Proceedings of the Inaugural IT & Systems Student Research Conference (ITSSRC) 2025

University of Canberra
7 February 2025

We are delighted to present the proceedings of the Inaugural IT & Systems Student Research Conference 2025 (ITSSRC25), held on 7 February 2025 at the University of Canberra. This conference marks an important milestone in showcasing emerging research talent and fostering dialogue in the fields of Information Technology, Systems, Artificial Intelligence (AI), Machine Learning (ML), and Robotics.

The conference is a part of a bigger initiative to enhance the teaching-research nexus among IT and Engineering students at the University of Canberra. This initiative is designed to actively engage undergraduate and postgraduate students in research, thereby fostering the development of essential academic and professional skills such as critical thinking, analytical reasoning, and problem-solving. Participation in the conference also provides students with a valuable platform to refine their technical communication abilities, both in written and oral forms, preparing them for future academic and industry-oriented endeavors.

The conference opened with a keynote presentation titled “R&D Opportunities and Capability Development for National Infrastructure and Innovation”, by Scott Carr, Chairman and Managing Director of Datapod. The keynote speech set the stage by highlighting the critical role of research and innovation in shaping future national capabilities and addressing complex societal needs.

Across two sessions of research talks, eighteen students presented a diverse and forward-looking body of work. Topics ranged from EEG-based emotion recognition, performance prediction in education, and smart city resource optimisation, to AI-enhanced therapeutic interventions, agricultural automation, robotic systems, and climate forecasting. The afternoon session further explored areas such as data center energy optimisation, soft robotics, consumer analytics, cyber-attack detection, and assistive technology for ageing populations.

The day concluded with a panel discussion titled “Shaping the Future: AI & Robotics Research in a Rapidly Evolving World”, by four academics from diverse backgrounds. The discussion offered reflections on the implications of emerging technologies and the trajectory of student-led innovation and highlighted the interplay between research and industry.

These proceedings reflect the breadth, creativity, and real-world relevance of the research undertaken by our student community. We commend all contributors for their dedication and insight, and we hope this collection inspires continued exploration, critical thinking, and collaboration within and beyond the university setting.

Machine Learning Models for EEG-based Emotion Recognition: a Review and Prototype Development

Sangay Wangmo, Namgay Tshering, Kuenzang Dorji, Sonam Thinley, Min Wang

***Abstract*—Emotions profoundly influence human behaviour, yet accurately identifying them through physiological signals remains a challenge. This paper introduces EmoSense, a novel emotion recognition system that leverages electroencephalogram (EEG) signals to identify four basic emotions: neutral, sad, fear, and happy. Drawing on theories from neuroscience, artificial intelligence, and human-computer interaction, this study addresses gaps in the literature by exploring machine learning approaches for EEG-based emotion recognition. We rigorously review the whole processing framework, including signal preprocessing, feature extraction using Power Spectral Density (PSD), and classification with Support Vector Machines (SVM), K-Nearest Neighbors (KNN), and Convolutional Neural Networks (CNNs). The results demonstrate that SVM achieves the highest accuracy, outperforming KNN and CNN, highlighting the potential of classical machine learning on small-scale EEG datasets. To translate these findings into practical utility, a prototype of the EmoSense system was implemented, enabling automatic recognition of users' emotions based on their EEG signals. The contributions lie in the integration of robust preprocessing, scalable feature extraction, and machine learning models while emphasizing practical applicability across diverse domains. The study concludes by proposing future enhancements, including advanced feature engineering, model optimization, and real-time application development. These improvements aim to expand EmoSense's impact across domains such as mental health monitoring, adaptive learning, and interactive entertainment technologies. By bridging theoretical insights and practical implementation, EmoSense represents a significant advancement in EEG-based emotion recognition, offering novel pathways for understanding and applying emotional intelligence in real-world contexts.**

***Index Terms*—Electroencephalogram (EEG), brain-computer interfaces (BCI), emotion recognition, machine learning, and convolutional neural networks (CNNs).**

I. Introduction

Emotions play a fundamental role in individuals' daily lives, influencing decisions, shaping interactions, and affecting perceptions. Despite their importance, accurately identifying and understanding emotions has long been a challenge in psychological and neuroscientific research [1]. Traditional methods, such as analysing facial expressions, body language, and vocal tones, provide valuable insights but often struggle with the subjective and transient nature of emotional experiences. Electroencephalogram (EEG) technology and brain-computer interface (BCI) systems have gained significant attention for their diverse applications, such as workload assessment, cognitive state monitoring and user authentication [2], [3]. These advancements are making EEG devices more accessible and practical, enabling their integration into various real-world tasks. The growing availability of EEG technology for applications beyond research highlights its potential for emotion recognition, as it can now be leveraged in everyday settings. This alignment of technology and need provides a strong foundation for the development of EEG-based emotion recognition systems, extending the capabilities of EEG to address the complex challenge of decoding emotions.

This paper explores the prediction of four basic emotions (neutral, sad, fear, and happy) using EEG signals and introduces EmoSense, a web application capable of real-time emotion recognition from EEG data. This research represents a significant advancement in the field of emotion recognition by offering an objective and precise approach to understanding emotional states. The motivation for this research stems from the growing need for accurate and scalable emotion recognition systems in domains such as mental health, education, and human-computer interaction [4]. Current emotion recognition techniques face limitations in reliably capturing internal emotional states, particularly in environments where observable behaviours might not fully reflect underlying feelings. EEG signals, which directly capture neural activity, provide a promising alternative by enabling researchers to move beyond external expressions and delve into the brain's electrical patterns [5]. The development of an intuitive, real-time system such as the EmoSense addresses both a scientific challenge and a practical need, paving the way for innovative applications in adaptive learning, therapy, and personalised technologies.

This study bridges neuroscience, artificial intelligence (AI), and human-computer interaction, leveraging EEG signals as a novel medium for decoding emotions. Unlike traditional methods that rely on observable external expressions, EEG-based approaches enable direct interpretation of neural signals, providing a more accurate and unbiased understanding of emotions. In this study, we rigorously review the entire processing framework required for EEG-based emotion recognition, encompassing signal preprocessing, feature extraction, and classification. Signal preprocessing is a critical step to remove noise and artifacts inherent in raw EEG signals, ensuring the reliability and quality of the data for further analysis. Feature extraction, a pivotal phase in the framework, is performed using Power Spectral Density (PSD), which captures the frequency domain characteristics of EEG signals linked to different emotional states. These extracted features are then used to train and evaluate machine learning models for emotion classification. Specifically, the study examines Support Vector

All the authors are with the Faculty of Science and Technology, University of Canberra, ACT 2617, Australia. *Corresponding author: Min Wang* (e-mail: min.wang@canberra.edu.au).

Machines (SVM), K-Nearest Neighbors (KNN), and Convolutional Neural Networks (CNNs) as classifiers to investigate their efficacy in predicting four basic emotions: neutral, sad, fear, and happy. Each classifier brings unique strengths to the task and this comparative analysis provides insights into the capabilities of classical machine learning and deep learning methods for emotion recognition, particularly when applied to small-scale EEG datasets. The results lay a solid foundation for the development of EmoSense, an emotion recognition system that combines robust preprocessing, effective feature extraction, and scalable classification techniques to deliver a practical and efficient solution.

The contributions of this study can be summarised as follows:

1) We review EEG-based emotion recognition systems and compare studies on the signal preprocessing, feature extraction and classification methods.
2) To translate these findings into practical utility, a prototype of the EmoSense system was implemented, enabling automatic recognition of users' emotions based on their EEG signals. It integrates robust signal preprocessing, scalable feature extraction, and machine learning models while emphasising practical applicability across diverse domains.

The structure of this paper is organised as follows: Section II provides a literature review of related studies on emotion recognition using EEG signals. Section III details the methodology, including data preprocessing, signal processing techniques, feature extraction, and classification methods. Section IV presents the experimental results and offers an in-depth discussion of the findings. Section V outlines the implementation details of the developed prototype system. Finally, Section VI concludes the study and discusses potential directions for future research.

II. Related Work

Emotion recognition using EEG signals has become a vital research area, bridging the gap between emotional psychology and computational neuroscience [6]. The non-invasive nature and high temporal resolution of EEG make it ideal for capturing the dynamic nature of emotions [7]. Recent studies leverage deep learning to improve recognition accuracy and efficiency, offering promise in mental health monitoring, interactive technologies, and beyond [8].

Early research relied on domain knowledge-based feature extraction and classical machine learning algorithms. Pioneering works explored the link between emotional states and frontal EEG asymmetry, power spectral features and entropy features [5]. As technology advanced, the focus shifted towards automated feature extraction with studies utilizing signal processing techniques for analysis [9], [10]. Table I summarises the details and performance of the existing works for EEG-based emotion recognition.

The use of EEG signals for emotion recognition has been extensively studied, with techniques ranging from traditional machine learning models to advanced deep learning approaches. A recent study presented a deep learning-based method using Bi-directional Long Short-Term Memory (Bi-LSTM) networks [8], showing promising results in classifying complex emotional states from EEG data. The advent of deep learning has revolutionized the field. Deep Belief Networks (DBNs), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs) can learn complex representations of EEG signals directly from data, bypassing manual feature engineering [11], [12]. Studies have demonstrated the efficacy of DBNs in identifying key frequency bands and channels for emotion recognition, achieving high accuracy levels [13]. Similarly, CNNs have been used to capture the spatial-temporal dynamics of EEG signals for emotion classification [14]. Fig. 1 and Fig. 2 show the number of studies using different machine learning models for EEG-based emotion recognition and the performance comparison, respectively. Among the related articles from 2009 to 2021, most of the studies used SVM, KNN, and CNN as the classification models, with average accuracy of 89.19%, 86.57%, and 89.03%, respectively [10].

Wearable technologies also play a critical role in advancing emotion recognition applications. For example, the authors investigated the use of wearable sensor technologies to aid special education [15], demonstrating how emotion recognition could be integrated into educational tools for students requiring special attention. Furthermore, the integration of emotion recognition into everyday consumer electronics was explored by Raja and Sigg [16], who developed RFexpress, a system that leverages wireless network edges for RF-based emotion sensing.

