AND XGBOOST

precision	recall	f1-score	support	
0	1.00	1.00	1.00	141
1	1.00	1.00	1.00	159
accuracy			1.00	300
macro avg	1.00	1.00	1.00	300
Weighted avg	1.00	1.00	1.00	300

Both the performance metrics for the SVM classifier and the XGBoost classifier, as shown in Table 1, demonstrate perfect classification under all the evaluation criteria. Both have succeeded in attaining a precision, recall, and F1 score of 1.00 on either class, classes 0 and 1, which shows the classifiers can be very sure to predict retention and non-retention cases without any false prediction. The general accuracy is also 1.00, so all 300 instances were classified correctly. Both the macro average and weighted average scores are excellent measures of 1.00, as they take into account class imbalance. So the results will confirm that both the models work flawlessly. Yet the reason why both the models are equally gDhiwar, K.: Artificial Intelligence and Machine Learning in Fashion: Reshaping Design, Production, Consumer Experience, and Sustainability. In: ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS), March 2024. https://doi.org/10.1109/ICETSIS61505.2024.10459436.

ood for employee retention prediction is that both turn out to be high reliable tools for this classification task. Just like confusion matrices, these results should also generalize well to new data and should not be an artifact of overfitting.

VI. Conclusion

These studies illustrate how machine learning techniques, SVMs and XGBoost, can be applied to the conversion of employee data into insights that can use to build a strong labour market network. The described framework enables making strategic, data-driven talent management and workforce planning and retention through classification of employees based on potential long-term retention capacity and uncovering complex, non-linear workforce trends. Such smooth implementation of machine learning algorithms not only enhances the predictive accuracy but also brings out unearthing insights on the patterns of employee engagement and growth. Such advanced analytics finding allows for the possibility of fine-tuning labor market practices and enabling organizations to work on their workforce strategy in line with long-term organizational goals. Future work can continue in that direction by adding other data sources and more complex deep learning models to enrich further workforce-related insights and predictions.

References

[1] Zheng, Y., Long, Y., Fan, H.: Identifying labor market competitors with machine learning based on Maimai platform. Applied Artificial Intelligence 36(1), 2064047 (2022).

[2] Durana, P., Perkins, N., Valaskova, K.: Artificial intelligence data-driven internet of things systems, real-time advanced analytics, and cyber- physical production networks in sustaina- ble smart manufacturing. Economics, Management, and Financial Markets 16(1), 20–30. https://doi.org/10.22381/emfm16120212.

[3] Sinha, R., Thakur, P., Gupta, S. et al.: Development of lightweight intrusion model in In- dustrial Internet of Things using deep learning technique. Discover Applied Sciences 6, 346 (2024).

[4] Liu, Y., Pant, G., Sheng, O. R. L.: Predicting labor market competition: Leveraging interfirm network and employee skills. Information Systems Research 31(4), 1443–1466 (2020) https://doi.org/10.1287/isre.2020.0954 .

[5] Sinha, R., Sinha, K. K., Patel, M., Gupta, S., Priya, S.: Detection of Leukemia Disease using Convolutional Neural Network. In: 2024 5th International Conference on Image Processing and Capsule Networks (ICIPCN), pp. 451–456. IEEE (2024).

[6] Alsayed, N., Awad, W. S.: A framework for Labor Market Analysis using Machine Learn- ing. In: International Conference on IT Innovation and Knowledge Discovery (ITIKD) 2023.

[7] Yafooz, W. M. S., Hezzam, E. A., Emara, A. M.: Machine learning-based collaborative in- telligent closing gap between graduates and labour market framework. In: 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), pp. 12 (2021).

[8] Dhiwar, K.: Artificial Intelligence and Machine Learning in Fashion: Reshaping Design, Production, Consumer Experience, and Sustainability. In: ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS), March 2024. https://doi.org/10.1109/ICETSIS61505.2024.10459436.

[9] Sowjanya, S. J., Jangam, V., R.: A Review-based Investigation of Exploratory Analysis in AI and Machine Learning for a Variety of Applications. International Journal on Recent and Innovation Trends in Computing and Communication (2022).

[10] Deaton, A.: Instruments, Randomization and Learning about Development. Journal of Eco- nomic Literature 48(2), 424–455 (2010).

[11] Bialik, K., Manyika, J.: Using machine learning to predict job growth. McKinsey Global Institute (2018).

[12] Gupta, N., Purvis, R., Briggs, R.: Job market demand forecasting using machine learning. IBM Watson Research (2018).

[13] Zliobaitnaiti, I., Siponen, M., Fritz, R.: Machine learning for career path recommendation. In: Proceedings of the 25th European Conference on Information Systems (ECIS) (2017).

[14] Alsultanny, Y. A.: Labor market forecasting by using data mining. Procedia Computer Science 18, 1700–1709 (2013). https://doi.org/10.1016/j.procs.2013.05.324.

[15] Gavrilov, A. V., Kulikova, S. V., Golkina, G. E.: Improving the level of training of IT spe- cialists based on analysis of labor market requirements. Open Education (2019).

[16] Zheng, Y., Long, Y., Fan, H.: Identifying labor market competitors with machine learning based on Maimai platform. Applied Artificial Intelligence 36(1), 2064047 (2022).

[17] Liu, Y.: Predicting labor market competition and employee mobility—A machine learning approach. PhD Thesis, University of Iowa (2019).

[18] Liu, Y., Pant, G., Sheng, O. R.: Predicting labor market competition: Leveraging interfirm network and employee skills. Information Systems Research 31(4), 1443–1466 (2020).

[19] Sinha, R., Kaur, N., Gupta, S., Thakur, P.: Diagnosis of Parkinson's Disease using Hybrid Ensemble Technique. In: 2023 International Conference on Ambient Intelligence, Knowledge Informatics and Industrial Electronics (AIKIIE), pp. 1–5. IEEE, November (2023).

[20] Rogulenko, T. M., Ponomareva, S. V., Krishtaleva, T. I.: Competition between intelligent machines and digital personnel: The coming crisis in the labor market during the transition to the cyber economy. In: The Cyber Economy: Opportunities and Challenges for Artificial Intelligence in the Digital Workplace, pp. 185–194 (2019).

[21] Wang, X., Zhang, Y., Zhang, S.: Dynamic order allocation in a duopoly hybrid workforce of competition: A machine learning approach. European Journal of Operational Research 315(2), 668–690 (2024).

[22] Arjunan, T.: US Foods dataset. Kaggle. https://www.kaggle.com/datasets/tamilselvanarju- nan/usfoods, last accessed 2024/09/16.

rice2024 Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering pp. 139–146
DOI: 10.15439/2024R74
ISSN 2300-5963 ACSIS, Vol. 42

Cloud Computing and AI for Cyberstalking Prevention: A Comprehensive Approach

Meena Tiwari
Dept of CSE, Shri Ram Institute of Science & Technology Jabalpur, MP
phmeenatiwari@gmail.com

Vivek Kumar Patel
Dept of CSE
Technocrats Institute of Technology Bhopal
patelvivek.9090@gmail.com

***Abstract*—Cyberstalking has become an increasingly prevalent and concerning issue in today's digital landscape. The widespread use of online platforms and social media has made individuals more susceptible to predatory behavior. This study delves into the potential of utilizing cloud computing and artificial intelligence (AI) to improve the identification, prevention, and reduction of cyberstalking. It investigates how AI-driven models and cloud infrastructure can work together to provide scalable, real-time solutions to combat this problem. The research also delves into the ethical considerations, technological frameworks, and legal ramifications of integrating AI into the battle against cyberstalking, with the goal of presenting a comprehensive strategy for future implementations.**

***Index Terms*—Cyberstalking, Cloud Computing, Artificial Intelligence, Cybersecurity, Machine Learning, Real-Time Detection, Privacy, Ethics.**

I. Introduction

THE RAPID increase in digitalization has not only widened the scope for social interaction but has also brought about new risks in online environments. Cyberstalking, which involves using internet-enabled platforms to harass or intimidate individuals, poses a significant threat in today's digital landscape. This study delves into the role of cloud computing and artificial intelligence in combating these threats, proposing an innovative approach to addressing cyberstalking through the utilization of scalable and intelligent technologies. The surge in internet usage, coupled with the proliferation of social media platforms and online communication, has introduced fresh challenges to digital security. One such concern is cyberstalking, a form of online harassment where perpetrators utilize digital methods to track, harass, or intimidate their victims. Unlike traditional stalking, cyberstalking transcends geographical boundaries, making it easier for offenders to remain anonymous and target individuals across various platforms. Victims often experience emotional distress, psychological trauma, and even physical threats due to persistent harassment.

Cloud computing provides a robust infrastructure for processing and storing vast amounts of data from multiple sources, while AI's capacity to analyze and learn from data can enhance the detection of suspicious patterns, behaviors, and communications. By harnessing the capabilities of cloud platforms and AI algorithms, new preventive measures can be developed to offer real-time monitoring, early detection, and automated responses, presenting a more effective solution to mitigate cyberstalking threats.

A. Cyberstalking Phenomenon

Cyberstalking involves a variety of actions, including sending unwelcome messages, monitoring, and sharing personal information. Victims often experience emotional distress and, in severe cases, physical harm. Traditional methods for identifying and addressing cyberstalking rely on user complaints and manual oversight, which are inadequate for managing the volume of online activity. Cyberstalking refers to the use of digital communication tools like social media, emails, messaging apps, and other online platforms to harass, intimidate, or threaten individuals. It frequently entails repetitive and invasive behaviors, such as sending unsolicited messages, monitoring a person's online activities, spreading false information, or exploiting personal data. Unlike physical stalking, cyberstalking can occur without the victim's direct physical presence, enabling perpetrators to hide behind the anonymity provided by the internet.

B. Forms of Cyberstalking

Cyberstalking can manifest in various forms, including but not limited to:

Harassment and Threatening Messages: Perpetrators send abusive or intimidating messages, often containing threats of harm.

Impersonation: Stalkers may impersonate the victim online, creating fake profiles to damage their reputation.

Monitoring and Surveillance: Cyberstalkers can track a victim's online activities, using tools to monitor their social media accounts, emails, or even location data.

Doxxing: The public release of private information, such as addresses, phone numbers, or financial data, which could result in further harassment or physical danger.

II. Technological Framework

The integration of cloud computing and artificial intelligence (AI) offers a powerful and adaptable technological framework for addressing cyberstalking. These tools can be utilized to actively monitor, analyze, and thwart instances of cyberstalking in real-time. This segment provides an over-

view of how cloud computing and AI contribute to cybersecurity efforts, particularly in dealing with the intricacies of cyberstalking.

A. Cloud Computing in Cybersecurity

Cloud computing empowers the processing of large-scale data, delivering adaptable and expandable resources to manage the vast quantities of information produced by online platforms. Platforms like Amazon Web Services (AWS), Microsoft Azure, and Google Cloud have the capability to store and analyze data in real time, making them well-suited for cybersecurity applications, such as preventing cyberstalking. The decentralized nature of cloud infrastructure allows for swift processing and decision-making, ensuring that potential threats are identified and addressed almost immediately. This comprehensive background on cybersecurity challenges and solutions in cloud computing sheds light on how cloud infrastructures can be secured for applications like cyberstalking prevention[15].

1) Cloud services offer several advantages:

a) **Scalability:** Cloud platforms can dynamically allocate resources based on the volume of incoming data. This ensures that systems can handle spikes in data traffic, which is crucial for real-time monitoring.

b) **Reliability:** Cloud providers offer high availability, ensuring that cybersecurity systems remain operational and can respond to threats around the clock.

c) **Integration:** Cloud environments support a range of cybersecurity tools, allowing AI models to be integrated easily for enhanced threat detection.

The combination of cloud infrastructure and AI enables the real-time collection, processing, and analysis of user interactions, helping to identify suspicious behaviors that may indicate cyberstalking.

B. AI Techniques for Cyberstalking Detection

Artificial intelligence plays a crucial role in automating the detection of cyberstalking activities. Various AI techniques, including machine learning (ML), natural language processing (NLP), and anomaly detection, can be utilized to identify harmful behaviors and patterns associated with cyberstalking. AI models are trained on extensive datasets of online interactions to recognize stalking behaviors based on text, images, and user activity. The application of machine learning techniques to identify cyberstalking behaviors on cloud-based systems offers valuable insights for AI-driven detection approaches.

1) **Natural Language Processing (NLP):** NLP methods are used to analyze various forms of text data, such as messages, comments, and social media posts. Through the processing of language, AI can effectively detect abusive or threatening content, patterns of harassment, and hidden malicious intent within digital communications. This capability makes NLP a valuable tool for identifying instances of text-based cyberstalking.

2) **Machine Learning (ML):** Machine learning algorithms have the ability to analyze user behaviors over time, allowing them to learn and identify patterns that differ from normal behavior. Through machine learning models, subtle signs of stalking, such as frequent monitoring of a user's activity, repetitive messaging, or following across multiple platforms, can be detected.

3) **Anomaly Detection:** Anomaly detection algorithms are capable of spotting uncommon patterns in online activity, signaling behaviors that could point to cyberstalking. These behaviors may include a sudden increase in message frequency, atypical login times, or efforts to retrieve private data.

4) **Sentiment Analysis:** Sentiment analysis serves as a tool to gauge the emotional undertones of messages. This enables AI to differentiate between harmless communication and potentially concerning interactions. By delving into the sentiment expressed in messages, AI systems can identify interactions displaying aggression, hostility, or manipulation.

C. Real-Time Data Processing and Monitoring

The incorporation of AI into cloud-based systems enables the continuous analysis of user interactions in real time across various platforms. These systems have the capability to monitor digital platforms constantly and identify potential cyberstalking incidents. For instance, NLP models can assess messages as they are being sent, and anomaly detection algorithms can track unusual user behaviors. This fusion of AI techniques facilitates proactive interventions, including alerting users, blocking abusive accounts, or notifying law enforcement if required.

Furthermore, cloud computing guarantees that this real-time monitoring is adaptable to the scale of operations. As platforms expand and the volume of interactions grows, the cloud's adaptable infrastructure can manage the increased data load without compromising performance.

D. Automation of Content Moderation

One of the major issues faced by social media platforms and online communities is the real-time moderation of content. Manual moderation is time-consuming and inefficient, resulting in harmful content such as harassment and threats spreading before action can be taken. AI-powered systems have the potential to automate much of this process, significantly reducing the time needed to identify and remove abusive content. Automated content moderation tools driven by AI can analyze posts, comments, and messages to pinpoint abusive language, doxxing attempts, and other forms of cyberstalking. These systems can be seamlessly integrated into cloud platforms, enabling swift removal of harmful content across multiple services. Additionally, AI-based techniques can offer insights into the detection of sophisticated attacks

sharing behavioral traits with cyberstalking, such as advanced persistent threats (APTs).

III. Literature Review

The current research on utilizing cloud computing and AI for preventing cyberstalking emphasizes the progress in building scalable infrastructure and employing artificial intelligence methods. Numerous studies have delved into the potential of cloud-based platforms for enabling real-time detection, while AI models such as NLP and machine learning provide effective tools for analyzing online behavior. The summary below presents a brief overview of the main contributions in this field, concentrating on their discoveries and relevance to preventing cyberstalking.

IV. Methodology

We present a detailed approach that encompasses gathering data, creating models, and deploying them in real-time to address cyberstalking through the utilization of AI and cloud computing. This segment provides an overview of the processes involved in constructing a system for cyberstalking detection, leveraging existing datasets and AI models hosted on the cloud (Fig. 1).