A significant development is the shift towards personalized emotion recognition systems. Recognizing individual variability in EEG signals, studies have developed personalized models that adjust to individual differences, achieving notable improvements in classification accuracy [17]. Furthermore, the application of EEG-based emotion recognition technologies in real-world scenarios is exemplified by studies implementing these systems in adaptive user interfaces, enhancing user experience through real-time emotional feedback [18]. The research also extends to emotion and stress recognition related sensors, where [19] reviewed the current technologies and methodologies, highlighting the need for more robust and ubiquitous sensing solutions for real-time applications.

With deep learning methods and wearable technologies, emotion recognition using EEG signals has made significant progress. Key directions for further progress could include enhancing real-time processing capabilities and integrating the emotion recognition technology into everyday applications such as mental health monitoring. However, many challenges still remain, such as improving the robustness and scalability of the models, ensuring performance reliability, and addressing the ethical and privacy concerns.

III. Methodology

A. Dataset and Pre-processing

The SEED dataset collected by Brain-like Computing and Machine Learning (BCMI), particularly the SEED IV extension, has become a cornerstone in EEG-based emotion

TABLE I: Summary of related works in EEG-based emotion recognition.

Year	Study	Stimuli	#Subjects	#Channels	Pre-processing	SR (Hz)	Feature Extraction+Selection	Classification	Emotions	Accuracy (%)
2011	[20]	Video	6	32	Bandpass (1-40 Hz), EMG (manually)	-	FFT	SVM	Positive, Negative	89.22
2011	[21]	Video	5	62	EOG, EMG (manually)	200	FFT+MRMR	SVM, KNN, MP	Joy, Relax, Sad, Fear	66.51
2012	[22]	Video	4	32	-	-	ASP, CSP, FBCSP	NB	Valence, Arousal	Valence: 65, Arousal: 80
2011	[23]	Video	30	32	Bandpass (4-45 Hz), CAR	256	PSD, AI	SVM, RBF	Valence, Arousal	Valence: 50.50, Arousal: 62.10
2013	[24]	Video	6	62	Manually	200	DE, DASM, RASM, ES +MRMR	SVM, KNN	Positive, Negative	76.56-84.22
2013	[25]	Audio-video	6	62	EOG, EMG (manually)	200	AE, FD, HE, PSD, WT +PCA, LDA CFS	SVM	Positive, Negative	91.77
2014	[25]	Video	6	62	Manually	200	PSD, WT, NDA	SVM	Positive, Negative	87.53
2015	[13]	SEED	15	62	Bandpass (0.3-50 Hz), EOG, EMG (manually)	200	PSD, DE, DASM, RASM, DCAU	DBN, KNN, LR, SVM	Positive, Neutral, Negative	83.99-86.08
2015	[26]	Audio-video	15	62	Bandpass (0.3-75 Hz)	200	DE	KNN, LR, LDA, QDA, SVM, NB, DT, SGD, RF, GB	Positive, Neutral, Negative	81.26
2016	[27]	Video	8	60	ICA, EOG (manually)	500	PSD (Welch's)	SVM Gaussian	Positive, Neutral, Negative	73
2016	[28]	Video	8	1	Wavelet transform	-	STFT	SVM	Fear, Relaxation	92.10
2018	[29]	Video	30	62	Bandpass (0.1-100 Hz), Notch filter (50 Hz), BSS (FICA)	-	DE, STFT, EMD+MRMR	SVM	Joy, Neutrality, Sadness, Anger, Disgust, Fear	87.36 binary, 54.52 six
2019	[30]	SEED	15	62	Bandpass (0.3-75 Hz), EOG, EMG (manually)	200	DE+ANOVA	DNNs	Positive, Neutral, Negative	93.29
2020	[31]	Audio-video	10	8	Bandpass (1-80 Hz), Notch filter	128	DE, DASM, RASM +Mutual Information, Chi-square	RF, DL	Valence, Arousal	71.22-83.18
2020	[32]	SEED	15	62	Filter (0-75 Hz)	200	Temporal entropy, Spectral entropy, Shannon entropy	AE+RF, DT, KNN, SVM, ELM	Positive, Neutral, Negative	94.4
2021	[33]	Audio-video	20	24	Filter (0.5-100 Hz), Notch filter, ICA	256	Hjorth complexity, Standard deviation value, Log energy	DT, KNN, SVM, ELM	Fear, Sad, Happy, Relax	97.24

recognition research, offering high-resolution EEG signals across a spectrum of emotional states. Seminal works by Zheng [13], and subsequent studies [34], [35] have leveraged this dataset to explore the depths of emotion recognition using deep learning techniques, highlighting the significance of critical frequency bands and the efficacy of different EEG features across subjects. Thus, in this study as well, the SEED IV dataset has been used for evaluation.

The SEED IV dataset was collected from a cohort of 15 participants across three sessions each separated by 2 weeks period. During each session, participants were made to watch a movie clip of approximately 2 minutes of 25 distinct movie clips of which each would evoke a specific emotion: neutral, sad, fear or happy [13]. The EEG signals were captured through 62 channels adhering to the 10–20 system, a standard used in neurophysiological research. In this system, the '10' and '20' denote that the distances between adjacent electrodes are either 10% or 20% of the total front–back or right–left distance of the skull [36].

EEG data are inherently noisy, requiring rigorous preprocessing to ensure signal fidelity [37]. It is a common practice to employ bandpass filters to remove artifacts and apply Independent Component Analysis (ICA) for artifact rejection [38]. The raw EEG data from SEED IV dataset had a sampling rate of 200 Hz samples per second. To eliminate background noise and artifacts from the raw data, a bandpass filter was

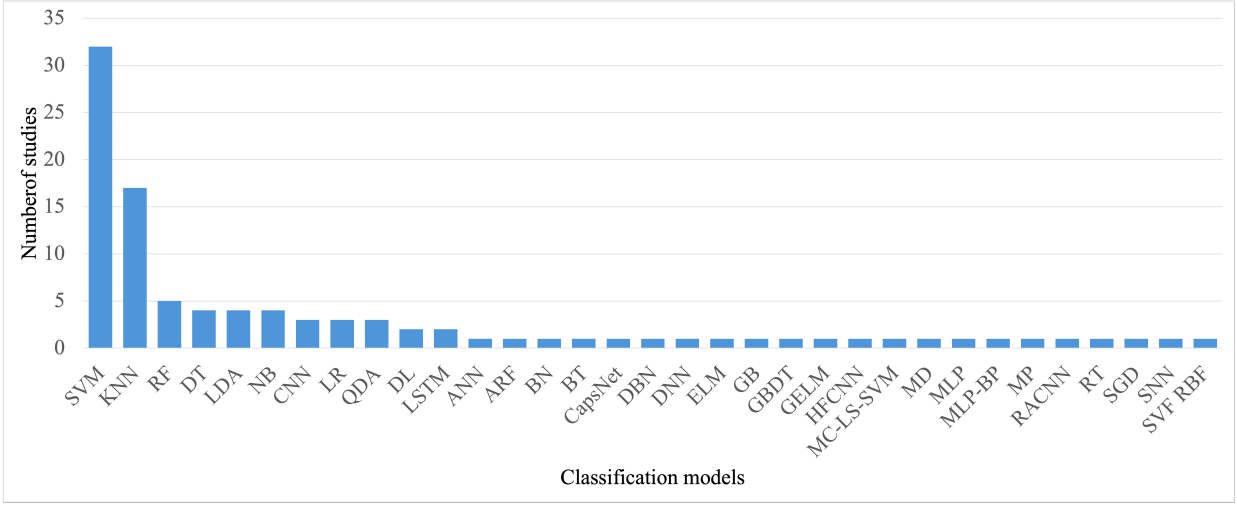


Fig. 1: Number of studies for different models for EEG-based emotion recognition based on reviewed papers between 2009 and 2021 in ScienceDirect.

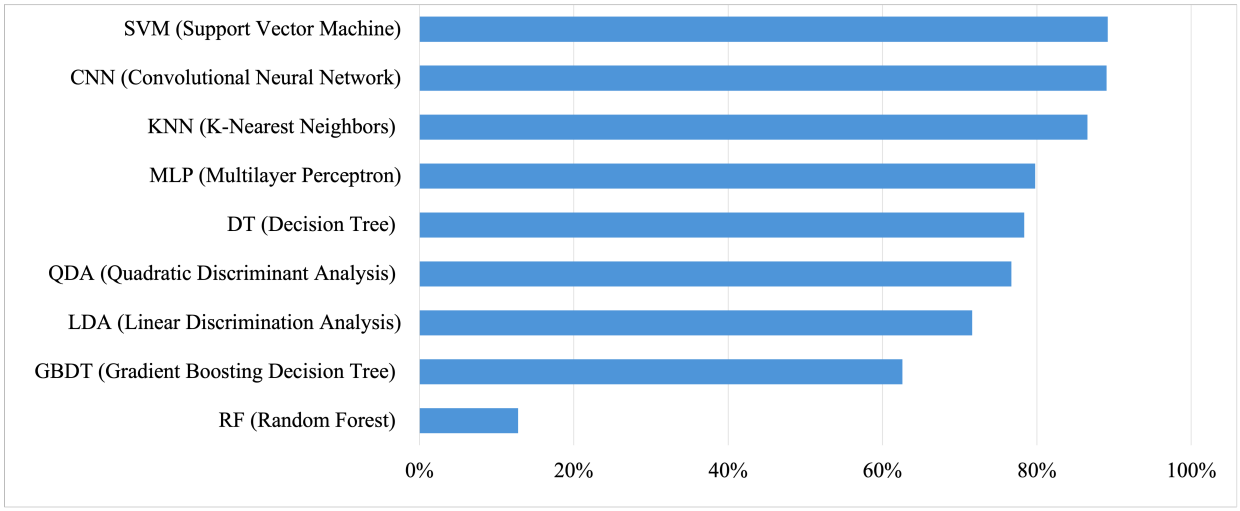


Fig. 2: Performance comparison of different models for EEG-based emotion recognition based on reported accuracies.

applied to restrict frequencies to [0.5, 42] Hz. The ICA was then performed on the filtered data, and artifact-contaminated components were ruled out. A sliding window of 5 seconds was applied to the filtered dataset, resulting in about 37,000 samples with around 12,000 for each recording session. The signal preprocessing was performed using Python and the MNE package.