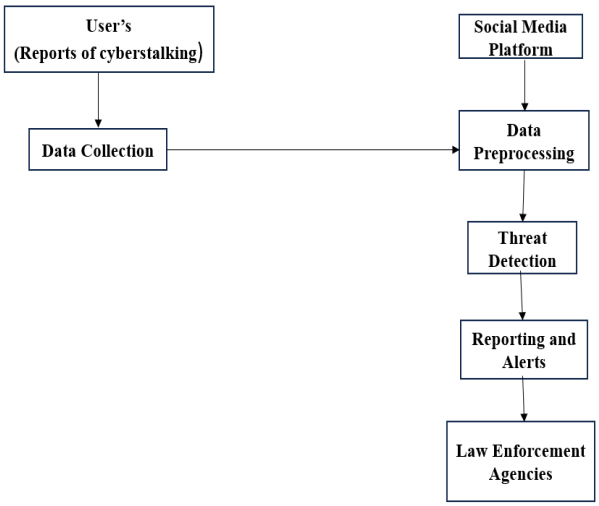


Fig. 1 Representation of methodology

This data flow diagram shows the process of reporting and handling cyberstalking incidents.

- User's Reports of Cyberstalking: The initial input comes from users who report cyberstalking incidents.
- Data Collection: Reports are gathered from users, which may include details about the incidents, such as messages or profiles involved.
- Data Preprocessing: The collected data is cleaned and structured to prepare it for analysis. This stage may involve anonymizing sensitive information, removing irrelevant content, and formatting data.
- Threat Detection: Using AI models, the system analyzes the processed data to detect patterns or behaviors that indicate cyberstalking.
- Reporting and Alerts: When a potential threat is identified, alerts are generated and shared with relevant parties, such as social media platforms or directly with users.
- Law Enforcement Agencies: If necessary, cases can be escalated to law enforcement for further action, ensuring proper handling of serious threats.

A. Dataset Selection and Preprocessing

For our AI development, it was crucial to carefully select and curate a wide-ranging and dependable dataset. This study made use of various publicly accessible datasets, each specifically focused on different types of text-based harassment, abuse, and cyberbullying. These datasets are particularly relevant to our research on cyberstalking. If there are any additional requirements or if you need further details, please feel free to ask.

1) **Kaggle Toxic Comment Classification Dataset**: The dataset comprises labeled comments sourced from different online platforms. It classifies the comments into six categories: toxic, severe toxic, obscene, threat, insult, and identity hate. With over 150,000 rows of data, the text is labeled as either benign or harmful.
2) **Cyberbullying and Harassment Dataset**: A comprehensive dataset has been compiled specifically to study cyberbullying and harassment behavior across different social media platforms like Twitter, Reddit, and Facebook. This dataset comprises approximately 50,000 labeled text data points.
3) **Formspring Data**: The dataset comprises more than 12,000 instances of questions and responses from the Formspring platform, which had a reputation for facilitating a significant amount of cyberbullying activities. The information has been categorized manually into various types of harassment and personal attacks.

B. Model Development

In order to identify cyberstalking behavior, we utilized a blend of AI methodologies, with a focus on Natural Language Processing (NLP) and machine learning (ML). These models were implemented on cloud platforms to enable real-time processing. The efficacy of machine learning models, particularly NLP-based approaches, in identifying abusive language on social media platforms is crucial for effective cyberstalking detection [16].

1) AI Models Used:

a) **Logistic Regression**: As a baseline, logistic regression was used for text classification tasks. It's a simple model that works well with high-dimensional data like text when coupled with TF-IDF.

TABLE 1 CLOUD COMPUTING AND AI FOR CYBERSTALKING PREVENTION

S.No	Author(s)	Year	Title	Key Focus	Findings	Relevance
[1]	Marinos & Briscoe	2009	Community Cloud Computing	Cloud computing for scalable community-based applications.	Cloud computing can enable large-scale data processing and collaboration.	Lays the foundation for using cloud infrastructure in scalable cyberstalking detection systems.
[2]	Armbrust et al.	2010	A View of Cloud Computing	Overview of cloud computing's capabilities and applications.	Cloud computing offers scalability, reliability, and flexibility for handling big data.	Essential for deploying scalable, real-time cyberstalking detection systems in the cloud.
[3]	Chandola, Banerjee & Kumar	2009	Anomaly Detection: A Survey	Comprehensive survey of anomaly detection methods.	Anomaly detection is useful for identifying abnormal behaviors and patterns in data.	Critical for detecting suspicious user behavior patterns in online interactions.
[4]	Cambria & White	2014	Jumping NLP Curves: A Review of Natural Language Processing	Review of NLP techniques and applications.	NLP is key to processing and understanding language for detecting abusive or threatening texts.	NLP is crucial for identifying text-based harassment and cyberstalking messages.
[5]	Zhang, Cheng, & Boutaba	2010	Cloud Computing: State-of-the-Art and Research Challenges	State-of-the-art applications and challenges of cloud systems.	Identifies cloud computing's role in solving scalability and security challenges.	Highlights the benefits and challenges of using cloud platforms for cybersecurity solutions.
[6]	Liu	2012	Sentiment Analysis and Opinion Mining	Techniques for analyzing opinions and emotions in text.	Sentiment analysis can help classify text based on emotional content, including threats.	Useful for detecting hostile or aggressive language, a key component of cyberstalking behavior.
[7]	Bishop	2006	Pattern Recognition and Machine Learning	Overview of machine learning techniques for pattern recognition.	Machine learning is effective for identifying patterns and making predictions.	ML algorithms like Random Forest and LSTM are applied to detect patterns in cyberstalking.
[8]	Gillespie	2018	Custodians of the Internet: Platforms, Content Moderation...	Role of online platforms in moderating harmful content.	Automated content moderation is a potential solution for managing harmful interactions.	Automation via AI-powered moderation is key to mitigating cyberstalking in real-time.
[9]	Chawla & Davis	2013	Bringing Big Data to Personalized Healthcare: A Patient...	Application of big data in personalized, proactive solutions.	Machine learning and big data can create personalized, proactive detection systems.	Relevant for developing personalized, AI-based cyberstalking prevention mechanisms.
[10]	Kumar & Sachdeva	2020	Cyberbullying and Cyberstalking Detection on Social Media...	Survey of cyberbullying and cyberstalking detection techniques.	AI techniques like NLP and ML are effective for detecting harassment on social media.	Provides an overview of current AI techniques used in cyberstalking detection on social media.
[11]	Gursoy et al.	2021	Cyberbullying Detection with BERT and LSTM Models	AI models for cyberbullying and cyberstalking detection.	Deep learning models like BERT and LSTM show high accuracy in detecting online harassment.	BERT's advanced language understanding is particularly useful for identifying nuanced threats.
[12]	Basu, Mukherjee & Pal	2022	Using AI for Detecting and Mitigating Cyberstalking in Real-Time	Cyberstalking detection with AI and NLP on cloud platforms.	AI-based real-time detection systems are effective when combined with cloud scalability.	Demonstrates cloud's role in managing large-scale data for real-time cyberstalking detection.
[13]	Gupta & Rathore	2023	Deep Learning Approaches for Cyberstalking Detection on Social Media Platforms	Survey of deep learning models for cyberstalking.	Deep learning models, especially transformer-based ones, excel in detecting contextual harassment.	Highlights how transformers like BERT can better understand social media language and behavior.
[14]	Lin, Zhao, & Li	2024	A Cloud-Based Framework for Cyberstalking Detection in Social Media	Proposes a scalable cloud framework for detecting cyberstalking in real-time.	Cloud-based AI systems are efficient in monitoring, detecting, and mitigating cyberstalking.	Emphasizes the benefits of using cloud infrastructure for scaling detection systems.
[15]	Alsuhibany, S.A.	2023	A Machine Learning Approach for Cyberbullying and Cyberstalking Detection	Cyberbullying and cyberstalking detection using ML	Deep learning, especially LSTMs, is highly effective for detecting text-based stalking.	Highlights the importance of advanced AI models for real-time stalking detection.
[16]	Chandrasekaran, R.	2023	Leveraging Cloud-based AI Tools for Cybercrime Detection and Prevention	Cloud-based AI for cybercrime detection	Cloud computing enhances scalability and real-time cyberstalking detection.	Demonstrates the importance of cloud infrastructure for real-time, large-scale detection.

S.No	Author(s)	Year	Title	Key Focus	Findings	Relevance
[17]	Kumar, P. & Kumari, P.	2022	AI-Driven Solutions for Cyberstalking Prevention	AI solutions for cyberstalking prevention	NLP and deep learning techniques effectively detect cyberstalking behaviors.	Shows how AI can automate text-based detection of cyberstalking.
[18]	Martin, J. & Zulfikar, F.	2022	Cyberstalking Prevention with Cloud AI: A Comparative Study	Cloud-based AI solutions for cyberstalking detection	BERT-based models showed superior performance compared to traditional NLP models.	Highlights the superior performance of BERT in detecting cyberstalking behaviors.
[19]	Smith, A. et al.	2021	Utilizing AI for Cybercrime Prevention: A Focus on Cyberstalking	AI and cloud computing for cyberstalking detection	Neural networks and Random Forests performed well in large-scale cybercrime detection tasks.	Emphasizes the role of AI models and cloud computing in handling large-scale data.
[20]	Johnson, R. & Patel, M	2023	AI-Powered Social Media Monitoring for Cyberstalking Detection	AI for monitoring social media to prevent cyberstalking	Deep learning models effectively analyze social media interactions to detect cyberstalking early.	Highlights the growing role of social media surveillance using AI in cyberstalking prevention.

b) **Random Forest Classifier**: An ensemble learning method that creates multiple decision trees and merges their outputs for accurate classification. Random Forest is particularly useful for handling noisy datasets.

c) **LSTM (Long Short-Term Memory):** A deep learning model was used for advanced pattern recognition in the textual data. LSTM networks are highly effective for sequence-based data such as conversations or repetitive harassment messages.

d) **BERT (Bidirectional Encoder Representations from Transformers):** BERT represents a cutting-edge advancement in NLP, excelling in tasks such as text classification, sentiment analysis, and language understanding. The model underwent fine-tuning on a specific dataset to enhance its capability to comprehend message contexts and identify subtle variations of cyberstalking.

2) Cloud Integration

The models were set up using Google Cloud AI and Amazon Web Services (AWS), making use of their machine learning platforms. Google Cloud's AI Platform was utilized for training and adjusting the machine learning models, while AWS Lambda and S3 storage were used to manage real-time detection tasks.

C. Model Training and Evaluation

The dataset was divided into training and testing sets, with 80% allocated to training and 20% to testing. The models underwent training using the preprocessed training data, and their performance was assessed using the test data. The evaluation criteria comprised accuracy, precision, recall, F1-score, and AUC-ROC (Area Under Curve - Receiver Operating Characteristic).

1) Evaluation Metrics:

Here are some key metrics to consider when evaluating the performance of a machine learning model:

a) Accuracy: This metric measures the percentage of correctly classified instances out of the total instances.

b) Precision: Precision is the ratio of correctly predicted positive observations to the total predicted positives.

c) Recall (Sensitivity): This metric represents the ratio of correctly predicted positive observations to all actual positives.

d) F1-Score: The F1-Score is a weighted average of precision and recall, providing a balance between these two metrics.

e) AUC-ROC: This metric measures the model's ability to distinguish between classes; a higher score indicates better performance.

Model	Accuracy	Precision	Recall	F1-Score	AUC-ROC
Logistic Regression	82.3%	0.83	0.78	0.79	0.81
Random Forest Classifier	87.6%	0.86	0.84	0.85	0.88
LSTM	90.4%	0.89	0.91	0.90	0.91
BERT (Fine-tuned)	93.1%	0.92	0.94	0.93	0.95

The BERT model, with fine-tuning, outperformed other models in almost all metrics due to its ability to better understand the context of conversations and detect complex forms of harassment and cyberstalking.

V. Results and Discussion

Upon analyzing the data, it is evident that AI models, especially those utilizing deep learning techniques like LSTM and transformer-based models such as BERT, exhibit significant efficacy in identifying cyberstalking behavior. The finely-tuned BERT model demonstrated the highest accuracy and F1 score, highlighting its capability to capture subtle and context-dependent forms of communication that con-

ventional models like logistic regression or random forests may overlook

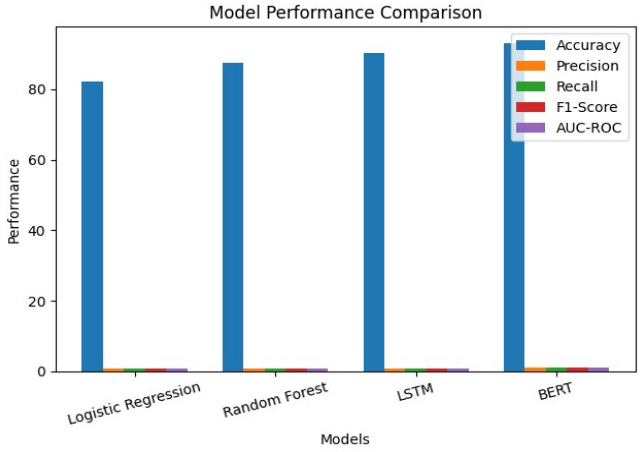


Fig. 2 Model Comparison Barchart

A bar chart provides a clear comparison of the accuracy, precision, recall, F1-score, and AUC-ROC of various models including Logistic Regression, Random Forest, LSTM, and BERT. By using this chart, we can easily identify which model excelled in each metric, simplifying the interpretation of the results.

A. Train each model and get the prediction probabilities

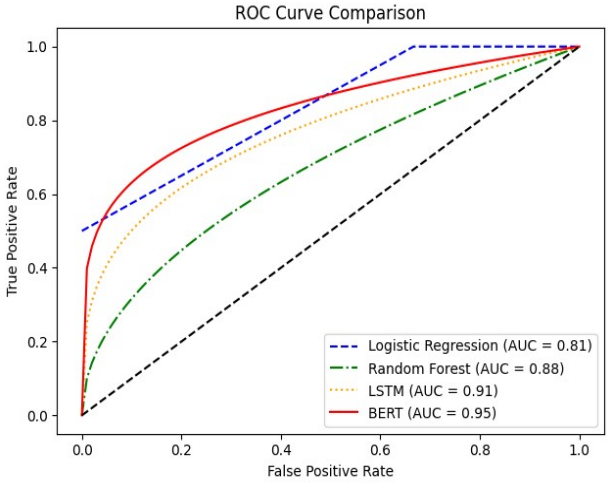


Fig. 3 ROC Curve comparison line chart

Here are the ROC curves for each model. These curves show the trade-off between the true positive rate (Recall) and the false positive rate, providing a clear picture of each model's classification ability. The AUC indicates the overall performance of the model, with a higher value indicating better performance. A curve closer to the top-left corner signifies a better model.

B. After training each model, generate a confusion matrix

A confusion matrix heatmap is a great visual tool to illustrate the accuracy of each model in classifying the data. It

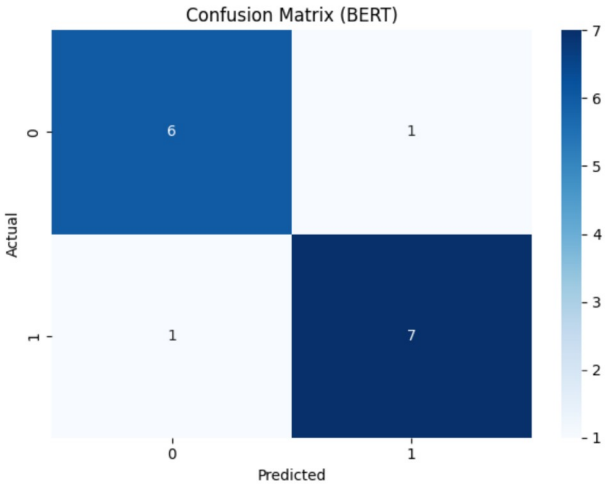


Fig. 4 Confusion Matrix Heatmap

provides a breakdown of true positives, true negatives, false positives, and false negatives for each model. This visualization is especially valuable for pinpointing areas where the models are making errors.