B. Feature Extraction

Feature extraction is a fundamental step, as it involves transforming raw brainwave or physiological data into meaningful features that can be used for emotion recognition [39], [40], [5]. Tang explored the application of PSD in the context of defect measurement, highlighting its utility in signal analysis which can be analogous to identifying nuanced changes in EEG signals for emotion recognition [41]. Alam focused on extracting features of EEG signals using PSD for a motor imagery-based BCI, emphasizing the importance of PSD in enhancing classification performance by effectively capturing the frequency-based characteristics of the EEG signals [42].

Differential Entropy (DE), on the other hand, offers insights into the complexity and uncertainty associated with EEG signals [43]. In this study, they proposed a method combining DE with brain connectivity features for EEG-based emotion recognition, demonstrating high accuracy and precision on established datasets such as DEAP and SEED. The KNIFE, a novel, differentiable kernel-based estimator of DE [44], was then proposed to provide a flexible approach to neural network training and feature extraction in EEG applications.

From the review studies, from among several feature extraction methods, the combination of Power Spectral Density (PSD) and Differential Entropy (DE) can be seen as an effective method for Brainwave-based emotion recognition. To compute traditional power spectral density (PSD) features from the brain wave data, PSD welch method was used with Short Time Fourier Transform (STFT) at 256, 20 seconds window with no overlap and a Hamming window function, in

five different frequency bands: delta (1-3Hz), theta (4-7 Hz), alpha (8-13 Hz), beta (14-30 Hz) and gamma (31-50 Hz). The DE feature is computed as follows:

$$h(X) = -\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} exp\frac{(x-\mu)^2}{2\sigma^2} ln(\frac{1}{\sqrt{2\pi\sigma}} exp\frac{(x-\mu)^2}{2\sigma^2})dx \quad (1)$$

$$= \frac{1}{2} ln 2\pi\sigma^2 \quad (2)$$

where X represents the Gaussian distribution $\mathcal{N}(\mu, \sigma)$, and e is a constant. DE offers a similar approach to analysing brain waves as the logarithmic version of the PSD.

C. Classification

The accurate classification of extracted and reduced features is essential for reliable emotion recognition. Zheng proposed an EEG-based emotion classification approach using deep belief networks [13], achieving promising accuracy results in discriminating different emotional states from brainwave signals. Additionally, Zhang developed a facial emotion recognition system based on biorthogonal wavelet entropy and fuzzy SVM, demonstrating high accuracy through stratified cross-validation [45]. Moreover, K-Nearest Neighbour (KNN) was applied for classification to multichannel EEG signals for emotion recognition [34], achieving notable accuracy in identifying diverse emotional states. These studies indicate that the traditional classification models, SVM, and KNN classification methods can effectively classify features extracted from brainwave and physiological data, yielding high accuracy for emotion recognition using EEG.

As evident from the above studies, SVM and KNN were the traditional classification method selected for this study. For the KNN classifier, the Minkowski distance is selected as the distance metric, and the number of nearest neighbours is set to 1. For SVM, the GridSearchCV was used to find the best parameters which were {'C': 10, 'gamma': 0.1, 'kernel': 'rbf'} where 'C' is the regularization parameter, 'gamma' is the kernel coefficient and 'kernel' is the kernel type selected which was 'rbf'.

The previous study also found that the deep learning method, paralytically CNN had a higher accuracy rate second to SVM. Therefore, this study also used the CNN model to classify the emotions. The result of a recent study [46] has demonstrated significant advancements in EEG analysis by achieving an accuracy of 90.41% through their proposed CNN model. Their approach utilized the tSNE visualization to showcase the extracted features, showcasing superior performance compared to traditional feature representations. Similarly, a multi-column CNN model [47] was proposed for EEG-based emotion recognition, showcasing enhanced accuracy compared to conventional methods using handcrafted features. In this study, we implement a 3-layer CNN, which consists of a convolutional layer, a fully connected layer and an output layer with softmax function.

IV. Results

We evaluate the selected models, which are SVM, KNN and CNN, on the SEEDIV dataset (15 subjects and 62 channels). A 20 second non-overlapping window was used to generate samples, and PSD was estimated for each channel of each sample using the Welch method. The band power features were then extracted, according the five canonical EEG frequency bands: delta (1-3 Hz), theta (4-7 Hz), alpha (8-13 Hz), beta (14-30 Hz) and gamma (31-50 Hz). SVM and KNN classifiers were used for the extracted band power features and CNN was used for raw EEG signals without any handcrafted feature extraction process.

The results are summarised in Table II. The accuracies in recognising four emotions are 65%, 58% and 50.7% for SVM, KNN and CNN, respectively, demonstrating strong predictive capabilities of the classic approach using handcrafted features and SVM classifier. Detailed performance of the SVM and KNN models are presented in Table III and Table IV, respectively. As shown in the SVM classification report, Class 1 has the highest recall indicating the model is better at identifying the Sad emotion. Class 2 has the highest precision suggesting that when the classifier predicts the Fear emotion, it is more likely to be correct. Class 1 also has the highest F1-score indicating a good balance between precision and recall. For KNN classifier, Class 1 (sad) has the highest precision, recall and F1-score. The CNN provided a 100% training accuracy whereas the validation accuracy ranged throughout the epoch from 40% to 55% and the resting accuracy was 50.7%. An early stopping criterion with patience of 10 epochs was applied.

As depicted in the research findings, the SVM model exhibited the highest accuracy ratings, establishing itself as the most accurate model. However, it's noteworthy that the accuracy falls short compared to benchmark accuracies achieved in other related research. To enhance accuracy, there's a crucial need for more extensive data preprocessing and feature extraction. Particularly, emphasis should be placed on robust noise detection and effective handling of EEG signals. Furthermore, in feature extraction, it's imperative to explore additional techniques beyond Power Spectral Density (PSD) and Differential Entropy (DE), allowing for comprehensive comparison and analysis of results. Some of the feature extraction techniques are time frequency distributions (TFD), fast Fourier transform (FFT), eigenvector methods (EM), wavelet transform (WT), and auto regressive method (ARM).

The CNN model generated a training accuracy of 100% whereas a validation accuracy of 50.7% indicating a over-fitting of the training process. To overcome over-fitting and improve the model's generalisation capacity, dropout, regularisation (e.g., L1 and L2 regularisation), and data augmentation techniques can be adopted [48]. Moreover, the CNN architecture should be optimised to improve the performance [49]. We will explore regularisation and data augmentation [50] and improve the architecture to futher improve the recognition accuracy in our future work.

V. Prototype Implementation and Discussion

To translate our findings into practical utility, a prototype web application system, EmoSense, was developed to enable real-time emotion recognition based on users' EEG signals.

TABLE II: Evaluation results on the SEEDIV dataset.

Pre-processing	**SR (Hz)**	**Feature Extraction**	**Frequency Bands**	**Classification Model**	**Accuracy (%)**
Bandpass Filtering (0.5-42 Hz)	200	PSD Welch method 20 second segment (no overlap) with Hamming window	Delta (1-3 Hz) Theta (4-7 Hz) Alpha (8-13 Hz) Beta (14-30 Hz) Gamma (31-50 Hz)	Support Vector Machine (SVM)	65
				K-Nearest Neighbours (KNN)	58
		5 second segment	-	Convolutional Neural Network (CNN)	50.70

TABLE III: SVM results

Emotion labels	Precision	Recall	F1-score	Support
0	0.64	0.66	0.65	123
1	0.63	0.76	0.69	147
2	0.75	0.47	0.58	99
3	0.64	0.64	0.64	108
accuracy			0.65	477
macro avg	0.66	0.63	0.64	477
weighted avg	0.66	0.65	0.64	477

TABLE IV: KNN results.

Emotion labels	Precision	Recall	F1-score	Support
0	0.57	0.54	0.56	123
1	0.61	0.68	0.64	147
2	0.58	0.45	0.51	99
3	0.57	0.62	0.59	108
accuracy			0.58	477
macro avg	0.58	0.57	0.58	477
weighted avg	0.58	0.58	0.58	477

The system features a user-friendly front-end with two primary interfaces: a signal uploading interface for users to upload their EEG recordings and a results display interface that visualizes the predicted emotional states, as illustrated in Fig. 3. The back-end is designed to handle the core processing pipeline, which includes signal preprocessing to remove noise and artifacts, feature extraction using PSD analysis to capture frequency-domain characteristics, and classification using a SVM model, which demonstrated the highest accuracy in our evaluation. The back-end is implemented in Python, leveraging modular and efficient libraries to ensure scalability and seamless integration of future updates. The lightweight design of the processing and classification functions allows the application to operate efficiently on consumer-grade hardware, ensuring accessibility and practicality. This end-to-end system bridges advanced EEG-based emotion recognition research with real-world usability, offering potential applications in areas such as mental health monitoring, adaptive learning environments, and interactive technologies. Our prototype demonstrates the feasibility of translating cutting-edge EEG-based emotion recognition research into a functional and user-centric application.

VI. Conclusion

This paper presents EmoSense, a novel emotion recognition system utilizing EEG signals to classify four basic emotions: neutral, sad, fear, and happy. By integrating neuroscience, artificial intelligence, and human-computer interaction, the study addresses existing gaps in emotion recognition literature, particularly in the application of machine learning techniques to EEG data. Our analysis of various classification models such as SVM, KNN, and CNN demonstrates the superiority of SVM in terms of accuracy, particularly when applied to small-scale EEG datasets. The development of the EmoSense prototype represents a significant step toward translating theoretic research into a practical tool, offering automatic emotion recognition capabilities. The system combines robust signal preprocessing, effective feature extraction, and scalable machine learning models, ensuring its applicability across diverse fields, including mental health monitoring, adaptive learning, and interactive human-machine systems. Future improvements in feature engineering, model optimization, and real-time processing are anticipated to further enhance the system's accuracy and usability, expanding its potential applications. In addition, we will explore data augmentation techniques to address the problem of limited data resources in the EEG domain to improve the performance of EEG-based emotion recognition systems, especially for deep learning-based approaches.

References

[1] B. Chakravarthi, S.-C. Ng, M. Ezilarasan, and M.-F. Leung, "EEG-based emotion recognition using hybrid CNN and LSTM classification," *Frontiers in Computational Neuroscience*, vol. 16, p. 1019776, 2022.

[2] M. Wang, H. El-Fiqi, J. Hu, and H. A. Abbass, "Convolutional neural networks using dynamic functional connectivity for EEG-based person identification in diverse human states," *IEEE Transactions on Information Forensics and Security*, vol. 14, no. 12, pp. 3259–3272, 2019.

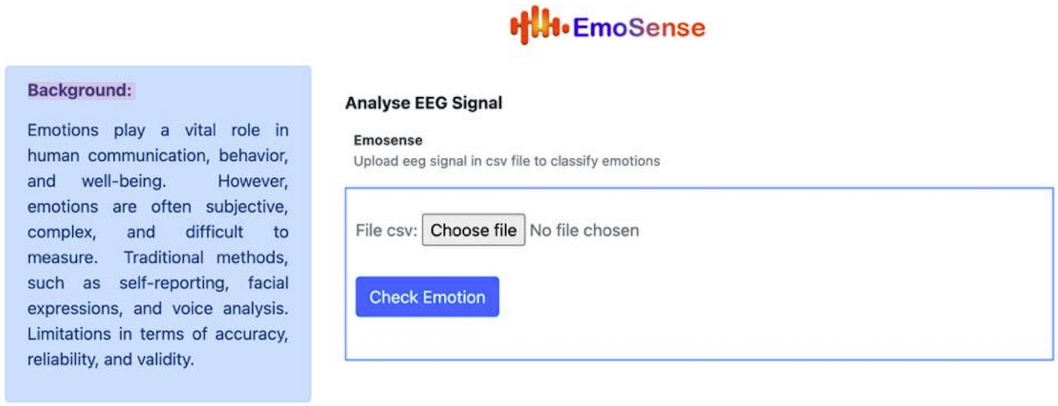
[3] M. Wang, S. Wang, and J. Hu, "Cancellable template design for privacy-preserving EEG biometric authentication systems," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 3350–3364, 2022.

[4] A. Hussein, L. Ghignone, T. Nguyen, N. Salimi, H. Nguyen, M. Wang, and H. A. Abbass, "Characterization of indicators for adaptive human-swarm teaming," *Frontiers in Robotics and AI*, vol. 9, p. 745958, 2022.

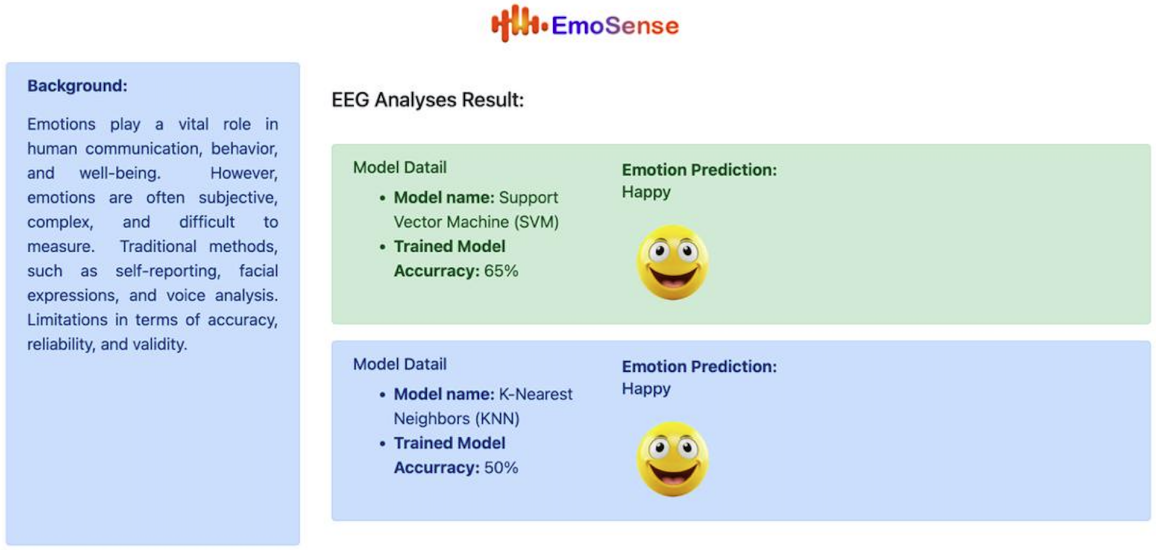
[5] M. Wang, X. Yin, Y. Zhu, and J. Hu, "Representation learning and pattern recognition in cognitive biometrics: A survey," *Sensors*, vol. 22, no. 14, p. 5111, 2022.

[6] D. W. Prabowo, H. A. Nugroho, N. A. Setiawan, and J. Debayle, "A systematic literature review of emotion recognition using EEG signals," *Cognitive Systems Research*, p. 101152, 2023.

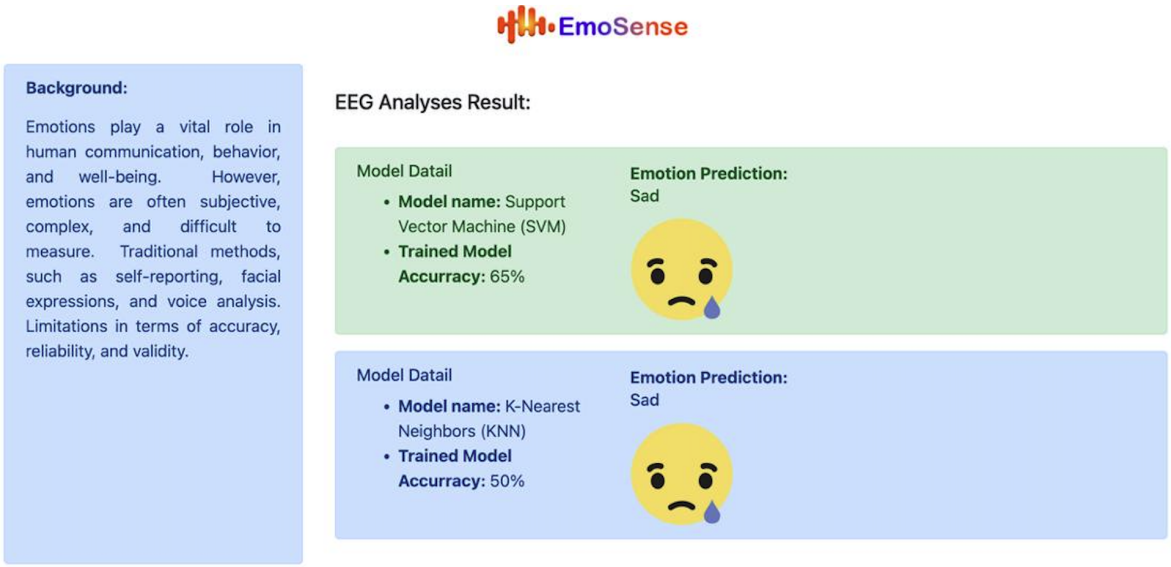
[7] S. Li, N. Scott, and G. Walters, "Current and potential methods for measuring emotion in tourism experiences: A review," *Current issues in Tourism*, vol. 18, no. 9, pp. 805–827, 2015.



(a) User interface of uploading a signal in EmoSense.



(b) Emotion recognition result of an EEG example (happy) using the EmoSense app.



(c) Emotion recognition result of an EEG example (sad) using the EmoSense app.

Fig. 3: Demonstration of the developed EmoSense application.

[8] M. Algarni, F. Saeed, T. Al-Hadhrami, F. Ghabban, and M. Al-Sarem, "Deep learning-based approach for emotion recognition using electroencephalography (EEG) signals using bi-directional long short-term memory (Bi-LSTM)," *Sensors*, vol. 22, no. 8, p. 2976, 2022.

[9] S. Koelstra, C. Muhl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras, "Deap: A database for emotion analysis; using physiological signals," *IEEE transactions on affective computing*, vol. 3, no. 1, pp. 18–31, 2011.

[10] M. M. Rahman, A. K. Sarkar, M. A. Hossain, M. S. Hossain, M. R. Islam, M. B. Hossain, J. M. Quinn, and M. A. Moni, "Recognition of human emotions using EEG signals: A review," *Computers in biology and medicine*, vol. 136, p. 104696, 2021.

[11] H.-M. Shao, J.-G. Wang, Y. Wang, Y. Yao, and J. Liu, "EEG-based emotion recognition with deep convolution neural network," in *2019 IEEE 8th Data Driven Control and Learning Systems Conference (DDCLS)*. IEEE, 2019, pp. 1225–1229.

[12] M. P. T and V. R, "Advanced EEG analysis based emotion recognition: A deep learning classifier with hybrid feature selection and artifact reduction," *International Journal of Electrical and Electronics Engineering*, vol. 10, pp. 128–141, 2023.

[13] W.-L. Zheng and B.-L. Lu, "Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks," *IEEE Transactions on autonomous mental development*, vol. 7, no. 3, pp. 162–175, 2015.

[14] P. Bashivan, I. Rish, M. Yeasin, and N. Codella, "Learning representations from EEG with deep recurrent-convolutional neural networks," *arXiv preprint arXiv:1511.06448*, 2015.

[15] R. M. Mehmood and H. Lee, "Towards building a computer aided education system for special students using wearable sensor technologies," *Sensors*, vol. 17, no. 2, p. 317, 2017.