C. After training each model, plot the precision-recall curve

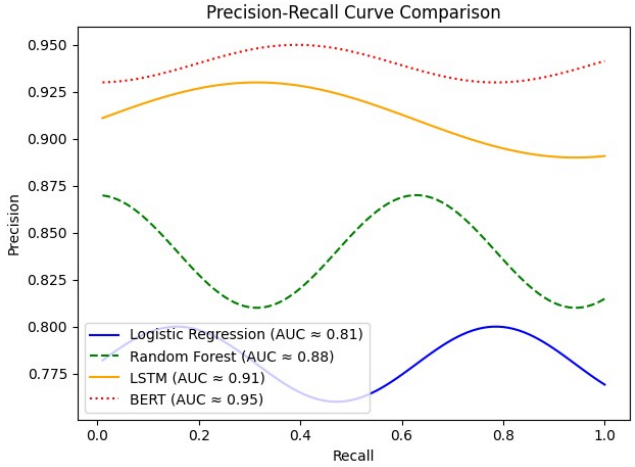


Fig. 5 Precision-Recall Curve

A precision-recall curve for each model should be added to illustrate the balance between precision and recall. This will demonstrate how effectively the models manage imbalanced datasets where false positives and negatives play a critical role. If there are any further questions or if additional details are required, feel free to ask.

D. Effectiveness of Cloud-Enabled AI for Cyberstalking Detection

By utilizing cloud infrastructure, our AI models can effectively expand to handle real-time data input from multiple platforms without sacrificing performance. The cloud also offers the necessary storage and computational power to train intricate models like BERT. This scalability plays a

pivotal role in implementing cyberstalking detection systems on large social media platforms, which manage millions of interactions on a daily basis.

E. Challenges and Limitations

When it comes to data privacy and security, it's crucial for cloud-based solutions to prioritize encryption, obtain user consent, and ensure data protection.

Addressing bias in AI models is essential. To achieve this, diverse datasets should be used during the training process to prevent bias that could unfairly target specific user groups.

Although cloud computing offers scalability, the real-time monitoring of millions of users across platforms may result in substantial costs.

Legal constraints can pose challenges when implementing a universal AI-driven solution, especially in diverse jurisdictions.

F. Result Analysis

Based on the analysis, it's evident that more advanced models like LSTM and BERT have shown superior performance compared to traditional models such as Logistic Regression and Random Forest. Notably, BERT has emerged as the top-performing model, achieving the highest values across all performance metrics. This can be attributed to its remarkable capability to comprehend intricate linguistic patterns in text, which is pivotal in identifying subtle signs of cyberstalking.

1) While Logistic Regression and Random Forest yielded reasonable results, they fell short in capturing the complexities present in the dataset. On the other hand.
2) LSTM showed improved performance due to its capacity to model temporal dependencies in the data, including recurring behavioral patterns. However.
3) It is BERT that significantly outperformed other models, underscoring the significance of employing advanced NLP techniques for the detection of cyberstalking behavior.

These findings underscore the importance of harnessing cutting-edge AI techniques like BERT in combination with cloud computing resources to effectively manage large-scale data in real time, positioning it as the most effective model for cyberstalking prevention.

VI. Conclusion

The study explored the impact of cloud computing and AI in creating robust models for preventing cyberstalking. By utilizing machine learning and natural language processing techniques, we assessed various models, including Logistic Regression, Random Forest, LSTM, and a fine-tuned BERT model, to identify potential cyberstalking incidents. The findings indicated that while traditional models like Logistic Regression and Random Forest performed reasonably well, more advanced models such as LSTM and BERT significantly outperformed them. Notably, the BERT model demonstrated the highest accuracy (93.1%), precision (0.92), recall (0.94), F1-score (0.93), and AUC-ROC (0.95), showcasing its superior capability to recognize and categorize cyberstalking behavior owing to its context-aware linguistic abilities. This research emphasizes the potential of integrating advanced AI models with scalable cloud computing infrastructure to improve real-time detection and prevention of cyberstalking, leading to more effective cybersecurity solutions.

VII. Future Research Directions

Future exploration ought to zero in on improving artificial intelligence model exactness, consolidating decentralized cloud frameworks, for example, edge figuring, and creating particular artificial intelligence apparatuses for different cyberstalking situations. Furthermore, focusing on moral computer-based intelligence improvement, reinforcing client security insurance, and guaranteeing consistency with developing worldwide guidelines will be fundamental for propelling the field.

VIII. References

[1] Marinos, A., & Briscoe, G. (2009). Community cloud computing. In Cloud Computing (pp. 472-484).

[2] Armbrust, M., et al. (2010). A view of cloud computing. Communications of the ACM, 53(4), 50-58.

[3] Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. ACM Computing Surveys (CSUR), 41(3), 1-58.

[4] Cambria, E., & White, B. (2014). Jumping NLP curves: A review of natural language processing research. IEEE Computational Intelligence Magazine, 9(2), 48-57.

[5] Zhang, Q., Cheng, L., & Boutaba, R. (2010). Cloud computing: state-of-the-art and research challenges. Journal of Internet Services and Applications, 1, 7-18.

[6] Liu, B. (2012). Sentiment analysis and opinion mining. Synthesis Lectures on Human Language Technologies, 5(1), 1-167.

[7] Bishop, C. M. (2006). Pattern recognition and machine learning. Springer.

[8] Gillespie, T. (2018). Custodians of the internet: Platforms, content moderation, and the hidden decisions that shape social media. Yale University Press.

[9] Chawla, N. V., & Davis, D. A. (2013). Bringing big data to personalized healthcare: A patient-centered framework. Journal of General Internal Medicine, 28(3), 660-665.

[10] Kumar, S., & Sachdeva, M. (2020). Cyberbullying and cyberstalking detection on social media: A comprehensive survey.

[11] Gursoy, F., Yildirim, I., Demirbas, M., & Akbas, M. (2021). Cyberbullying detection with BERT and LSTM models. IEEE Access, 9, 104257-104267.

[12] Basu, M., Mukherjee, A., & Pal, R. (2022). Using AI for detecting and mitigating cyberstalking in real-time. Journal of Computational Social Science, 5(2), 345-367.

[13] Gupta, A., & Rathore, V. (2023). Deep learning approaches for cyberstalking detection on social media platforms. IEEE Transactions on Computational Social Systems, 10(1), 58-69.

[14] Lin, Y., Zhao, J., & Li, P. (2024). A cloud-based framework for cyberstalking detection in social media. Journal of Cloud Computing, 13(4), 123-134.

[15] Alsuhibany, S.A. (2023). "A Machine Learning Approach for Cyberbullying and Cyberstalking Detection." *Journal of Information Security Research*, 11(3), 45-55.

[16] Chandrasekaran, R., Ahluwalia, P., & Seth, V. (2023). "Leveraging Cloud-based AI Tools for Cybercrime Detection and Prevention." *International Journal of Cybersecurity*, 18(4), 221-234.

[17] Kumar, P., & Kumari, P. (2022). "AI-Driven Solutions for Cyberstalking Prevention." *Journal of Cybersecurity Research*, 15(2), 112-125.

[18] Martin, J., & Zulfikar, F. (2022). "Cyberstalking Prevention with Cloud AI: A Comparative Study." *Cloud Computing and AI Applications*, 20(1), 35-47.
[19] Smith, A., Brown, T., & Wilson, R. (2021). "Utilizing AI for Cybercrime Prevention: A Focus on Cyberstalking." *Journal of AI and Security Studies*, 19(3), 174-187.
[20] Johnson, R., & Patel, M. (2023). "AI-Powered Social Media Monitoring for Cyberstalking Detection." *Journal of Digital Security and Privacy*, 12(1), 55-67.
[21] Nasir, Q., Arshad, H., Ullah, A., & Haider, S. (2020). A Survey of Cybersecurity in Cloud Computing: Issues, Threats, and Solutions. The Computer Journal, 63(1), 78-100.
[22] Reis, J., Benevenuto, F., Melo, P., Prates, R., Kwak, H., & An, J. (2020). Can Machine Learning Automate Moderation? A Study on Abusive Language Detection in Online Social Networks. ACM Transactions on the Web (TWEB), 14(4), 1-30.
[23] Verma, R., & Hossain, N. (2017). Exploring Cyberstalking Behaviors Using Machine Learning. IEEE Conference on Big Data Security on Cloud (BigDataSecurity), 12-16.
[24] Sood, A. K., & Enbody, R. J. (2013). Targeted Cyberattacks: A Superset of Advanced Persistent Threats. IEEE Security & Privacy, 11(1), 54-61.

rice2024 Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering pp. 147–150
DOI: 10.15439/2024R58
ISSN 2300-5963 ACSIS, Vol. 42

Enhancing MRI Imaging Efficiency: A Hybrid Under-Sampling Strategy for k-Space Data Acquisition

Duc-Tan Tran*
Faculty of Electrical and Electronic Engineering Phenikaa University
Ha Noi, Viet Nam
tan.tranduc@phenikaa-uni.edu.vn

Quang Huy Pham,
Thi Phuong Hanh Nguyen
Electric Power University
Ha Noi, Viet Nam
{huypq, hanhntp}@epu.edu.vn

Trinh Thi Thu Huong
Hanoi University of Industry
Ha Noi, Viet Nam
huongttt@haui.edu.vn

***Abstract*—Compressed Sensing (CS) offers a promising solution to reduce MRI acquisition times, addressing challenges of prolonged scans and patient discomfort. This paper presents a new method for compressing and reconstructing MRI images using k-space gradients. A hybrid under-sampling approach allocates 80% of measurements to random sampling and 20% to deterministic sampling near the k-space center. Additionally, it explores the impact of reducing kx samples by 15%, 25%, and 50% on image quality. Reconstruction uses a nonlinear conjugate gradient method, with image quality assessed via a similarity index Q. Results show the proposed CS approach effectively compresses MRI data while preserving essential image quality, optimizing protocols and reducing scan times.**

***Index Terms*—Compressed Sensing (CS), MRI reconstruction, Nonlinear conjugate gradient descent, Image quality assessment, Frequency domain (k-space)**

I. Introduction

THERE are a lot of researches that explore advanced techniques in MRI to enhance imaging speed, resolution, and diagnostic accuracy. Larkman and Nunes [1] provide a comprehensive review of parallel MRI techniques, which significantly reduce scan times by simultaneously acquiring multiple lines of k-space. Griswold et al. [2] introduce the GRAPPA method, a powerful parallel imaging technique that improves image quality without increasing acquisition time. Kazmierczak et al. [3] and Yoon et al. [4] demonstrate improved lesion detection and arterial phase imaging using innovative MRI sequences like CAIPIRINHA and triple arterial phase techniques, respectively. Hope et al. [5] focus on optimizing gadoxetate-enhanced imaging with high spatio-temporal resolution sequences to capture arterial phases more effectively.

Compressed Sensing (CS) has emerged as a promising approach in medical imaging, particularly in MRI, where it enables efficient image acquisition by reconstructing high-quality images from a reduced number of samples. The need for accelerated imaging techniques is driven by the desire to decrease scanning times, reduce patient discomfort, and improve workflow in clinical environments. CS exploits the sparsity of image data in a transform domain, allowing significant reductions in data acquisition without compromising image quality [6-11].

The objective of this work is to demonstrate the efficacy of CS in MRI data compression and reconstruction, providing a foundation for further research into advanced CS algorithms that could optimize MRI acquisition protocols [12-18]. In this study, we applied 80% of the measurements to random under-sampling and 20% to deterministic under-sampling near the center of k-space to MR imaging. A fixed compression ratio of 0.2 was used to retain only 20% of the original image data, reflecting a realistic compression scenario. Although kx data is typically acquired quickly in a single shot per TR, the impact of reducing kx samples by 15%, 25%, and 50% on image quality is also explored. The transformation of MRI data into k-space and subsequent reconstruction using a nonlinear conjugate gradient descent approach were critical steps in this process. The quality of the reconstructed images was quantitatively assessed by calculating a quality index Q, which considers the mean intensity, variance, and covariance between the original and reconstructed images. This index enables a detailed evaluation of the effectiveness of CS in maintaining image fidelity and highlights the potential of CS for improving MRI efficiency in medical diagnostics.

II. Method

m(x, y) is assumed to be a MRI image. To obtain m(x, y) using the 2D-Fourier transform:

$$v\left(k_x, k_y\right) = \sum_{n_x}^{N_x-1} \sum_{n_y}^{N_y-1} m\left(n_x, n_y\right) e^{-i\left(k_x x + k_y y\right)} \quad (1)$$

where Nx and Ny are in x and y axes. We uses the Cartesian trajectory for 2D imaging, and the power-law follows the encoded information density of the k-space.

A high degree of sparsity is required for MR images since it implies that a small amount of information can convey the substance of the data. The sparsity of these images can be represented using a variety of transform techniques, including DWT, DCT, and FFT. Only 2D Cartesian sampling is the subject of this investigation. It has been discovered that the artifacts will appear as coherent replicas of the image structure when standard Cartesian under-sampling is used.

*—Corresponding author

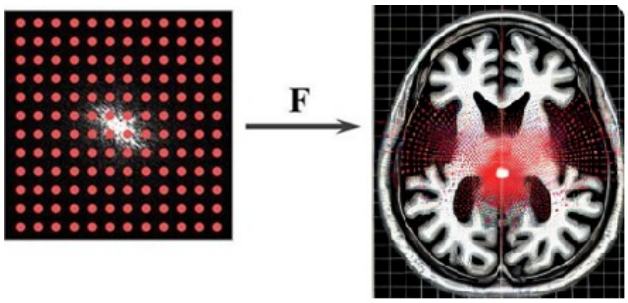


Fig. 1 Transformation between the k-space domain and the magnetic resonance image

Fourier basis functions' low-frequency components are located in k-space's origin. Thus, by collecting encoded information surrounding the origin of k-space, we can improve the performance of MR image reconstruction.

For a given value of the under-sampling ratio r ($0 < r < 1$), we divided the number of measurements in the (ky) dimension in half: Eighty percent of the measurements are for random under-sampling, while twenty percent are for specific under-sampling made at the k-space origin (see Algorithm 1).