[16] M. Raja and S. Sigg, "RFexpress!—exploiting the wireless network edge for RF-based emotion sensing," in *2017 22nd IEEE international conference on emerging technologies and factory automation (ETFA)*. IEEE, 2017, pp. 1–8.

[17] H. Amrani, D. Micucci, M. Nalin, and P. Napoletano, "Emotion personalization with machine learning using EEG signals and dry electrodes," in *2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE)*. IEEE, 2023, pp. 132–137.

[18] W.-C. Fang, K.-Y. Wang, N. Fahier, Y.-L. Ho, and Y.-D. Huang, "Development and validation of an EEG-based real-time emotion recognition system using edge AI computing platform with convolutional neural network system-on-chip design," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, vol. 9, no. 4, pp. 645–657, 2019.

[19] K. Kyamakya, F. Al-Machot, A. Haj Mosa, H. Bouchachia, J. C. Chedjou, and A. Bagula, "Emotion and stress recognition related sensors and machine learning technologies," p. 2273, 2021.

[20] L.-C. S. Dan Nie, Xiao-Wei Wang and B.-L. Lu, "EEG-based emotion recognition during watching movies," *International Journal of Electrical and Electronics Engineering*, p. 667–670, 2011.

[21] D. N. Xiao-Wei Wang and B.-L. Lu, *EEG-Based Emotion Recognition Using Frequency Domain Features and Support Vector Machines*. Springer, Berlin, Heidelberg, 2011.

[22] D. Huang, C. Guan, K. K. Ang, H. Zhang, and Y. Pan, "Asymmetric spatial pattern for EEG-based emotion detection," in *The 2012 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2012, pp. 1–7.

[23] T. P. Mohammad Soleymani, Jeroen Lichtenauer and M. Pantic, "A multimodal database for affect recognition and implicit tagging," *Transactions on Affective Computing*, vol. 3, pp. 42–55, 2011.

[24] J.-Y. Z. Ruo-Nan Duan and B.-L. Lu, "Differential entropy feature for EEG-based emotion classification," *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*, 2013.

[25] X.-W. Wang, D. Nie, and B.-L. Lu, "Emotional state classification from EEG data using machine learning approach," *Neurocomputing*, vol. 129, pp. 94–106, 2014.

[26] W.-L. Zheng, R. Santana, and B.-L. Lu, "Comparison of classification methods for EEG-based emotion recognition," in *World Congress on Medical Physics and Biomedical Engineering, June 7-12, 2015, Toronto, Canada*. Springer, 2015, pp. 1184–1187.

[27] S. Liu, J. Tong, M. Xu, J. Yang, H. Qi, and D. Ming, "Improve the generalization of emotional classifiers across time by using training samples from different days," in *2016 38th Annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 2016, pp. 841–844.

[28] E. B. P. Amir Jalilifard and M. K. Islam, "Emotion classification using single-channel scalp-EEG recording," *IEEE Engineering in Medicine and Biology Society (EMBC)*, p. 845–849, 2016.

[29] N. Zhuang, Y. Zeng, K. Yang, C. Zhang, L. Tong, and B. Yan, "Investigating patterns for self-induced emotion recognition from EEG signals," *Sensors*, vol. 18, no. 3, p. 841, 2018.

[30] Y. Wang, S. Qiu, C. Zhao, W. Yang, J. Li, X. Ma, and H. He, "EEG-based emotion recognition with prototype-based data representation," in *2019 41st annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 2019, pp. 684–689.

[31] E. P. Torres, E. A. Torres, M. Hernández-Álvarez, and S. G. Yoo, "Emotion recognition related to stock trading using machine learning algorithms with feature selection," *Ieee Access*, vol. 8, pp. 199 719–199 732, 2020.

[32] L. G. Abhijit Bhattacharyya, Rajesh Kumar Tripathy and R. B. Pachori, "A novel multivariate-multiscale approach for computing EEG spectral and temporal complexity for human emotion recognition," *IEEE Sensors Journal*, vol. 21, no. 3, p. 3579–3591, 2020.

[33] S. K. Khare and V. Bajaj, "An evolutionary optimized variational mode decomposition for emotion recognition," *IEEE Sensors*, vol. 21, no. 2, p. 2035–2042, 2021.

[34] M. Li, H. Xu, X. Liu, and S. Lu, "Emotion recognition from multichannel EEG signals using K-nearest neighbor classification," *Technology and health care*, vol. 26, no. S1, pp. 509–519, 2018.

[35] T. Duan, M. A. Shaikh, M. Chauhan, J. Chu, R. K. Srihari, A. Pathak, and S. N. Srihari, "Meta learn on constrained transfer learning for low resource cross subject EEG classification," *IEEE Access*, vol. 8, pp. 224 791–224 802, 2020.

[36] TMSi, "What is the 1020 system for EEG?" 2022.

[37] J. Xu, M. N. Potenza, and V. D. Calhoun, "Spatial ICA reveals functional activity hidden from traditional fMRI GLM-based analyses," p. 154, 2013.

[38] I. Winkler, S. Debener, K.-R. Müller, and M. Tangermann, "On the influence of high-pass filtering on ICA-based artifact reduction in EEG-ERP," in *2015 37th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 2015, pp. 4101–4105.

[39] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," in *2014 science and information conference*. IEEE, 2014, pp. 372–378.

[40] Z. M. Hira and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data," *Advances in bioinformatics*, vol. 2015, no. 1, p. 198363, 2015.

[41] Z. Tang, K. Zhou, P. Meng, and Y. Li, "A frequency-domain location method for defects in cables based on power spectral density," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–10, 2022.

[42] M. N. Alam, M. I. Ibrahimy, and S. Motakabber, "Feature extraction of EEG signal by power spectral density for motor imagery based BCI," in *2021 8th International Conference on Computer and Communication Engineering (ICCCE)*. IEEE, 2021, pp. 234–237.

[43] W.-L. Zheng, J.-Y. Zhu, and B.-L. Lu, "Identifying stable patterns over time for emotion recognition from EEG," *IEEE transactions on affective computing*, vol. 10, no. 3, pp. 417–429, 2017.

[44] G. Pichler, P. Colombo, M. Boudiaf, G. Koliander, and P. Piantanida, "KNIFE: Kernelized-neural differential entropy estimation," 2022.

[45] Y. Zhang, C. Cheng, and Y. Zhang, "Multimodal emotion recognition using a hierarchical fusion convolutional neural network," *IEEE access*, vol. 9, pp. 7943–7951, 2021.

[46] S. Hwang, K. Hong, G. Son, and H. Byun, "Learning CNN features from DE features for EEG-based emotion recognition," *Pattern Analysis and Applications*, vol. 23, pp. 1323–1335, 2019.

[47] H. Yang, J. Han, and K. Min, "A multi-column CNN model for emotion recognition from EEG signals," *Sensors*, vol. 19, no. 21, p. 4736, 2019.

[48] P. Thanapol, K. Lavangnananda, P. Bouvry, F. Pinel, and F. Leprévost, "Reducing overfitting and improving generalization in training convolutional neural network (CNN) under limited sample sizes in image recognition," in *2020-5th International Conference on Information Technology (InCIT)*. IEEE, 2020, pp. 300–305.

[49] R. Ruizendaal, "Deep learning #3: More on cnns & handling overfitting," *Towards Data Science*, 2017.

[50] S. M. Abdelfattah, G. M. Abdelrahman, and M. Wang, "Augmenting the size of EEG datasets using generative adversarial networks," in *2018 International joint conference on neural networks (IJCNN)*. IEEE, 2018, pp. 1–6.

Automated Classification of Rice Varieties

Dehai Lu
University Of Canberra
Canberra, Australia
U3276283@uni.canberra.edu.au

Jia Cui
University Of Canberra
Canberra, Australia
U3271434@uni.canberra.edu.au

***Abstract*— This study investigates the application of machine learning techniques for the classification of rice varieties based on morphological features. Utilizing a dataset enriched with diverse characteristics, we performed data preprocessing to eliminate redundant features. We evaluated five classification algorithms: Logistic Regression, Support Vector Machines (SVM), Random Forest, K-Nearest Neighbors (KNN), and Decision Tree. Model performance was assessed through 5-fold cross-validation, generating classification reports and visualizations. The results indicate that the performance of the SVM model is significantly better than other models. Therefore, based on the existing data, we developed an automated rice variety classification model using the SVM algorithm, achieving an accuracy of up to 92%. This study provides a robust model for agricultural applications, with future directions focusing on dataset expansion and real-time implementation strategies.**

Keywords—rice classification, machine learning, k-fold cross validation

I. INTRODUCTION

A. Background and Motivation

Rice is an essential food worldwide, and accurately identifying its different types is crucial for farming, quality control, and the rice trade. Each type of rice has unique characteristics like cooking qualities, nutritional value [1], and market price [2]. Creating a system that automatically categorizes rice based on measurable physical traits can improve efficiency, reduce human error, and maintain consistent quality in processing and distribution [3]. This project addresses a practical agricultural challenge while also providing an opportunity to explore key concepts in pattern recognition and machine learning, including feature extraction, model selection, and performance evaluation.

B. Problem Description

The aim of this study is to create a machine learning model that can accurately classify two rice varieties, Cammeo and Osmancik [4]. The model will analyze various physical traits of the rice grains, including area, perimeter, and eccentricity. This task represents a typical pattern recognition problem in machine learning, where the model must learn from grain features to effectively distinguish between the two varieties.

C. Implementationt Strategy

To devise a PRML solution for this issue, there are several crucial steps that need to be taken:

- Data Collection and Preprocessing: Assemble a dataset containing measurements of rice grains, with each one labelled as either Cammeo or Osmancik. Next, the data is checked for any missing values. If missing values are present, we choose to use mode imputation, as this is a classification model. Mode imputation is particularly advantageous because it retains the most common category in the feature, ensuring consistency in categorical data and minimizing distortion of the dataset's distribution. Then prepare the data by normalizing the feature values and dividing it into training and test sets.
- Feature Extraction and Selection: Identify and extract relevant features (such as area, perimeter, etc.) that describe each rice grain. Determine which features are most useful for the classification task.
- Model Selection and Training: Select appropriate machine learning algorithms for training on the dataset. Train the selected models on the training data, adjusting hyperparameters as necessary for optimal performance.
- Model Evaluation: Assess the model's performance using cross-validation and evaluate metrics such as accuracy, precision, recall, and F1 score. Validate the model's ability to generalize by testing it on unseen test data.
- Build a predictive model using the entire dataset for future use.