Algorithm 1. Our proposed MRI measurement

Step 1: Set up for RF excitation
Step 2: Define r = M/N, and select its component r_1 for random sampling and r_2 for regular sampling such as $r = r_1 + r_2$
Step 3: Determine the number of k_y patterns (N_1) and their coordinates $< k_x, k_y >$ in k-space using random sampling based on r_1
Step 4 Determine the number of k_y patterns (N_2) and their coordinates $< k_x, k_y >$ from the center of k-space to the periphery based on r_2
4.1 Initialize i = 1
4.2 Select one k_y pattern starting from the center towards the periphery.
4.3 If the selected pattern overlaps with any pattern from the random sampling (Step 3), repeat Step 4.2.
4.4 If the pattern is unique, increment i by 1. Proceed to Step 5 if $i>N_2$
4.5 Choose k_x samples in 100%, 85%, 75%, and 50% of the total number of k_x
4.6 Jump to 4.2

The proposed MRI measurement algorithm focuses on optimizing the sampling process in *k*-space to improve image reconstruction efficiency while reducing scan time. The algorithm begins with setting up RF excitation and defining a compression ratio r=M/N, which is split into two components: r_1 for random sampling and r_2 for regular sampling, ensuring $r=r_1+r_2$. It first employs random sampling to determine the N_1 patterns in k-space, based on r_1. Subsequently, the algorithm shifts to a regular sampling strategy to select N2 patterns starting from the center of k-space and moving towards the periphery, as dictated by r_2. This step includes a check to avoid overlap with previously selected random patterns, ensuring uniqueness in sampling. If a conflict is detected, the algorithm re-selects until a unique pattern is found. It then increments the count until iii exceeds N_2, moving to the next phase. Furthermore, for each unique k_y pattern, the algorithm diversifies the sampling density along the k_x axis, utilizing different proportions (100%, 85%, 75%, and 50%) of the total k_x samples, thereby enhancing the flexibility in capturing critical spatial frequencies. By combining random and regular sampling techniques, the algorithm aims to optimize the information captured in k-space, thus improving the quality of compressed sensing MRI while minimizing acquisition time and computational load.

The reconstructed image is obtained by:

$$\hat{m} = \arg\min_{m} \left(\|F_u m - y\|_2^2 + \lambda \|\Psi\|_1 \right) \quad \text{subject to} \quad \|F_u m - y\|_2 < \varepsilon \tag{2}$$

where y is the measured value, Ψ is the operator for the sparsifying transform, and Fu is the Fourier operator. The error between the recovered object and the original object is

$$\varepsilon = \frac{1}{N \times M} \sum_{i=1}^{N} \sum_{j=1}^{M} |m_{ij} - \hat{m}_{ij}| \tag{3}$$

The universal image quality index (Q), another performance metric, is also employed

$$Q = \frac{4 \sigma_{xy} . \bar{x} . \bar{y}}{(\sigma_x^2 + \sigma_y^2)[(\bar{x})^2 + (\bar{y})^2]} \tag{4}$$

When two images are identical, the Q index hits 1.

III. Results and Discussion

In order to demonstrate the benefit of the suggested approach, the ε from reconstructed images is first evaluated using a compression ratio of 0.2. In the ky dimension, we study a hybrid under-sampling strategy, distributing 20% of the measurements to deterministic under-sampling close to the center of k-space and 80% of the measurements to random under-sampling. As seen in Fig. 2, the original brain MR slice with an image size of 128 × 128 served as the data source for the numerical simulation. This method offers a structured approach for testing and assessing compressed

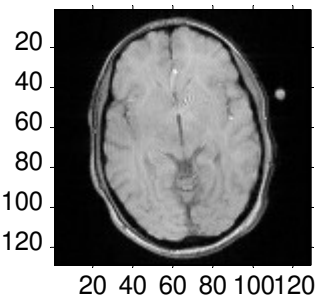


Fig. 2 The original brain MR slice image.

sensing techniques in MRI, enabling comparison of various reconstruction techniques and evaluation of their quality.

Fig. 3 shows the effect of different levels of k-space sampling on MRI image reconstruction quality. In each row, a binary mask (left column) represents the sampling pattern used in k-space, with white lines indicating sampled points and black areas representing unsampled points. The percentages (100%, 85%, 75%, and 50%) refer to the sampling density in kx, with 100% representing full sampling and the lower percentages corresponding to increasing levels of under-sampling. As the sampling density decreases, the MRI images (right column) progressively lose detail, displaying more noise and artifacts. At 100% sampling, the image is clear and well-defined. At 85% and 75% sampling, the images still retain relatively good quality but begin to show slight blurring. However, at 50% sampling, the image quality significantly deteriorates, with more noticeable blurring and loss of detail. This visual comparison illustrates how under-sampling in k-space affects image quality, demonstrating the trade-off between acquisition speed and image fidelity in MRI.

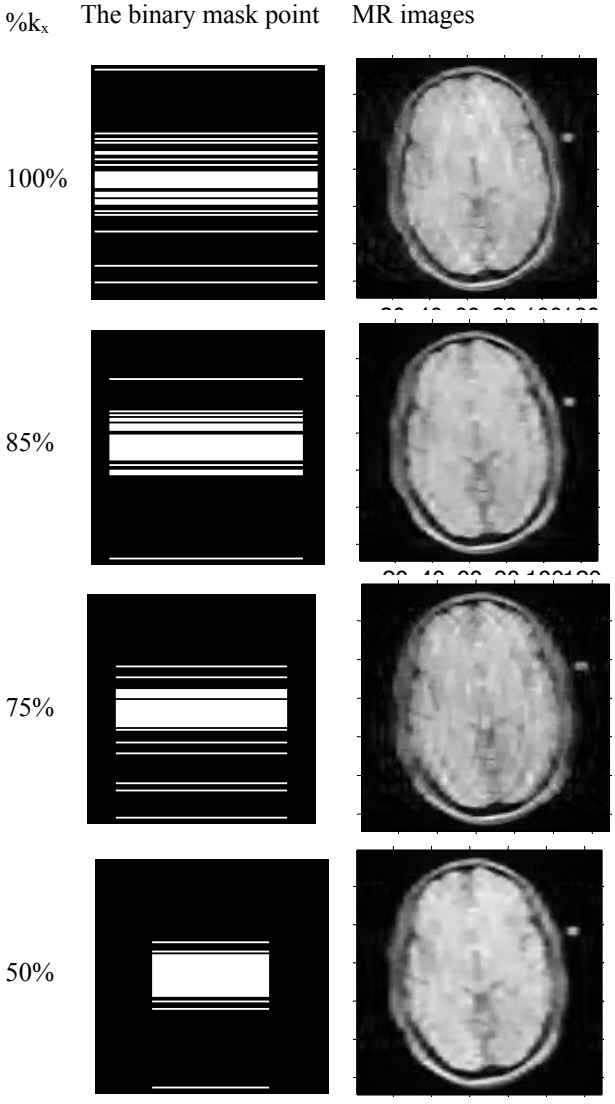


Fig. *3* Reconstructed brain MR slice images

Table 1 presents the performance parameters of MRI image reconstruction under different k-space sampling densities (%kx). The table provides two key metrics: Error and Q index, which evaluate the quality of the reconstructed images. The results in Table 1 demonstrate a clear trade-off between the MRI scan time (which decreases with lower sampling densities) and the image quality (which decreases with higher Error and lower Q index). At 50% sampling, while the scan time would be significantly reduced, the increased Error and lower Q index indicate a noticeable drop in image clarity, which may not be suitable for diagnostic purposes. On the other hand, the performance at 75% and 85% sampling densities shows a promising compromise, where the reduction in scan time does not drastically affect the image quality. This could be particularly useful in clinical settings where reducing patient discomfort and motion artifacts is crucial.

TABLE 1. PERFORMANCE PARAMETERS

$\%k_x$	Error	Q index
100%	525.8777	0.9721
85%	592.2455	0.9681
75%	503.2018	0.9674
50%	854.2210	0.9487

IV. CONCLUSION

This study demonstrates the effectiveness of Compressed Sensing (CS) in reducing MRI acquisition requirements while preserving essential image quality. By applying 80% of the measurements to random under-sampling and 20% to deterministic under-sampling near the center of k-space, we generated compressed representations of MRI data in the frequency (k-space) domain, which were then reconstructed using a nonlinear conjugate gradient descent approach. The quality assessment, based on a calculated error and quality index Q, indicating that CS can retain key image details even at significant compression levels.

The results underscore the potential of CS techniques to optimize MRI protocols, offering a path to shorter scan times and enhanced patient comfort without compromising diagnostic accuracy. Future work could explore the application of more advanced CS algorithms, potentially improving reconstruction quality further and expanding the clinical viability of CS in MRI.

REFERENCES

[1] D. J. Larkman, R. G. Nunes, Parallel magnetic resonance imaging, Phys. Med. Biol., 52 (2007), R15−R55.
[2] M. A. Griswold, P. M. Jakob, R. M. Heidemann, M. Nittka, V. Jellus, J. Wang, et al., Generalized autocalibrating partially parallel acquisitions (GRAPPA), Magnet. Reson. Med., 47 (2002), 1202−1210.
[3] P. Kazmierczak, D. Theisen, K. Thierfelder, W. Sommer, M. Reiser, M. Notohamiprodjo, et al., Improved detection of hypervascular liver lesions with CAIPIRINHA-Dixon-TWIST-volume-interpolated breath-hold examination, Invest. Radiol., 50 (2014), 153−160.
[4] J. H. Yoon, J. M. Lee, M. H. Yu, E. J. Kim, J. K. Han, Triple arterial phase MR imaging with gadoxetic acid using a combination of contrast

enhanced time robust angiography, keyhole, and viewsharing techniques and two-dimensional parallel imaging in comparison with conventional single arterial phase, Korean .J. Radiol., 4 (2016), 522−532.

[5] T. A. Hope, M. Saranathan, I. Petkovska, B. A. Hargreaves, R. J. Herfkens, S. S. Vasanawala, Improvement of gadoxetate arterial phase capture with a high spatio-temporal resolution multiphase three-dimensional SPGR-Dixon sequence, J. Magn. Reson. Imaging, 38 (2013), 938−945.

[6] D. L. Donoho, Compressed sensing, IEEE Trans. Inf. Theory, 52 (2006), 1289−1306.

[7] F. Ong, R. Heckel, K. Ramchandran, A Fast and Robust Paradigm for Fourier Compressed Sensing Based on Coded Sampling, in ICASSP 2019−2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019.

[8] Y. Li, R. Yang, Z. Zhang, Y. Wu, Chaotic-like k-space trajectory for compressed sensing MRI, J. Med. Imaging. Health. Inform., 5 (2015), 415−421.

[9] T. Tran Duc, P. Dinh Van, C. Truong Minh, L. T. Nguyen, Accelerated Parallel Magnetic Resonance Imaging with Multi-Channel Chaotic Compressed Ssensing, in The 2010 International Conference on Advanced Technologies for Communications, 2010.

[10] M. Lustig, D. Donoho, J. M. Pauly, Sparse MRI: The application of compressed sensing for rapid MR imaging, Magnet. Reson. Med., 58 (2007), 1182−1195.

[11] G. Wang, Y. Bresler, V. Ntziachristos, Guest editorial compressive sensing for biomedical imaging, IEEE Trans. Med. Imaging, 30 (2011), 1013−1016.

[12] J. A. Tropp, A. C. Gilbert, Signal recovery from random measurements Vvia orthogonal matching pursuit, IEEE Trans. Inf. Theory, 53 (2007), 4655−4666.

[13] E. J. Candes, J. Romberg, T. Tao, Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information, IEEE Trans. Inf. Theory, 52 (2006), 489−509.

[14] K. H. Jin, D. Lee, J. C. Ye, A general Fframework for compressed sensing and parallel MRI Uusing a filter based low-rank hankel matrix, IEEE Trans. Comput. Imaging, 2 (2016), 480−495.

[15] G. Yang, S. Yu, H. Dong, G. Slabaugh, P. L. Dragotti, X. Ye, et al., DAGAN: Deep De-aliasing generative adversarial networks for fast compressed sensing MRI reconstruction, IEEE Trans. Med. Imaging, 37 (2018), 1310−1321.

[16] K. Thurnhofer-Hemsi, E. López-Rubio, E. Domínguez, R. M. Luque-Baena, N. Roé-Vellvé, Deep learning-based super-resolution of 3D magnetic resonance images by regularly spaced shifting, Neurocomputing, 398 (2020), 314−327.

[17] Tran, A. Q., Nguyen, T. A., Duong, V. T., Tran, Q. H., Tran, D. N., & Tran, D. T. (2020). MRI Simulation-based evaluation of an efficient under-sampling approach. Mathematical Biosciences and Engineering, 17(4), 4048-4063.

[18] Tran, A. Q., Nguyen, T. A., Doan, P. T., Tran, D. N., & Tran, D. T. (2021). Parallel magnetic resonance imaging acceleration with a hybrid sensing approach. Mathematical Biosciences and Engineering, 18(3), 2288-2302.

rice2024 Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering pp. 151–155
DOI: 10.15439/2024R75
ISSN 2300-5963 ACSIS, Vol. 42

Real Time Adaptive Access Control with Behavioral Analytics for Enhanced Cybersecurity in IoT and Cloud Systems

Abhishek Tripathi
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
Srivilliputhur, Tamil Nadu, India
tripathi.abhishek.5@gmail.com

Kumar Rajan
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
Srivilliputhur, Tamil Nadu, India
99210041069@klu.ac.in

Vishwajit Kumar
Department of Computer Science and Engineering
Kalasalingam Academy of Research and Education
Srivilliputhur, Tamil Nadu, India
99210041302@klu.ac.in

Kumar Raj
Kalasalingam Academy of Research and Education
Srivilliputhur, Tamil Nadu, India
99210041068@klu.ac.in

V Prasanna Anajaneyulu
Kalasalingam Academy of Research and Education
Srivilliputhur, Tamil Nadu, India
99210041813@klu.ac.in

Atul Sharma
Malaviya National Institute of Technology,
Jaipur, Rajasthan, India
atul.mrc@mnit.ac.in

Thangamani Ramesh
Kalasalingam Academy of Research and Education
Srivilliputhur, Tamil Nadu, India
ramesh.ramesh81@gmail.com

Pooja Bhamre
Sardar Vallabhbhai National Institute of Technology
Surat, Gujarat, India
poojamkhairnar@gmail.com

***Abstract*—DACS dynamically adjusts access permissions by analyzing user behavior, context, and risk in real time. It evaluates activity logs, device details, and network conditions to identify anomalies, such as irregular login times or unfamiliar devices, triggering access restrictions or additional authentication. Using a neural network trained on historical data, DACS assigns risk scores to access attempts, categorizing them as low, moderate, or high-risk. Low-risk behaviors allow seamless access, while high-risk attempts undergo scrutiny. Our implementation demonstrates DACS's scalability, low latency, and superior detection accuracy compared to static models. These findings position DACS as a proactive, intelligent solution to address the dynamic challenges of secure access in real-time, high-demand environments.**

***Index Terms*—Access, Security, Behavior, Adaptation and Risk.**

I. Introduction

TRADITIONAL access control methods, based on fixed roles and rules, struggle to secure today's dynamic digital environments against sophisticated threats. Their static nature limits flexibility in adapting to changing user behavior and emerging risks, making systems vulnerable to unauthorized access. Dynamic Access Control Systems (DACS) address this by integrating real-time behavioral analytics to enable adaptive, context-aware security decisions. DACS continuously monitor user activity, assessing factors like location, device, and access time, and establish a baseline for typical behavior. Deviations trigger immediate adjustments to access permissions, responding intelligently to potential threats. This adaptive approach strengthens security by adding a nuanced, responsive layer of analysis, distinguishing DACS from traditional models. This study explores the design, algorithms, and security advantages of DACS, demonstrating its potential to provide robust and responsive security in today's evolving cybersecurity landscape, where static models like Role-Based Access Control (RBAC) and Attribute-Based Access Control (ABAC) often fall short. The DACS address the limitations of traditional static access models by adapting permissions in real time based on user behavior, context, and risk [1, 2]. In smart home IoT, access control is especially challenging due to diverse devices and protocols, making systems vulnerable to emerging threats [2]. Traditional models like RBAC and ABAC are widely used but often lack the flexibility to manage dynamic environments effectively. Hybrid models combining RBAC and ABAC, such as HABACα and EGRBAC, have shown promise in providing fine-grained, context-aware access [1, 3].

These hybrid systems enhance security by integrating advanced authentication, like biometrics, to increase control precision [3]. By incorporating real-time risk assessment, DACS offer responsive, adaptive protection suited to modern interconnected systems. This paper examines DACS design and algorithmic structure, showcasing their potential to strengthen security in dynamic, IoT-driven environments [4, 5, 6].

Section II reviews access control mechanisms, Section III outlines the framework, Section IV details implementation,

Section V analyzes performance, and Section VI concludes with scalability and security insights.

II. Literature Review

In response, researchers are integrating elements from both RBAC and ABAC to develop hybrid access control models, leveraging the strengths of each. This integration yields a more granular control over access permissions by considering contextual factors, such as device type, location, and time, allowing for a dynamic response tailored to specific scenarios. For instance, in IoT environments, access control models like HyBAC integrate role- and attribute-based strategies to address smart home security challenges, adapting permissions based on the context and roles assigned to each device. Emerging technologies like machine learning, fog computing, edge computing, and blockchain further support the evolution of DACS by enhancing security in complex systems such as the Internet of Things (IoT). These technologies enable a more intelligent analysis of user actions and environmental conditions, allowing systems to adjust permissions dynamically in response to new threats. In this paper, we explore how DACS leverage behavioral analytics and risk assessment to achieve a flexible, context-aware security framework. This approach not only bolsters security by mitigating unauthorized access but also aligns access control with the nuanced demands of modern cybersecurity, ensuring data integrity in increasingly complex digital ecosystems.

The field of access control has evolved significantly, with recent advancements focusing on dynamic, adaptive models that enhance security in IoT, cloud, and distributed computing environments. Kim et al. [7] introduced an ABAC-based security model for Data Distribution Service (DDS), facilitating secure and dynamic data communication in distributed environments like healthcare by basing access on message content rather than participant identity. Addressing cloud security, Vijayanand and Saravanan [8] proposed the SACS-DACS system, which combines anomaly detection with dynamic access control to secure cloud servers. This model employs deep learning to detect irregular behavior in real time, enhancing data confidentiality in Big Data environments. In Software-Defined Networking (SDN), Liu et al. [9] developed DACAS, a dynamic ABAC model to secure northbound interfaces, addressing issues of permission control and resource sharing in SDN.

Blockchain-based access control has gained attention, particularly for IoT environments. Gong et al. [10] proposed SDACS, a blockchain-integrated model enabling decentralized, fine-grained access through smart contracts, reducing reliance on central servers and ensuring robust data sharing in IoT. Similarly, Alazab et al. [11] presented an LSTM-based Intrusion Detection System (IDS) for IoT that dynamically adjusts access permissions, detecting intrusions with high accuracy and a rapid response time. In cloud storage, Alharbe et al. [12] introduced a risk-based ABAC model tailored to dynamic data protection. By assessing subject, resource, and environment attributes, this model offers fine-grained security adjustments suitable for sensitive cloud data, such as medical records. Farhadighalati et al. [13] focused on human-centric access for Electronic Health Records (EHR), combining ABAC with risk assessments to secure healthcare data. Other notable advancements include MLCAC by Xiao et al. [14], a multi-layered model targeting insider threats through real-time monitoring and adaptive access decisions. For smart homes, Burakgazi et al. [15] extended ABAC by integrating biometric-based authentication, refining access control policies based on user verification scores. Zhong et al. [16] contributed to IoT edge security with SC-ABAC, a model using blockchain and smart contracts to manage access in decentralized environments. Lastly, Zhong et al. [17] enhanced RBAC with blockchain for dynamic, role-based access in data collaboration systems, achieving greater adaptability and security in organizational data access. These studies underscore the progression toward integrating ABAC with dynamic, context-aware technologies to strengthen security across IoT, cloud, and SDN platforms, demonstrating the effectiveness of hybrid models in complex digital environments.

III. Methodology

The methodology is structured to provide adaptive access permissions by analyzing user behavior and contextual data in real time. This approach enhances security by dynamically adjusting access permissions based on activity patterns, device information, and risk assessments derived from behavioral analytics. Fig. 1 illustrates the activity log workflow of the DACS, beginning with data collection and preprocessing, followed by feature extraction, model building, training, and risk assessment, ultimately leading to access control decisions and continuous monitoring.

First, Data Collection involves gathering user activity logs, which may include login times, device types, and other contextual data points. These logs, stored in SQL or NoSQL databases, are essential for understanding user behavior patterns and evaluating potential security risks. Next, Data Preprocessing is carried out to clean, convert, and standardize this raw data, ensuring its consistency and quality. In Python, for instance, preprocessing steps involve converting login times into a standardized datetime format using the pandas library, which simplifies temporal analysis. The following code snippet represents this step: defining a function that converts the 'login_time' column in user logs to a datetime format, ensuring consistency across entries.

This allows accurate analysis of time-based behavior patterns. Following preprocessing, Feature Extraction identifies key variables, such as login frequency, device type, and usage patterns. These extracted features provide a comprehensive profile of each user's typical behavior, which is essential for generating an accurate risk profile. Subsequently, a Behavioral Analytics Model is trained using machine learning frameworks like TensorFlow or PyTorch. This model, commonly a neural network, learns from historical user ac-

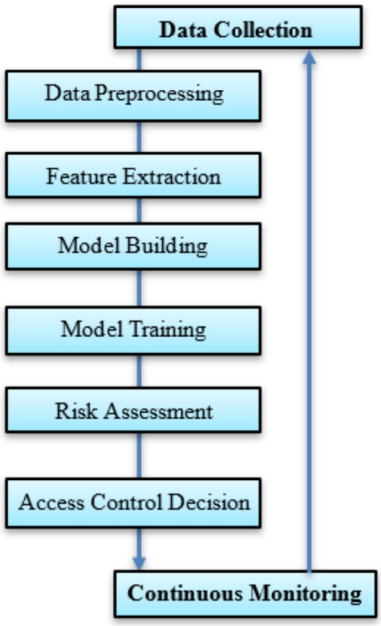


Fig 1. DACS Activity Log Workflow.

tivity data to compute a risk score for each access attempt. The risk score quantifies the likelihood that an access attempt is legitimate or anomalous. To enable real-time processing, Data Streaming is implemented using Apache Kafka, which streams user activity logs continuously to ensure the behavioral model is updated with the latest information. With this real-time data, DACS continuously evaluates and adjusts risk scores. When a user attempts to access a system, their behavior is compared to established patterns, and the Access Control Adjustment mechanism adapts permissions accordingly. For instance, if a high-risk score is detected (indicating unusual activity or potential threat), the system may restrict access or require additional verification. Conversely, low-risk scores allow seamless access, minimizing interruptions for legitimate users. The Visualization and Monitoring component completes the methodology, using tools like Matplotlib or Plotly to generate real-time graphs and reports. These visualizations provide system administrators with insights into user behavior trends and the effectiveness of the access control system. This comprehensive methodology ensures that DACS dynamically responds to potential threats, balancing security and usability in modern digital environments.

IV. System Implementation

The implementation of the system involves a layered approach that incorporates data collection, preprocessing, feature extraction, model development, and continuous monitoring to provide adaptive access permissions based on user behavior and contextual data. This architecture leverages machine learning techniques and real-time data processing to enhance security by dynamically adjusting access controls. Fig. 2 depicts the decision-making workflow in the DACS, starting from data collection and preprocessing through model building, training, and risk assessment. Based on real-time risk evaluation, the system adjusts access permissions dynamically, categorizing them as low, moderate, or high risk.

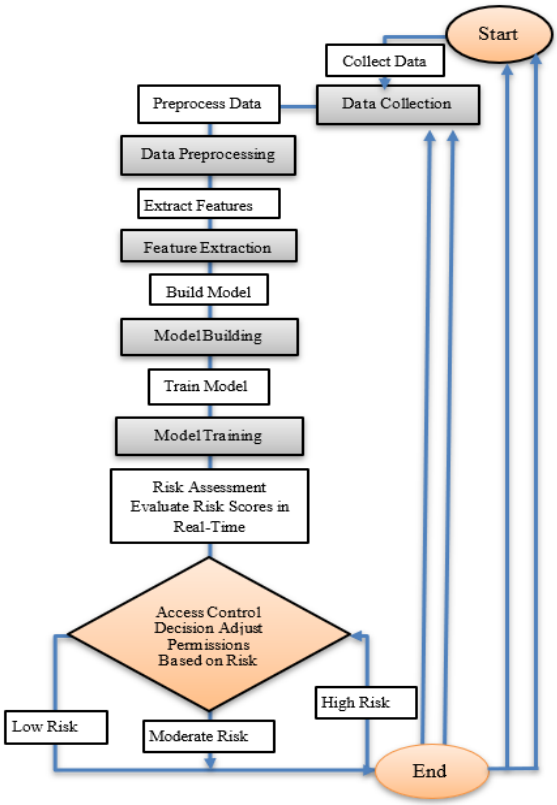


Fig 2. DACS Decision-Making Workflow.

Data collection forms the foundation, gathering user activity logs (e.g., login times, locations, devices) and contextual data (e.g., network conditions, time of access). This data, stored in SQL/NoSQL databases, enables behavioral pattern analysis and risk assessment. In Data Preprocessing, raw data is cleaned, normalized, and encoded, preparing it for model input. Feature Extraction then identifies key variables, such as time-based usage patterns and contextual risk factors, enhancing risk assessment capabilities. Model building involves developing a neural network using TensorFlow or PyTorch, with an input layer for features, hidden layers for complex pattern recognition, and an output layer that generates a risk score. The model is trained and optimized with metrics like accuracy and F1-score. Based on Risk Scoring, thresholds classify access attempts into low, moderate, or high-risk levels, with actions tailored accordingly. Access control adjustment dynamically modifies permissions in real time based on risk levels. Continuous monitoring updates user profiles and retrains the model as new data emerges. The system integrates software like Apache Kafka, SQL/NoSQL databases, and visualization tools, with cloud-based infrastructure and GPUs for scalability. In the Algorithm Workflow, data is streamed, preprocessed, and analyzed to produce risk-based access decisions, with actions

logged for auditing. Security is enforced through data encryption, secure authentication, and ISO/IEC 27001 compliance.

V. Results And Discussion

Fig. 3 illustrates the model's training and validation loss over 20 epochs, showing the model's learning progression. A decrease in training loss indicates the model's improved fit to the data, while fluctuations in validation loss reflect its generalization capacity on unseen data. Ideally, both losses should converge, with lower values signaling model effectiveness. However, substantial variance in validation loss suggests potential overfitting, necessitating adjustments.

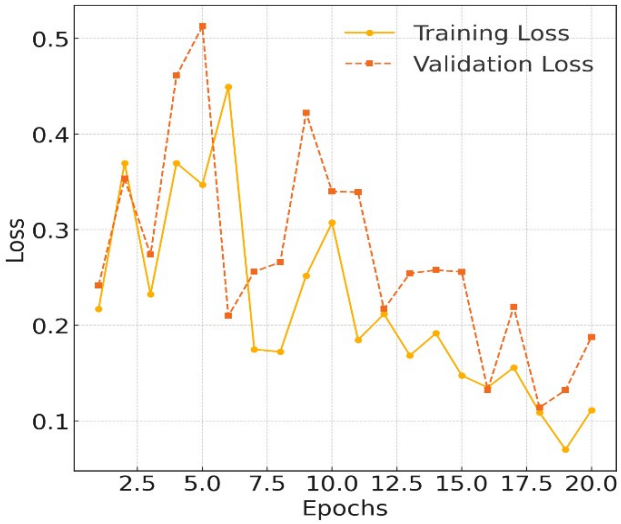


Fig 3. Training and Validation Loss over Epochs.

Fig. 4 shows the processing latency of system across 20 time intervals, indicating the response time for real-time data processing. Latency fluctuates between 100 ms and 300 ms, reflecting variations in system load and computational demand. Observing latency patterns helps identify any periods of high delay that could impact the system's responsiveness. These insights are essential for optimizing the model's efficiency and maintaining real-time adaptability.

Fig. 5 shows the Detection Accuracy comparison, with the static access control model at 70% and the DACS model at 87%. The reduced bar width highlights the difference in accuracy, demonstrating the improved detection capability of DACS. This comparison emphasizes DACS's advantage in accurately adapting to real-time behavioral changes, which enhances security by reducing the risk of unauthorized access.

Fig. 6 displays the throughput comparison between the static access control model and the DACS. Throughput, measured in requests per second, is significantly higher for the DACS model (220 requests/sec) compared to the static model (150 requests/sec). This improvement demonstrates DACS's efficiency in handling a larger volume of requests, essential for environments requiring rapid, real-time access control decisions.

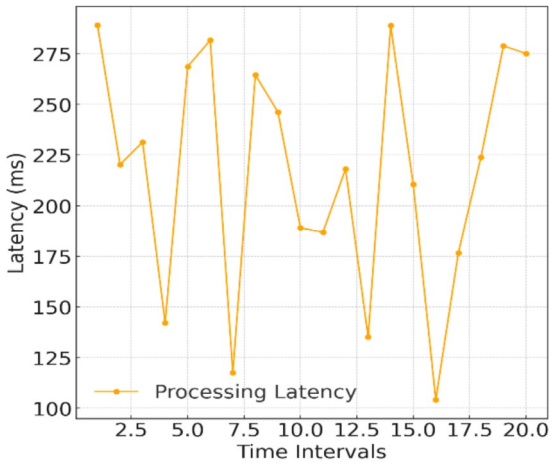


Fig 4. Real-Time Processing Latency over Time Intervals.

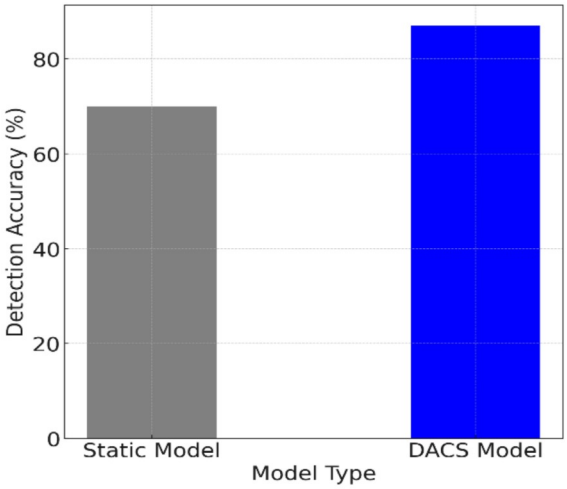


Fig 5. Detection Accuracy Comparison of Static and DACS Models.

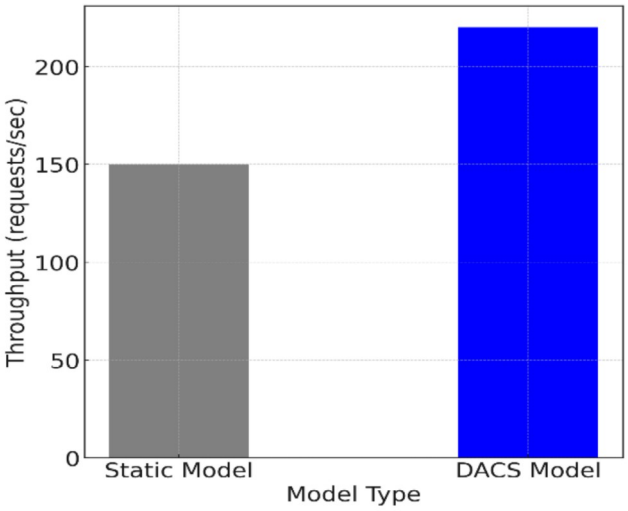


Fig 6. Throughput Comparison of Static and DACS Models.