II. RESEARCH DESIGN

A. Dataset selection

The work utilizes the Rice Cammeo and Osmancik rice classification datasets from the UCI machine learning repository [5]. The dataset consists of 3510 samples, with 2010 samples for Osmancik rice and 1500 samples for Cammeo rice, and an additional 300 unseen samples. Given that this dataset meets the 10 times rule for feature-to-sample ratio, it is sufficiently large to train a model effectively. The dataset is well-balanced, presenting a binary classification problem.

The dataset contains the following features, the physical properties of each sample are:

- Area: The area of the rice grains.
- Perimeter: The perimeter of a grain of rice.
- Major Axis Length: the maximum axis length of a grain of rice.
- Minor Axis Length: The minimum axis length of a grain of rice.
- Eccentricity: Describes how stretched a grain of rice is.
- Convex Area: the area of the convex envelope of the rice grain.
- Extent: The ratio of the area of the grain to the area of the bounding box.
- Target: Cammeo or Osmancik

B. *Feature selection*

In the feature selection process, a heatmap can be used to visualize the correlation between different features [6]. If two features exhibit a very high correlation, such as a correlation value of 1 between 'Convex_Area' and 'Area,' it indicates that these features carry redundant information. Since they contribute similarly to the model, it becomes logical to consider reducing or removing one of these features to avoid multicollinearity. In this case, 'Convex_Area' could be excluded from the model without losing significant predictive power, simplifying the model and potentially improving its performance by reducing noise and overfitting risks.

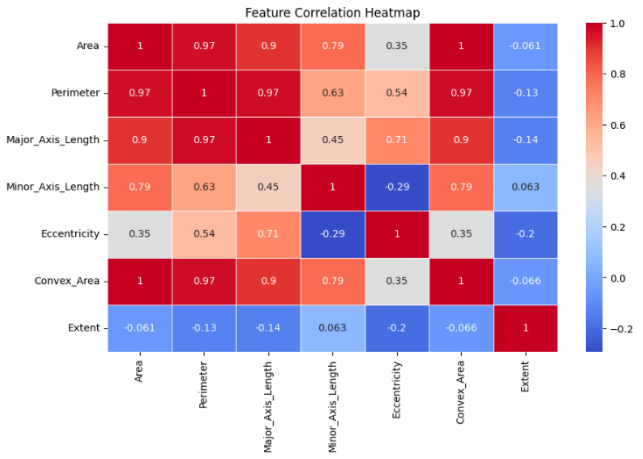


Fig. 1. Feature Correlation Heatmap

In the feature selection process, analyzing the area distribution plot reveals that only 'Extent' has a similar range across both rice categories, suggesting that it may not contribute significantly to distinguishing between them. Furthermore, examining the heatmap shows that the correlation values between 'Extent' and other features are close to 0, indicating minimal relationship with the remaining features. Given this information, it is reasonable to consider reducing or removing 'Extent' from the model. This can help simplify the model without losing critical predictive information, while potentially enhancing model performance by reducing unnecessary complexity.

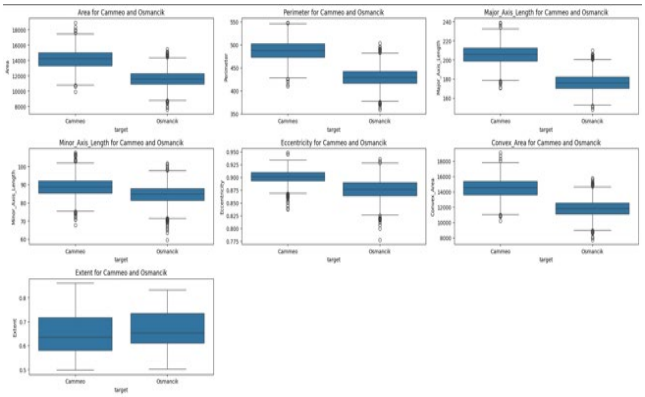


Fig. 2. Area Distribution Plot

The numerical range and scale differences between features can be visually observed from a distribution plot. If the distribution ranges of different features vary significantly (for example, the Perimeter feature ranges from 1 to 550, while Eccentricity ranges from 0 to 0.95), standardization becomes essential. This process ensures that all features are scaled consistently, preventing certain features from having an outsized impact on the model training results, while others may be diminished or overlooked. Standardization helps maintain the integrity of the model by giving equal importance to all features.

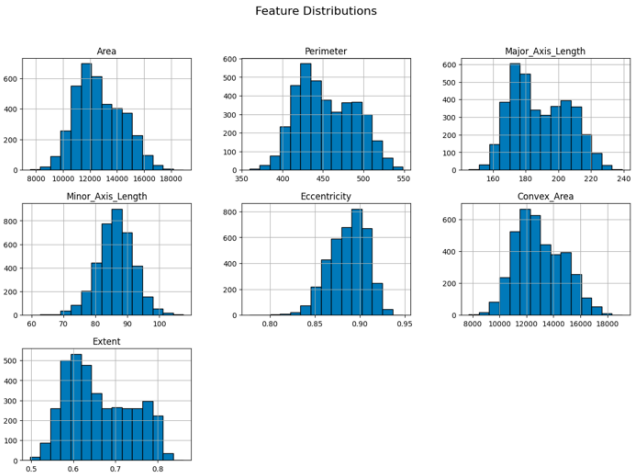


Fig. 3. Feature Distribution

C. *Data Preparation*

Data Cleaning: The work checks the data set to make sure there are no missing values, so we don't have to deal with any missing data. During the data cleansing step, our main goal is to ensure that the data is complete and in a consistent format for building models later.

Data conversion and feature scaling: During the data preprocessing phase, it is crucial to scale the features, such as through standardization, because the dataset contains multiple physical characteristics that are measured in different units or dimensions. By standardizing each feature to have a mean of 0 and a standard deviation of 1, we ensure that no single feature disproportionately influences the model. This process ensures that the model learns equally from all features, improving overall performance and helping it avoid biases toward features with larger scales.

D. *Machine Learning Algorithms for the Problem*

Logistic regression is a straightforward and effective linear classifier that works well for binary classification tasks. It provides insight into how each feature contributes to the classification. Since the rice classification task is binary (Cammeo vs. Osmancik), logistic regression can efficiently model the likelihood of a grain belonging to either category based on its attributes.

Support vector machines (SVMs) are well-known for identifying the optimal hyperplane that maximizes the separation between classes, particularly in high-dimensional spaces where the classes aren't easily divided by a straight line. SVM adapts to the dimensionality of the data by using linear separation in two-dimensional spaces, planar separation in three-dimensional spaces, and hyperplanar separation in higher dimensions [7]. By finding the ideal hyperplane for dividing data, SVM is effective for managing complex class boundaries, especially when rice grain characteristics overlap.

Random Forest, an ensemble learning technique, predicts outcomes by generating multiple decision trees [8]. In rice classification, Random Forest can handle intricate relationships between features and assess their importance. With various features like area, perimeter, and major axis length in the dataset, Random Forest can determine which

ones are most influential in distinguishing between rice types. Although Random Forest generally performs well, its results may fluctuate based on the dataset, requiring cross-validation to ensure the model's stability and performance.

The K-NN algorithm is a simple yet effective classification method that predicts the class of a test sample by examining the K-nearest neighbors from the training data [9]. In the rice classification context, k-NN can identify patterns in the feature space, making it suitable for smaller datasets like those containing rice grain measurements. It classifies samples based on the distance between test data points and their nearest neighbors in the training set. However, k-NN can become computationally intensive with large datasets and requires careful tuning of the K value to achieve optimal performance. For rice datasets with multiple features, k-NN helps identify the variety by comparing each test sample with its closest neighbors.

E. Evaluation Harness Developed

We selected five classification models for evaluation:

- Logistic Regression
- Decision Tree Classifier
- Support Vector Machine (SVM)
- Random Forest Classifier
- K-Nearest Neighbors (KNN)

To evaluate these models, we developed an evaluation harness based on 5-fold cross-validation using GridSearchCV [10]. This approach allows for tuning hyperparameters systematically for each model while assessing the models' generalization capabilities across multiple validation folds. The cross-validation process splits the training data into five subsets, where each subset is used as a validation set while the others serve as the training set. This prevents overfitting and provides a robust accuracy estimate.

For each model, we tested various hyperparameter combinations to identify the best configuration, optimizing the models for accuracy. After obtaining the best parameters through cross-validation, we trained the final model on the training set and evaluated it on the test set.

F. Optimization Methods to Improve Results

we applied standardization to the data before model training, which rescales features to have a mean of 0 and a standard deviation of 1. Standardization ensures that all features contribute equally to the model and prevents features with larger numerical ranges from dominating the model's decisions. This was particularly crucial for models like Logistic Regression and SVM, which are sensitive to feature scaling.

In addition to standardization, GridSearchCV was employed to tune hyperparameters, which enhanced the performance of each algorithm by exploring the most effective parameter combinations.

- For Logistic Regression, it was adjusted with different values of C (0.1, 1, and 10) and kernel types (including 'linear' and 'rbf').
- For the Decision Tree, we conducted hyperparameter tuning with a focus on two parameters: C, which varied between 0.1, 1, and 10, and solver, which included options such as 'liblinear' and 'saga'.
- For SVM, we tuned both C (0.1,1,10) and kernel (linear or RBF).
- For Random Forest, we fine-tuned the hyperparameters for n_estimators (choosing from 10, 50, and 100), max_depth (including None, 5, 10, and 15), and min_samples_split (set to 2, 5, and 10).
- For KNN, it was fine-tuned by adjusting the hyperparameters n_neighbors (ranging from 5 to 20) and weights (with options of 'uniform' and 'distance').