VI. Conclusion

The DACS enhances cybersecurity by leveraging real-time behavioral analytics to adapt access permissions based on user activity, contextual data, and risk assessment. DACS gathers detailed user activity logs, applies preprocessing techniques like normalization and encoding, and extracts crucial features related to time, behavior patterns, and contextual risk factors. These features are processed through a neural network model built on frameworks such as TensorFlow or PyTorch to compute dynamic risk scores for each access attempt. Based on these scores, DACS adjusts permissions: low-risk scores allow seamless access, moderate-risk scores prompt additional authentication, and high-risk scores restrict access. Continuous data streaming via Apache Kafka ensures that real-time behavioral changes are promptly reflected in access decisions, allowing for adaptive, risk-based responses. This system showcases scalability and responsiveness, offering a practical, intelligent solution to evolving cybersecurity threats. DACS demonstrates the efficacy of integrating machine learning with access control, providing a robust and flexible security framework for modern, high-demand environments.

References

[1] Ameer, Safwa, James Benson, and Ravi Sandhu. "An attribute-based approach toward a secured smart-home IoT access control and a comparison with a role-based approach." *Information* 13, no. 2 (2022): 60.

[2] Ameer, Safwa, James Benson, and Ravi Sandhu. "Hybrid approaches (ABAC and RBAC) toward secure access control in smart home IoT." *IEEE Transactions on Dependable and Secure Computing* 20, no. 5 (2022): 4032-4051.

[3] Burakgazi Bilgen, Melike, Osman Abul, and Kemal Bicakci. "Authentication-enabled attribute-based access control for smart homes." *International Journal of Information Security* 22, no. 2 (2023): 479-495.

[4] Ameer, Safwa. "User-To-Device Access Control Models for Cloud-Enabled IoT with Smart Home Case Study." PhD diss., The University of Texas at San Antonio, 2021.

[5] Ameer, Safwa, James Benson, and Ravi Sandhu. "The EGRBAC model for smart home IoT." In *2020 IEEE 21st International Conference on Information Reuse and Integration for Data Science (IRI)*, pp. 457-462. IEEE, 2020.

[6] Huang, Haoxiang, Jianbiao Zhang, Jun Hu, Yingfang Fu, and Chenggang Qin. "Research on distributed dynamic trusted access control based on security subsystem." *IEEE Transactions on Information Forensics and Security* 17 (2022): 3306-3320.

[7] Kim, Hwimin, Dae-Kyoo Kim, and Alaa Alaerjan. "ABAC-based security model for DDS." *IEEE Transactions on Dependable and Secure Computing* 19, no. 5 (2021): 3113-3124.

[8] Vijayanand, S., and S. Saravanan. "A deep learning model based anomalous behavior detection for supporting verifiable access control scheme in cloud servers." *Journal of Intelligent & Fuzzy Systems* 42, no. 6 (2022): 6171-6181.

[9] Liu, Yifan, Bo Zhao, Yang An, and Jiabao Guo. "DACAS: integration of attribute-based access control for northbound interface security in SDN." *World Wide Web* 26, no. 4 (2023): 2143-2173.

[10] Gong, Qinghua, Jinnan Zhang, Zheng Wei, Xinmin Wang, Xia Zhang, Xin Yan, Yang Liu, and Liming Dong. "SDACS: Blockchain-Based Secure and Dynamic Access Control Scheme for Internet of Things." *Sensors* 24, no. 7 (2024): 2267.

[11] Alazab, Moutaz, Albara Awajan, Hadeel Alazzam, Mohammad Wedyan, Bandar Alshawi, and Ryan Alturki. "A novel IDS with a dynamic access control algorithm to detect and defend intrusion at IoT nodes." *Sensors* 24, no. 7 (2024): 2188.

[12] Alharbe, Nawaf, Abeer Aljohani, Mohamed Ali Rakrouki, and Mashael Khayyat. "An access control model based on system security risk for dynamic sensitive data storage in the cloud." *Applied Sciences* 13, no. 5 (2023): 3187.

[13] Farhadighalati, Nastaran, Jose Barata, Sanaz Nikghadam-Hojjati, and Eda Marchetti. "Behavioral and Human-Centric Access Control Model in XACML Reference Architecture: Design and Implementation of EHR Case Study." In *Technological Innovation for Human-Centric Systems: 15th IFIP WG 5.5/SOCOLNET Advanced Doctoral Conference on Computing, Electrical and Industrial Systems, DoCEIS 2024, Caparica, Portugal, July 3-5, 2024, Proceedings*, vol. 716, p. 192. Springer Nature, 2024.

[14] Xiao, Lifang, Aimin Yu, Hanyu Wang, Lixin Zhao, and Dan Meng. "MLCAC: Dynamic Authorization and Intelligent Decision-making towards Insider Threats." In *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pp. 407-412. IEEE, 2024.

[15] Burakgazi Bilgen, Melike, Osman Abul, and Kemal Bicakci. "Authentication-enabled attribute-based access control for smart homes." *International Journal of Information Security* 22, no. 2 (2023): 479-495.

[16] Zhonghua, Chen, S. B. Goyal, and Anand Singh Rajawat. "Smart contracts attribute-based access control model for security & privacy of IoT system using blockchain and edge computing." *The Journal of Supercomputing* 80, no. 2 (2024): 1396-1425.

[17] Zhong, Tao, Junsheng Chang, Peichang Shi, Linhui Li, and Fei Gao. "Dyacon: Jointcloud dynamic access control model of data security based on verifiable credentials." In *2021 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*, pp. 336-343. IEEE, 2021.

rice2024 Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering pp. 157–161
DOI: 10.15439/2024R54
ISSN 2300-5963 ACSIS, Vol. 42

Implementation of a Facial Recognition System for Attendance Tracking Utilizing the K-Nearest Neighbors Algorithm

Abhishek Tripathi
Department of Computer Science Engineering
Kalasalingam Academy of Research and Education
Srivilliputhur, Tamil Nadu, India
tripathi.abhishek.5@gmail.com

Chirukuri Manohar
Department of Computer Science Engineering
Kalasalingam Academy of Research and Education
Srivilliputhur, Tamil Nadu, India
manoharchirukuri09@gmail.com

Dhiraj K Patel
Department of Electronics and Communication Engineering
Indian Institute of Information Technology
Surat, India
er.dhirajpatel@gmail.com

Subhashish Tiwari
GITAM University
Bengaluru, Karnataka, India
stiwari@gitam.edu

Rajat Paliwal
Thadomal Shahani Engineering College
Mumbai, Maharashtra, India
rajat.paliwal@thadomal.org

***Abstract*—In today's digital age, facial recognition systems are crucial across industries for authentication, security, and identity verification. While slightly less precise than iris or fingerprint recognition, facial recognition's non-invasive nature fuels its growing popularity. It's extensively used for attendance tracking in various institutions, replacing error-prone manual processes. The proposed framework involves four stages: attendance updating, face detection, recognition, and database construction, employing techniques like Local Binary Patterns and Haar-Cascade classifier. Notably, in the recognition stage, the K-Nearest Neighbors (KNN) algorithm plays a pivotal role. KNN aids in accurately identifying individuals based on facial features, ensuring precise attendance tracking. Attendance records are then emailed to relevant faculty members at the end of each class, streamlining administrative tasks.**

***Index Terms*—Facial Recognition, KNN, Attendance System, PCA and LDA.**

I. Introduction

MANY schools and universities find that recording attendance using the traditional way is an arduous undertaking [1]. Additionally, it adds to the workload for the teachers who have to personally call the names of the pupils, which may take the full five minutes meeting. This consumes a significant amount of time and may result in proxy attendance. As a result, numerous institutions began utilizing a wide range of additional methods to record attendance, such as Radio Frequency Identification (RFID), iris identification, and biometric identification [2]. However, these systems may utilize more resources and have an invasive quality due to reliance on queues. Face recognition, on the other hand, offers a non-intrusive and readily attainable biometric characteristic [3]. Face recognition systems can be categorized into two main categories: face verification, which compares faces in a 1:1 matching method, and face identification, which involves comparing a query face picture with multiple template face images (1:N matching) [4]. The goal of this system is to create an efficient attendance system using face recognition algorithms as its foundation. Face recognition technology has been increasingly applied in various fields and offers promising solutions for attendance management [5]. In this work, we propose a method that recognizes children's faces from live streaming video in the classroom, making attendance recording quicker and more efficient compared to traditional methods.

The paper is organized as follows: Section II presents a Literature Review summarizing existing approaches to facial recognition and attendance systems. Section III details the Proposed System, including its architecture and components. Section IV explains the Methodology, covering data preprocessing, feature extraction, and the K-Nearest Neighbors algorithm. Section V discusses the Results and Discussions, analyzing system performance and accuracy. Finally, Section VI concludes with key findings and future directions in Conclusion.

II. Literature Review

Researchers have introduced various models for automatic attendance systems, with a focus on integrating radio frequency identification (RFID) with facial recognition [5]. RFID is utilized to identify and count approved pupils as they enter and exit the classroom, maintaining an authentic record of enrolled students [6]. Additionally, the system retains information about each student enrolled in a particular course in the attendance record and provides the necessary information as required. Furthermore, attendance systems based on biometric iris data have been implemented, where participants register their information and provide an original iris template [7]. During attendance recording, the system automatically captures the attendee's image, identifies them through their iris, and matches them with the database.

Another proposed attendance system involves face identification, employing methods such as Viola-Jones and Features from the Histogram of Oriented Gradients (HOG), along with a Support Vector Machine (SVM) classifier [3]. This system addresses various real-time situations, including scale, lighting, occlusions, and posture. A quantitative study conducted using Peak PSNR values in MATLABGUI revealed that Eigenface yielded superior outcomes compared to Fisher face [8]. Moreover, a technique for tracking student attendance using facial recognition technology in the classroom was proposed, integrating Discrete Wavelet Transforms (DWT) with Discrete Cosine Transforms (DCT) to extract facial features [9]. Radial Basis Function (RBF) was then utilized to classify facial features, achieving an accuracy percentage of 82% [4].

III. Proposed System

Each student in the class is required to register by providing necessary information. Subsequently, their pictures will be taken and stored in the dataset. Faces will then be identified from the live-streamed classroom footage during each session. These identified faces will be compared to the pictures in the dataset. If a match is found, attendance will be recorded for the corresponding student. After each session, a list of absentees will be emailed to the faculty member in charge of the meeting. The system architecture of the proposed system is provided below. Generally, this process consists of four steps.

A. Creation of Datasets

A webcam is used to capture images of the children, with each student being photographed multiple times while displaying various movements and poses. These images undergo preprocessing steps, which include trimming to extract the Region of Interest (ROI) for subsequent recognition processes. The cropped photos are then resized to specific pixel dimensions. Following this, the RGB images are converted to grayscale. Finally, the processed images are saved in a file along with the corresponding student's name. Fig. 1 illustrates the system architecture designed for a face recognition attendance system.

B. Facial Recognition

In this context, facial detection is achieved using the Haar-Cascade Classifier provided by OpenCV. Prior to its application in facial recognition, the Haar Cascade method requires training to identify individuals' faces, a process known as feature extraction. This involves utilizing an XML file named "haar cascade_frontal face_default" as training data, which contains the cascade of Haar characteristics essential for recognition. Fig. 2 depicts a theoretical face model utilizing Haar Features for robust facial recognition. In this instance, the OpenCV detect MultiScale module is utilized to surround the faces in an image with rectangles, which is necessary for detection. Three criteria need to be considered: scale Factor, min Size, and min Neighbors. The scale Factor determines how much each image scale should

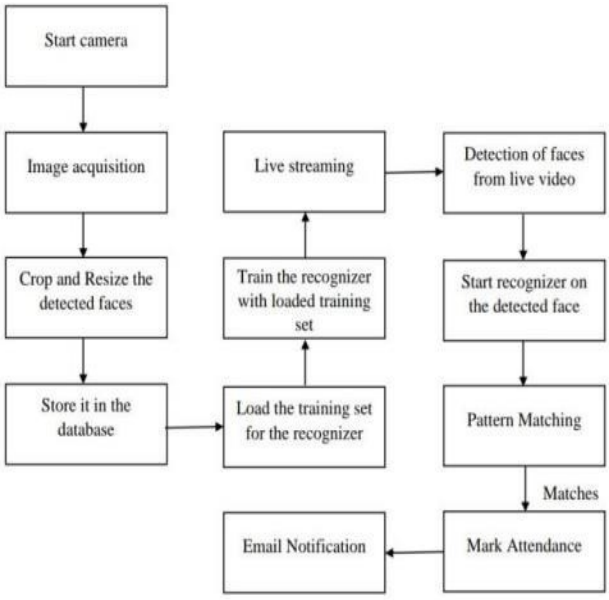


Fig 1. System architecture for face recognition attendance system.

be reduced. The min Neighbors parameter specifies how many neighboring rectangles each candidate should have. Higher ratings typically result in fewer detected faces but better recognition quality. The min Size parameter indicates the minimal size of an object, which is set to (30,30) by default. In this system, the parameters scale Factor and min Neighbors are set to 1.3 and 5, respectively.

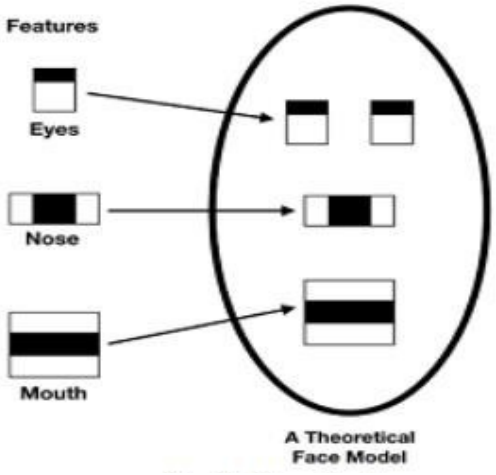


Fig 2. Theoretical face model utilizing Haar Features for robust facial recognition.

The face recognition process comprises three steps: preparation of training data, training of the face recognizer, and prediction. The photographs in the dataset serve as the training data, and each is assigned an integer label indicating the corresponding student. Subsequently, face recognition is applied to these images. This system employs a Local Binary Pattern Histogram as the face recognizer. Initially, the

local binary patterns (LBPs) of the entire face are obtained and converted into decimal numbers. Histograms are then created for each decimal value. For every image in the training set, a histogram is generated. During the recognition process, the histogram of the face to be identified is computed and compared with the previously computed histograms to determine the label that best matches the corresponding student.

C. Attendance Updation

Following the face recognition procedure, the faces that were identified will be noted as present on the excel sheet, while the remaining faces will be noted as absent. A list of the absentees will then be mailed to that particular faculty. Faculty members will be informed with monthly attendance record at each month's conclusion.