III. Implementation and Results Analysis

A. Reconfirming Feature Selection

Before implementing machine learning algorithms, reconfirming feature selection is an important step. In the initial phase, exploratory analysis of the dataset was performed to examine feature correlations, determining which features might be redundant. The features 'Convex_Area' and 'Extent' were identified as having high linear relationships with the target variable, leading me to verify their impact.

Multiple processed versions of the dataset were created by gradually removing redundant features. The specific versions included:

- Unchanged Version (train_test_data_1)
- Remove 'Convex_Area' (train_test_data_2)
- Remove 'Extent' (train_test_data_3)
- Remove 'Convex_Area' and 'Extent' (train_test_data_4)

Dataset Version	Logistic Regression	Support Vector Machine	Random Forest	K-Nearest Neighbors
Version 1 (Original Dataset)	0.9302	0.8704	0.9217	0.8689
Version 2 (Remove 'Convex_Area')	0.9245	0.8675	0.9188	0.9088
Version 3 (Remove 'Extent')	0.9302	0.8704	0.9174	0.8689
Version 4 (Remove Both)	0.9259	0.8675	0.9160	0.9088

Fig. 4. Reconfirming Feature Result

These results indicate that removing the feature 'Extent' had a relatively minor impact on model performance, particularly for the logistic regression model, which did not show a significant decrease in accuracy. This further validates the rationale and effectiveness of our feature selection process.

B. Classification Report

We generated classification reports for each model, which provided detailed statistics on precision, recall, F1-score, and support for each class. These metrics allowed for a more nuanced understanding of model performance beyond accuracy.

Metric	Logistic Regression	Decision Tree	SVM	Random Forest	KNN
Precision	0.95 (Cammeo)	0.94 (Cammeo)	0.96 (Cammeo)	0.95 (Cammeo)	0.95 (Cammeo)
	0.91 (Osmancik)	0.91 (Osmancik)	0.91 (Osmancik)	0.91 (Osmancik)	0.91 (Osmancik)
Recall	0.91 (Cammeo)	0.91 (Cammeo)	0.91 (Cammeo)	0.91 (Cammeo)	0.91 (Cammeo)
	0.95 (Osmancik)	0.94 (Osmancik)	0.96 (Osmancik)	0.95 (Osmancik)	0.95 (Osmancik)
F1-score	0.93 (Cammeo)	0.93 (Cammeo)	0.93 (Cammeo)	0.93 (Cammeo)	0.93 (Cammeo)
	0.93 (Osmancik)	0.92 (Osmancik)	0.93 (Osmancik)	0.93 (Osmancik)	0.93 (Osmancik)
Support	360 (Cammeo)	360 (Cammeo)	360 (Cammeo)	360 (Cammeo)	360 (Cammeo)
	342 (Osmancik)	342 (Osmancik)	342 (Osmancik)	342 (Osmancik)	342 (Osmancik)

Fig. 5. Classification Report

SVM achieved the best overall performance in terms of both F1-score and recall, indicating that it balanced precision and recall effectively, reducing false positives and false negatives.

C. *Visualization of Results and Insights from the Best-Performing Model*

We visualized the classification performance using bar plots of precision, recall, F1-score, and support across models. The visual comparison highlighted that SVM consistently achieved the highest F1-scores, indicating strong performance in both identifying true positives and minimizing false positives and negatives.

The SVM's higher recall for the minority class indicates it is particularly good at identifying more of the positive instances in the dataset. Meanwhile, its precision ensures that a high proportion of the predicted positive instances are indeed true positives. This balance makes it an excellent choice for the classification problem.

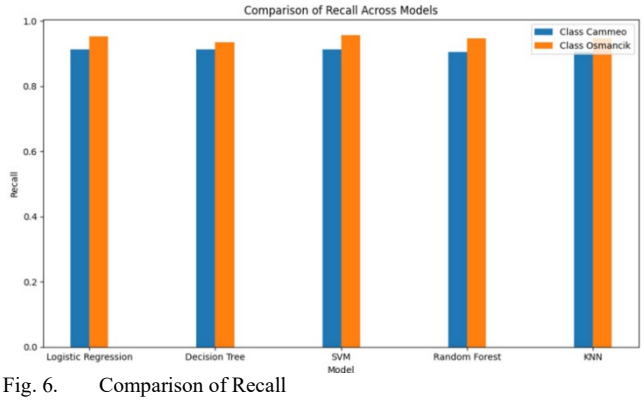


Fig. 6. Comparison of Recall

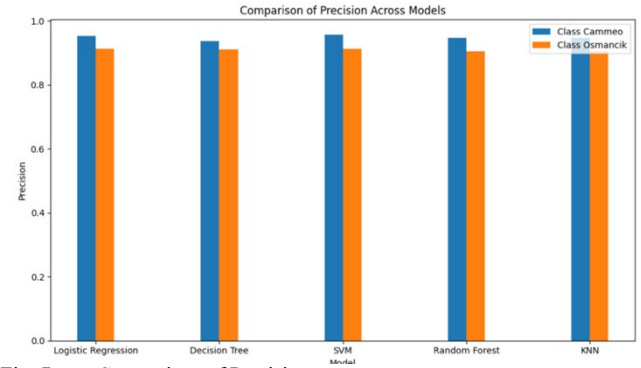


Fig. 7. Comparison of Precision

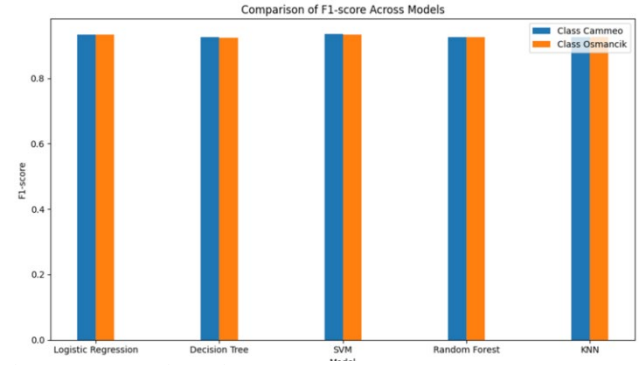


Fig. 8. Comparison of F-1 Score

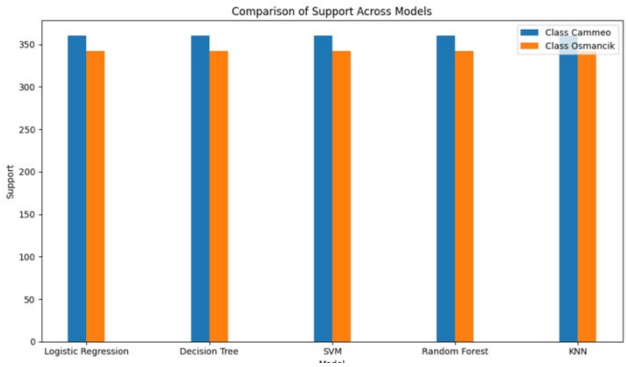


Fig. 9. Comparison of Support

In conclusion, SVM was determined as the best-performing model due to its strong F1-scores and balanced precision and recall, making it highly effective for the classification task after standardization and hyperparameter optimization.

D. *Visualizing Cross-Validation Accuracy*

To ensure the robustness of my models, I also visualized the 5-fold cross-validation accuracy. The results showed minimal differences in accuracy across the folds for each model, indicating that the models generalized well across different subsets of the data. This consistency suggests that the dataset is stable, and none of the models were significantly overfitting or underfitting.