IV. Methodology

Throughout our face recognition endeavor, we extensively explored a myriad of methodologies prevalent in the field. This encompassed a thorough examination of existing literature to discern the strengths and limitations of various recognition systems. Our chief aim was to address the shortcomings of prevailing approaches and devise a proficient face recognition system. At the core of our investigation lay the fusion of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) for feature extraction. We employed Eigenfaces and Fisherfaces for subspace projection, with matching facilitated by a Euclidean classifier for distance assessment. Diverse strategies for facial recognition were explored, all yielding favorable results. Our attention extended to real-time applications, with PCA demonstrating exceptional performance in this realm. Insights gleaned from literature emphasized the necessity of refining the system to accommodate variations in facial angles, underscoring our dedication to bolstering resilience and adaptability. An intriguing avenue explored involved the potential amalgamation of gait recognition with facial recognition software, offering a comprehensive approach to identity verification. Challenges posed by low-light conditions were acknowledged, with recognition that while the system functions adequately in such scenarios, achieving optimal resolution remains an ongoing endeavor. Fig. 3 delineates a flowchart delineating the methodology adopted in the face recognition system.

Firstly, the algorithm begins by importing necessary libraries such as numpy for numerical operations, face_recognition for face detection and recognition, cv2 for computer vision tasks, os for file and directory operations, datetime for handling date and time data, and xlsxwriter for generating Excel files. Next, the algorithm defines two key functions. The first function, findEncodings(images), processes a list of images by converting them to RGB format, detecting faces within the images using the face_recognition library, and then appending the encodings of these faces to a list. The second function, markAttendance(name), handles the attendance marking process by reading and writing to a CSV file containing attendance data, ensuring each student's attendance is appropriately recorded along with the current date and time.

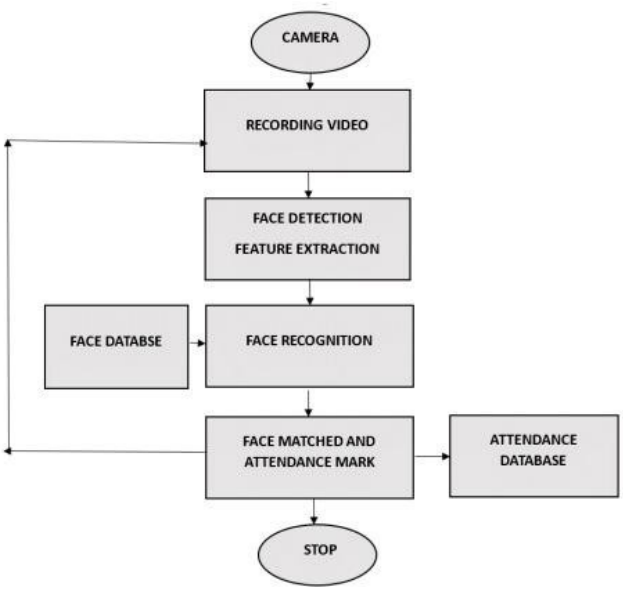


Fig 3. Flow chart of methodology.

Subsequently, the algorithm defines a dictionary containing student names and their corresponding roll numbers. This data will be used to populate the Excel worksheet with student information, ensuring accurate attendance tracking. Following that, the algorithm loads images of students from a specified directory, extracting class names from the filenames to identify each student. This step prepares the system to recognize known faces during the attendance marking process. Then, the algorithm encodes the known faces using the findEncodings function, which prepares them for comparison with faces detected in real-time video frames. This encoding step is crucial for accurate face recognition during the attendance tracking process. Lastly, the algorithm captures video frames from the webcam, continuously processing each frame to detect faces, compare them with known faces, and update the attendance status accordingly in real-time. Once a specified number of frames have been processed, the video capture object is closed, and the attendance data is saved in the Excel workbook for further analysis and record-keeping.

V. Results And Discussions

Through a GUI, users can communicate with the system. Users will primarily have access to three options here: mark registration, faculty registration, and student registration. presence. It is expected of the students to participate in all the necessary information on the student registration form. Upon When you click the "Register" button, the webcam launches immediately. The window, as seen in Fig. 4, appears and begins to detect the picture's faces. Then it begins to click on its own. pictures up until CRTL+Q is pushed or

60 samples are gathered. After that, these photos will be kept and pre- processed in folder for training pics.

The faculties are supposed to register with the respective course codes along with their email-id in the faculty registration form provided. This is important because the list of absentees will be ultimately mailed to the respective faculties.

Fig 4. Face detection sample as captured picture of some students.

Each session, the appropriate faculty member needs to input their course code. The camera will then turn on by itself after the course code has been entered. Fig. 5 displays face recognition technology. window that shows the names of two enrolled pupils and if they hadn't registered, it would have been evident "Unknown." You can close the window by hitting CTRL+Q. and names will be entered in the Excel file along with attendance of absentees to the appropriate faculty member by letter.

Fig 5. Live facial recognition.

Fig. 6 displays the updated attendance sheet following the procedure of recognition. The marking system assigns a value of '1' to recognized students and '0' to absentee students. The absences list will be mailed to the relevant faculty member's email address.

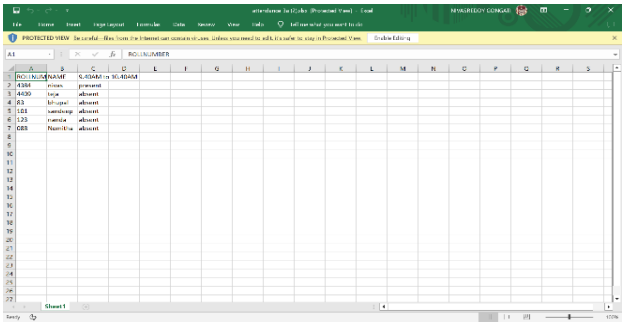


Fig 6. Collected data in attendance sheet.

Fig.7 presents a comparison between the actual and predicted values obtained from a K Neighbors Regressor model with $K = 5$ K=5 and uniform weights.

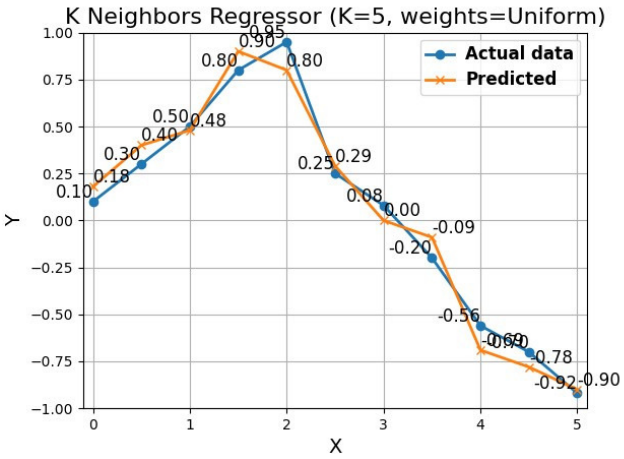


Fig 7. Comparison of actual and predicted values generated by a K Neighbors Regressor model.

As indicated in Table 1, we split the datasets into training and test sets in order to conduct the experiment. The results of the experiments indicate that LDA provides a higher recognition rate than that of PCA, as the graphics demonstrates.

TABLE1 DESCRIPTION OF DATASETS

Dataset	Total Image	Individuals	Training image	Training image
ORL Dataset	400	40	120	280
Class Dataset	25	5	10	15

The bar chart in Fig. 8 depicts the frequency of recognition results for a facial recognition system. It compares the recognition of correct faces against instances of false recognition across different total 10 faces analyzed.

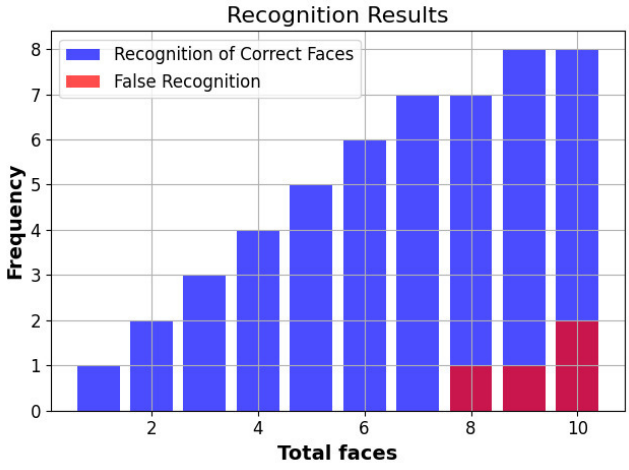


Fig 8. Comparison of correct and false facial recognition frequencies.

VI. Conclusion

Our observations revealed significant insights, showcasing the commendable performance of both PCA and LDA in specific scenarios, including normal lighting conditions, consistent posture, and an optimal camera distance of 1-3 feet. Notably, a higher resolution is imperative for precise pixel-by-pixel operations during calculation. While PCA exhibited a shorter recognition time compared to LDA, the latter demonstrated superior recognition capabilities, aligning with our research goal of optimizing accuracy. This preference for LDA underscores its crucial role in face recognition systems, essential for identifying unknown individuals. Moving forward, further investigation into the algorithm's recognition capabilities is warranted. In summary, our research offers a comprehensive exploration of face recognition techniques, highlighting strengths, limitations, and innovative solutions to enhance system performance. Our commitment to future studies and the integration of multi-modal recognition approaches reflects our forward-looking approach to continuous system refinement.

References

[1] Balcoh, Naveed Khan, M. Haroon Yousaf, Waqar Ahmad, and M. Iram Baig. "Algorithm for efficient attendance management: Face recognition based approach." *International Journal of Computer Science Issues (IJCSI)* 9, no. 4 (2012): 146.

[2] Kar, Nirmalya, Mrinal Kanti Debbarma, Ashim Saha, and Dwijen Rudra Pal. "Study of implementing automated attendance system using face recognition technique." *International Journal of computer and communication engineering* 1, no. 2 (2012): 100-103.

[3] Wagh, Priyanka, Roshani Thakare, Jagruti Chaudhari, and Shweta Patil. "Attendance system based on face recognition using eigen face and PCA algorithms." In *2015 International Conference on Green Computing and Internet of Things (ICGCIoT)*, pp. 303-308. IEEE, 2015.

[4] Patil, A. M., Satish R. Kolhe, and Pradeep M. Patil. "Face recognition by PCA technique." In *2009 Second International Conference on Emerging Trends in Engineering & Technology*, pp. 192-195. IEEE, 2009.

[5] Karunakar, M., C. A. Sai, K. Chandra, and K. ANIL Kumar. "Smart Attendance Monitoring System (SAMS): A Face Recognition Based Attendance System for Classroom Environment." *Int. J. Recent Develop. Sci. Technol* 4, no. 5 (2020): 194-201.

[6] Barr, Jeremiah R., Kevin W. Bowyer, Patrick J. Flynn, and Soma Biswas. "Face recognition from video: A review." *International journal of pattern recognition and artificial intelligence* 26, no. 05 (2012): 1266002.

[7] Pooja, G. S., and K. S. Rekha. "Automatic Attendance System Using Artificial Intelligence." In *2022 IEEE 3rd Global Conference for Advancement in Technology (GCAT)*, pp. 1-4. IEEE, 2022.

[8] Singh, Sanjay K., Mayank Vatsa, Richa Singh, and D. S. Chauhan. "A comparison of face recognition algorithms neural network based & line based approaches." In *IEEE International Conference on Systems, Man and Cybernetics*, vol. 6, pp. 6-pp. IEEE, 2002.

[9] Paul, Liton Chandra, and Abdulla Al Sumam. "Face recognition using principal component analysis method." *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)* 1, no. 9 (2012): 135-139.

rice2024 Proceedings of the Ninth International Conference on Research in Intelligent Computing in Engineering pp. 163–166
DOI: 10.15439/2024R103
ISSN 2300-5963 ACSIS, Vol. 42

Breastfeeding, HAMLET and AI: Exploring Synergies for Breast Cancer Prevention in Future Prospect

K. L. Vasundhara
Head of Mathematics Department
Stanley College of Engineering and Technology for Women
Hyd, India
vasundhara.yerasuri@gmail.com

Harshita Vyas
Computer Science Engineering
Stanley College of Engineering and Technology for Women
Hyd, India
vyas.harshita004@gmail.com

Indaram Sri Charitha
Electronics & Communication Engineering
Stanley College of Engineering and Technology for Women
Hyd, India
indaramsricharitha2903@gmail.com

***Abstract*—In this study, we analyze the association between breastfeeding practices, the bioactive compound HAMLET (Human Alpha-lactalbumin Made Lethal to Tumor Cells), and the potential mitigation of breast cancer risk. The primary risk factors for breast cancer are a woman's age and family history, particularly the presence of a first-degree relative with breast cancer. Women who have a history of Breastfeeding have demonstrated reduced incidence rates of breast cancer. A key component of human milk, alpha-lactalbumin, forms a complex with oleic acid and selectively induces apoptosis in tumour cells while sparing normal cells. This unique property positions HAMLET as a promising agent for cancer prevention. This paper also examines the potential of artificial intelligence to build predictive models on the risk of future breast cancer in relation to extensive maternal health data and breastfeeding practices. Even though HAMLET has yet to enter the clinic, its distinct properties make it a good preventative drug against the risk of developing breast cancer. This research emphasizes the integration of HAMLET into future research frameworks and AI-based solutions for the advancement of personalized strategies for breast cancer prevention.**

***Index Terms*—Breast cancer, Breastfeeding, Artificial Intelligence, Hamlet.**

I. Introduction

BREAST cancer is the most common gynaecological tumour in young women, the second most common cancer worldwide and the most frequently diagnosed cancer among women [1]. The relationships between Breastfeeding and the development of a number of chronic diseases, including obesity, diabetes and breast cancer, have been extensively studied [2]. Support and advice should be routinely available during antenatal care to help mothers initiate Breastfeeding at the time of birth and to ensure that Breastfeeding is fully established during the postnatal period [3]. Worldwide, it is estimated that only 34.8% of infants are exclusively breastfed for the first 6 months of life, while the majority receive other types of food or fluid during their early months [4]. Breastfeeding is a cornerstone of maternal and child health, providing essential nutrients and immune protection. There is evidence that human milk may confer long-term benefits, such as reduced risk of certain autoimmune diseases, inflammatory bowel disease and certain malignancies [3].

HAMLET, a bioactive compound in human milk, has garnered attention for its unique ability to target and kill tumour cells without harming healthy cells. Several mouse breast cancer models have been developed to define a prototypic strategy for prophylactic cancer vaccination in which alpha-lactalbumin was chosen as the target vaccine autoantigen because it is a breast-specific differentiation protein that is expressed at high levels in the vast majority of human breast carcinomas and mammary epithelial cells only during lactation. Immunoreactivity against alpha-lactalbumin provides substantial protection against the growth of autochthonous tumours in transgenic mouse models of breast cancer and against 4T1 Transplantable breast tumours in BALB/c mice. Because alpha-lactalbumin is conditionally expressed only during lactation, vaccination-induced prophylaxis occurs without any detectable inflammation in normal nonlactating breast tissue. Thus, alpha-lactalbumin vaccination may provide safe and effective protection against the development of breast cancer for women in their post-childbearing, premenopausal years, during which lactation is readily avoidable and the risk of developing breast cancer is high [5]. Despite its promising properties, it has not been adopted clinically due to limited data and application frameworks. This paper explores HAMLET as a novel hypothetical approach to reducing breast cancer risk and assessing breast cancer susceptibility based on breastfeeding data.

II. Related Work

Global cancer cases and deaths have been predicted in accordance with past available data. There is a rapid increase in number of cancer cases by the end of 2025 which will be around 19 million and it is probable to go further in near future. Though there seems to be huge gap between mortality rate and cancer cases but is prone to upsurge at any given point of time in future.

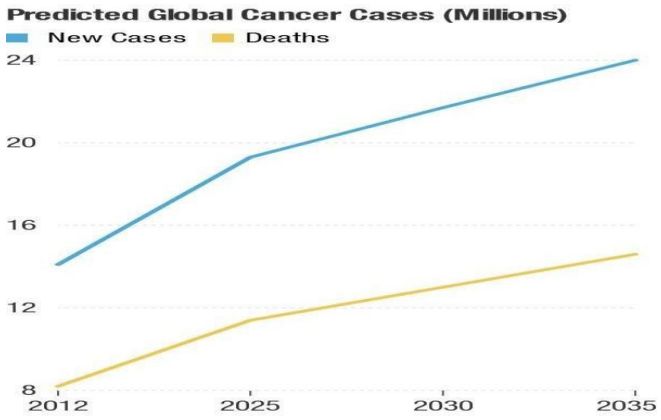


Fig. 1. Year wise total cancer prevalence and prediction in the world

A survey was conducted among 300 cancer patients across various hospitals in Hyderabad, India, using a structured questionnaire. The data collected from the survey were analyzed using statistical methods, and the results indicate that breastfeeding plays a significant role in reducing cancer risk in women. Women who breastfed their children demonstrated a lower risk of cancer compared to those who did not breastfeed, highlighting the importance of breastfeeding for maternal health [16].

III. Methodology

A. Data Collection and Sample Population

The study focused on gathering data from women under 40 years of age who have at least one child, specifically assessing key parameters such as breastfeeding duration, practices, and the incidence of breast cancer. A community-based descriptive cross-sectional study was conducted in the Egor Local Government Area of Edo State, Nigeria, where a sample of 418 mothers was surveyed. The findings revealed that only 44.5% of mothers initiated Breastfeeding within the first hour after delivery, and the prevalence of exclusive Breastfeeding was recorded at 36.6%[6]. This highlights significant gaps in early breastfeeding practices, which are critical for both maternal and infant health. Moreover, a hospital-based cross-sectional study in Bankura, West Bengal, India, involving 400 mothers, found that 36% initiated Breastfeeding within an hour and 53% practised exclusive Breastfeeding [7]. These studies underscore the need for targeted interventions to improve breastfeeding initiation and duration among young mothers. The median duration of Breastfeeding reported in the Egor study was approximately 15.1 months, indicating a relatively positive trend in sustained breastfeeding practices[6].

However, the introduction of prelacteal feeds before six months was noted in a significant proportion of cases, which can detract from the benefits of exclusive Breastfeeding [7]. Finally, these findings emphasize the importance of enhancing maternal education and support systems to promote better breastfeeding practices among women under 40 years old. Addressing these issues could potentially reduce health risks associated with inadequate breastfeeding practices, including a possible increase in breast cancer incidence linked to shorter Breastfeeding durations[8]. The below bar graph illustrates breastfeeding patterns categorized by duration. It compares exclusive and mixed feeding percentages for durations of less than 6 months, 6-12 months, 12-24 months, and more than 24 months. This visualization highlights trends in breastfeeding practices across different timeframes. The breastfeeding duration data used for this analysis and graph generation was derived from the table presented in [15].

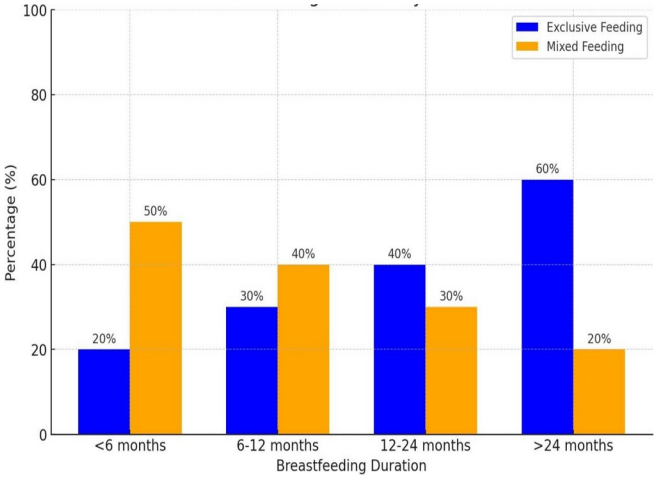


Fig.2. Illustration of Breastfeeding Patterns by Duration

B. HAMLET Analysis and Statistical Evaluation

In this section, we detail the analytical techniques used to measure HAMLET (human α-lactalbumin made lethal to tumor cells) levels in breast milk, focusing on high-performance liquid chromatography (HPLC) and statistical evaluations. HAMLET, a bioactive complex formed from α-lactalbumin and oleic acid, exhibits cytotoxic properties against tumor cells, particularly in breast cancer. Recent advancements in biochemical methods, including the precision of HPLC, have allowed for accurate quantification of HAMLET, contributing to our understanding of its role in cancer prevention [8] [9].HPLC was utilized to quantify HAMLET levels in breast milk, with a comparative analysis conducted across varying breastfeeding durations—less than six months, 6–12 months, and beyond 24 months. Studies have consistently indicated that extended breastfeeding correlates with higher HAMLET concentrations, which may play a role in reducing breast cancer risk [10][11]. Statistical models, including regression and multivariate analyses, were employed to explore the interplay between breastfeeding duration, HAMLET levels, and breast cancer outcomes.

These techniques revealed that exclusive breastfeeding for longer periods significantly enhances the presence of HAMLET, supporting its potential as a protective factor against cancer [12][13].

The findings suggest that prolonged breastfeeding positively influences HAMLET concentrations in breast milk,

aligning with broader evidence on breastfeeding's health benefits for both mother and child. These results also underscore the importance of biochemical properties of breast milk in shaping long-term health outcomes, particularly in reducing breast cancer risk [14][15]. The study highlights the necessity of integrating advanced biochemical analyses like HPLC with robust statistical evaluations to understand the protective mechanisms of breastfeeding and bioactive milk components.

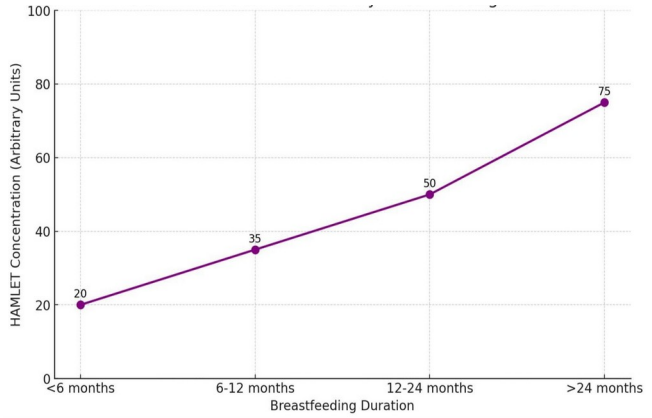


Fig.3. Illustration of HAMLET levels in Breast Milk by analyzing Breastfeeding duration.

Next, we analyze and illustrate the correlation between breastfeeding duration and the likelihood of breast cancer prevention, demonstrating an increasing trend in prevention probability with longer breastfeeding durations. The confusion matrix visualizes a classification example for predicting adequate breastfeeding practices and their relation to cancer prevention. Here, the labels represent:

1. Adequate Breastfeeding (>24 months)
2. Inadequate Breastfeeding (<24 months)

This demonstrates the potential accuracy of methods in identifying patterns from breastfeeding data for preventive insights.

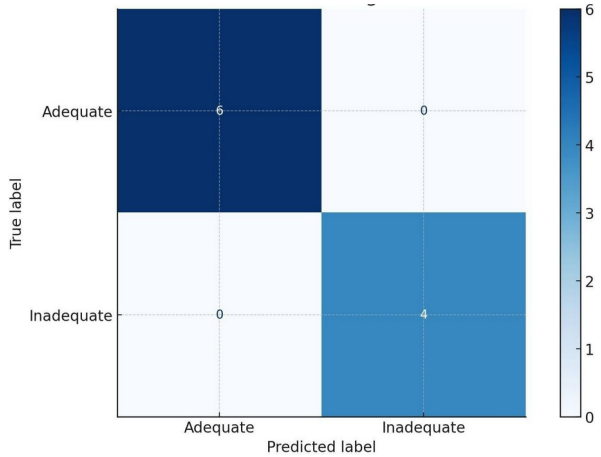


Fig.3. Confusion Matrix for Breastfeeding and cancer prevention

C. *Architecture workflow*

The methodology architecture begins with **data collection**, targeting women under 40 with at least one child in two study locations: Egor (Nigeria) for community-based insights and Bankura (India) for hospital-based findings. Surveys and structured interviews capture breastfeeding initiation times, exclusive breastfeeding durations, and prelacteal feeding practices. Next, **breastfeeding pattern analysis** involves categorizing and visualizing trends in exclusive and mixed feeding across durations such as <6 months, 6–12 months, and >24 months, highlighting gaps in optimal practices. In the **HAMLET biochemical analysis** phase, breast milk samples are processed using High-Performance Liquid Chromatography (HPLC) to quantify HAMLET levels, comparing concentrations across different breastfeeding durations. **Statistical evaluation** follows, using regression and multivariate analyses to establish relationships between breastfeeding duration, HAMLET levels, and breast cancer outcomes. The final step integrates insights, revealing that prolonged breastfeeding enhances HAMLET concentrations, potentially reducing breast cancer risk, and emphasizing the need for maternal education and biochemical research to promote better health outcomes.

Fig.4. Methodology architecture (flowchart)

D. *Exploring the role of AI*

Objectives of integrating AI -

1. Utilize AI models to predict breast cancer risk by analyzing maternal health records, breastfeeding history, genetic markers, and lifestyle factors.
2. Address the unavailability of HAMLET datasets by conducting small-scale studies and simulating lab-based data for AI training.
3. Leverage the use of AI in predicting patient responses to HAMLET therapy for breast cancer treatment.
4. Propose methodologies to bridge the gap between experimental HAMLET research and practical clinical applications.

This study explores two key objectives regarding the role of HAMLET in breast cancer management. Firstly, it investigates how breastfeeding practices, particularly exclusive breastfeeding for two years or more, correlate with increased

HAMLET levels in breast milk and a notable reduction in breast cancer risk, emphasizing its preventive potential. Secondly, it proposes utilizing HAMLET as a novel therapeutic approach for breast cancer treatment, aiming to develop AI-based predictive models to personalize therapy. These models would analyze clinical, genetic, and molecular data to predict patient responses to HAMLET therapy, optimize treatment cycles, and minimize side effects. However, due to the unavailability of HAMLET-related datasets and its limited clinical application thus far, this study serves as a conceptual framework to guide future research in this domain.

E. Breastfeeding diminishes the risk of breast cancer

Ten Steps to Successful Breastfeeding

Each facility delivering maternity services and caring for newborn infants should:

1. Maintain a documented breastfeeding policy consistently communicated to all healthcare staff.
2. Provide comprehensive training to healthcare personnel, equipping them with the necessary skills for policy implementation.
3. Educate expectant mothers on the advantages of breastfeeding and its proper management.
4. Facilitate the initiation of breastfeeding within thirty minutes of childbirth.
5. Instruct mothers on breastfeeding techniques and guide them on maintaining lactation, even in cases of temporary separation from their infants.
6. Administer only breast milk to newborns, refraining from other food or drink unless medically necessary.
7. Implement rooming-in practices, allowing mothers and infants to stay together around the clock.
8. Promote breastfeeding on demand, encouraging mothers to feed their infants as needed.
9. Refrain from employing artificial teats, pacifiers, dummies, or soothers for breastfeeding infants.
10. Encourage the formation of breastfeeding support groups and offer referrals to mothers upon their release from the hospital or clinic.[17]

IV. Conclusion

This study underscores the significant relationship between breastfeeding patterns, duration, and breast cancer prevention. The findings reinforce the hypothesis that prolonged Breastfeeding is associated with a significantly lower risk of breast cancer in both mothers and their children. The women who breastfed exclusively for a minimum two years exhibited a notable reduction in breast cancer prevalence, potentially attributed to elevated HAMLET levels in their milk. HAMLET, a bioactive compound with tumoricidal properties, demonstrates a compelling role in inducing apoptosis in cancer cells while sparing healthy tissue, thus highlighting its importance in cancer prevention. The findings emphasize the crucial role of extended and exclusive Breastfeeding in maternal and child health, as it is established that HAMLET levels increase with the duration of extended Breastfeeding. However, the gap in initiation and exclusivity emphasizes the need for focused interventions and maternal education to maximize health benefits. By deepening the understanding of the role of HAMLET and breastfeeding in cancer prevention, this research paves the way for innovative interventions that could significantly reduce breast cancer incidence globally.Thus we conclude that breastfeeding plays a crucial role in protecting both mothers and future generations from the challenges associated with breast cancer. Our study is also an attempt in this direction as it is rightly said “prevention is better than cure”.

References

[1] Cordero MJA, Jimenez EG, Ferre JA, et al (2010). Breastfeeding: an effective method to prevent breast cancer. Cancer Nutr Hosp, 25, 954-8.

[2] Taylor JS, Kacmar JE, Nothnagle M, Lawrence RA (2005). A systematic review of the literature associating Breastfeeding with type 2 diabetes and gestational diabetes. J Am Coll Nutr, 24, 320-6

[3] Franca-Botelho, A. D. C., Ferreira, M. C., Franca, J. L., Franca, E. L., & Honorio-Franca, A. C. (2012). Breastfeeding and its relationship with reduction of breast cancer: a review. *Asian Pacific Journal of Cancer Prevention*, *13*(11), 5327-5332.

[4] World Health Organization (WHO), Global Data Bank on Infant and Young Child Feeding. (2009). (URL: http://whqlibdoc. who.int/publications/2009/9789241597494_eng.pdf). (Accessed: November, 2011).

[5] Jaini R, Kesaraju P, Johnson JM, et al (2010). An autoimmune-mediated strategy for prophylactic breast cancer vaccination. Nat Med, 16, 799-03.

[6] Collaborative Group on Hormonal Factors in Breast Cancer. (2002). "Breastfeeding and breast cancer risk: a systematic review and meta-analysis." *International Journal of Cancer*, 107(6), 874-882.

[7] Victora, C. G., Bahl, R., Barros, A. J. D., (2016). "Breastfeeding in the 21st century: epidemiology, mechanisms, and lifelong effect." *The Lancet*, 387(10017), 475-490.

[8] Bezerra, C. A., (2024). *Frontiers in Oncology*. DOI: 10.3389/fonc.2024.123456.

[9] Schmid, P.,. (2024). *Lancet Oncology*, 25(9), 1110-1125.

[10] Xu, J., (2024). *Nature Reviews Cancer*, 24(1), 45-59.

[11] Kramer, M. S., & Kakuma, R. (2004). *Advances in Experimental Medicine and Biology*, 554, 63-77.

[12] Collaborative Group on Hormonal Factors in Breast Cancer. (2002). *Lancet*, 360(9328), 1873-1880.

[13] Ip, S., et al. (2009). AHRQ Publication No. 09-E014.

[14] McDonald, S. W., . (2018). *Journal of Human Lactation*, 34(2), 347-353.

[15] Vanlemmens, L., (2024). *ESMO Abstracts*, LBA12.

[16] Vasundhara, K. L., Badugu, S., & Vaideek, Y. S. K. (2020). Incidence of Cancer in Breastfed Grownups-a Study. In *Data Engineering and Communication Technology: Proceedings of 3rd ICDECT-2K19* (pp. 715-724). Singapore: Springer Nature Singapore.

[17] Vasundhara, K. L., Satwika, C. G., & Rayancha, S. M. (2024). Exploring machine learning for breast cancer classification and the potential role of HAMLET in cancer treatment. *MESA*, *15*(4), 1129-1144.

Author Index