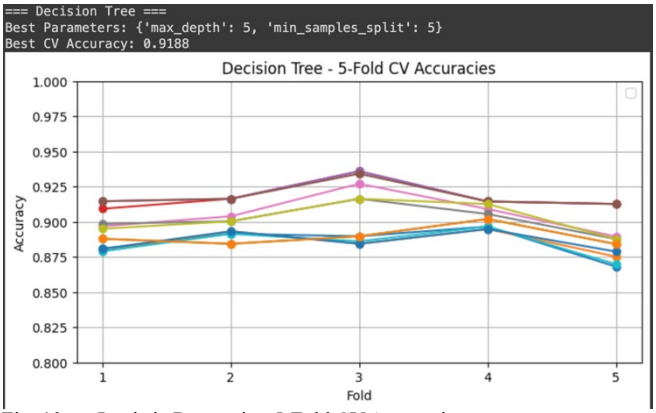


Fig. 10. Logistic Regression 5-Fold CV Accuracies

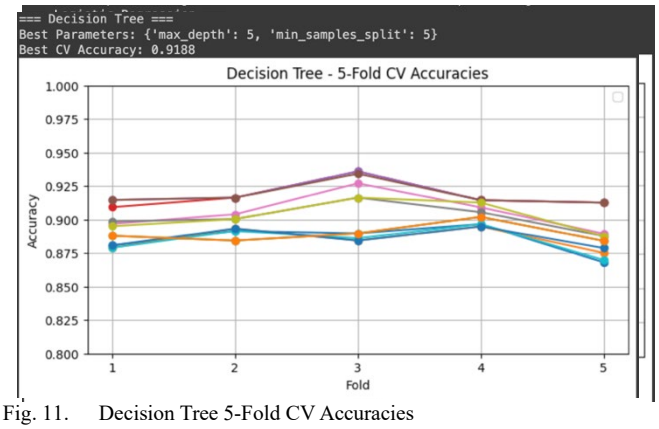


Fig. 11. Decision Tree 5-Fold CV Accuracies

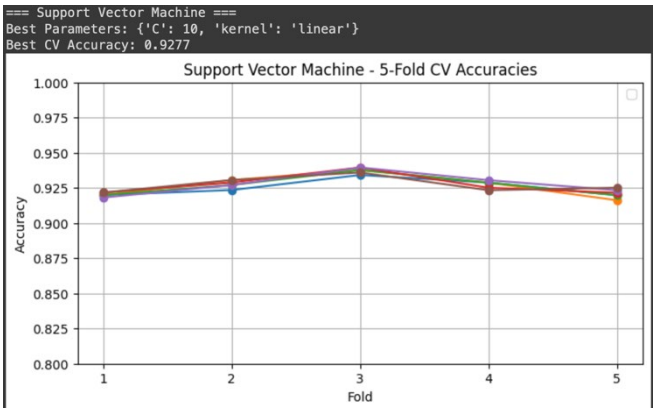


Fig. 12. Support Vector Machine 5-Fold CV Accuracies

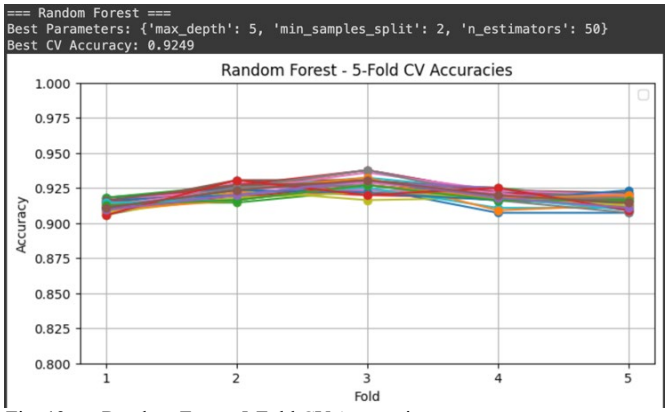


Fig. 13. Random Forest 5-Fold CV Accuracies

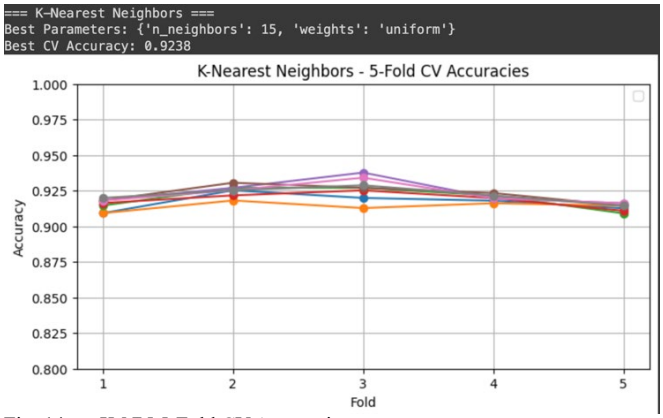


Fig. 14. K-NN 5-Fold CV Accuracies

By visualizing this, we could confirm that the model performance was reliable across different data splits, providing a solid estimate of real-world performance and avoiding misleading results from a single data split. This step reinforced our confidence in the robustness and reliability of the final models, especially the SVM, as the best choice for this classification task.

E. Make predictions on sample unseen data

- Removing Columns: remove both the 'Extent' and 'target' columns from unseen data.
- Standardization: The unseen data is then scaled using the scaler that was fitted on training data.
- Predictions: The predictions are made using the best_svm model, and the results are added to the unseen_data DataFrame as a new column called 'predicted_target'.
- Print Results: Finally, the first 20 rows of the DataFrame with the predictions are printed.

	predicted_target
1	Osmancik
2	Osmancik
3	Osmancik
4	Osmancik
5	Osmancik
6	Osmancik
7	Osmancik
8	Osmancik
9	Osmancik
10	Osmancik
11	Osmancik
12	Osmancik
13	Osmancik
14	Osmancik
15	Osmancik
16	Osmancik
17	Osmancik
18	Cammeo
19	Osmancik
20	Osmancik

Fig. 15. Predicted Result

F. Create a standalone model on entire dataset

- Process the data: Remove the 'Extent' column and separate the features from the target variable.
- Standardize the features: Use `StandardScaler` to standardize all features.
- Define the best parameters: Define the hyperparameters of the SVM model based on the best parameters obtained from previous training.
- Create and train the model: Train the SVM model using the entire dataset with the best parameters.
- Save the model: Use `joblib` to save the trained model as 'final_best_svm_model.pkl' for future use.

G. Using the Saved Model for Predictions

The 'final_best_svm_model.pkl' can be loaded for future predictions using the following steps:

- Load the Model: Use joblib.load () to retrieve the saved model from the file.
- Prepare Unseen Data: Ensure that the new data follows the same preprocessing steps as the training data, which includes removing any unnecessary columns and standardizing the features.
- Make Predictions: With the loaded model and preprocessed data, use the predict () method to generate predictions.

H. Further Tuning the Model with New Data

As new data becomes available, the model can be further tuned to improve its performance. This process involves:

- Collecting New Data: Gather additional labelled data to enhance the training set.
- Retraining the Model: Combine the new data with the existing training data and retrain the model. This can help the model adapt to changes and improve accuracy.
- Re-saving the Model: After retraining, save the updated model again to ensure that the improvements are preserved for future use.

IV. Ethical and Privacy Issues

In this study, although the dataset primarily focuses on the physical characteristics of rice grains, privacy concerns may arise if personal data is involved, such as linking grain data to farmers or producers. To address these concerns, it is essential to implement measures that ensure sensitive data cannot be traced back to individuals or specific organizations. Failure to adequately handle such data could lead to data

breaches or misuse, infringing on individuals' privacy rights. To safeguard privacy, sensitive data must be anonymized, and compliance with data protection regulations, such as GDPR, is crucial to ensure that data collection, storage, and processing meet legal requirements. Additionally, it is vital to inform data providers of the intended use of their data and obtain their consent during the data collection phase. By adopting these measures, personal privacy can be effectively protected, reducing potential risks.

Even in rice grain classification, biases may emerge from the data collection process, such as variations in sample quality or sources. If the training dataset is not representative, it may lead to biased predictions favoring certain types or regions of rice, potentially affecting the model's accuracy and resulting in unfair outcomes for stakeholders in the agricultural sector. To address this issue, it is essential to ensure that the dataset used is representative of various rice varieties, regions, and growing conditions. During data collection, diversity must be prioritized to encompass different sample types. Additionally, employing data augmentation techniques or increasing sample collection can help balance the dataset. Regular evaluation and validation of the model's performance are crucial for identifying and correcting any potential biases. By implementing these strategies, the fairness and generalizability of the model can be enhanced, improving its effectiveness across various application contexts.

V. Future Work

Expansion of the Dataset: Increasing the dataset size by collecting more samples of Cammeo and Osmancik rice varieties would enhance the model's generalization ability. More comprehensive data collection could also involve additional features or even different varieties of rice to make the model applicable in broader contexts.

Testing Advanced Models: While the current study explores several classification algorithms, future research can investigate more complex models such as Convolutional Neural Networks (CNNs) or ensemble methods like Gradient Boosting Machines (GBMs) [4]. These methods may capture more nuanced patterns in the data, leading to improved accuracy in classification.

Real-time Implementation: Another potential area of focus is the development of a real-time classification system. By integrating the trained model into a live environment, such as automated quality control systems in rice milling factories, the model can offer practical benefits. This step would require further optimization of both model efficiency and computational performance.

BIBLIOGRAPHY

[1] B. Y. A. S. P. S. A. A. W. Khalid Gul, "Rice bran: Nutritional values and its emerging potential for development of functional food—A review," *Bioactive Carbohydrates and Dietary Fibre,* vol. 6, no. 1, pp. 24-30, 2015.

[2] "The state of rice value chain upgrading in West Africa," *Global Food Security,* vol. 25, p. 100365, 2020.

[3] I. C. Y. S. T. Murat Koklu, "Classification of rice varieties with deep learning methods," *ELSEVIER,* vol. 187, p. 106285, 2021.

[4] A. I. K. U. a. E. I. I. U. Ilhan, "Classification of Osmancik and Cammeo Rice Varieties using Deep Neural Networks," *International Symposium on Multidisciplinary Studies and Innovative Technologies,* vol. 10, pp. 587-590, 2021.

[5] UC irvine, "rice+cammeo+and+osmancik," 2019. [Online]. Available: https://archive.ics.uci.edu/dataset/545/rice+cammeo+and+osmancik. [Accessed 2024].

[6] D. G. W. Z. Wubin Ding, "PyComplexHeatmap: A Python Package to Visualize Multimodal Genomics Data," *iMeta,* vol. 2, p. 3, 2023.

[7] K. S. Yu H, "SVM Tutorial-Classification," *Regression and Ranking,* vol. 1, pp. 479-506, 2012.

[8] S. E. Biau G, "A random forest guided tour," *Test,* vol. 25, pp. 197-227, 2016.

[9] J. A. Pandey A, "Comparative analysis of KNN algorithm using various normalization techniques," *International Journal of Computer Network and Information Security,* vol. 10, no. 11, p. 36, 2017.

[10] scikit learn developers, "GridSearchCV," 2024. [Online]. Available: https://scikit-learn.org/dev/modules/generated/sklearn.model_selection.GridSearchCV.html. [Accessed 2024].

Leveraging Technology to Bridge the Digital Divide and Optimize Resource Management in Smart Cities: A Case Study of Amaravati

Narsimha Rao Bandi
Faculty of Science and Technology
University of Canberra
Email: U3033322@uni.canberra.edu.au

***Abstract*—This study explores the potential of digital infrastructure and artificial intelligence (AI) in optimizing resource management, enhancing efficiency, and promoting sustainability in smart cities. The focus is Amaravati, India's greenfield smart capital city of the state of Andhra Pradesh. Addressing the digital divide and its socio-economic implications, the research examines how AI-driven technologies can be harnessed to create equitable, inclusive, and sustainable urban environments. By employing a mixed-methods approach, the study combines quantitative and qualitative data to investigate the interplay between AI, digital infrastructure, and social transformation. Drawing upon theoretical frameworks such as diffusion of innovation theory, actor-network theory, and digital divide theory, this research aims to inform policymakers, urban planners, and technologists on effective strategies to ensure the benefits of AI are accessible to all.**

INTRODUCTION

Smart cities, powered by digital infrastructure and AI, represent a paradigm shift in urban development, promising innovation, efficiency, and sustainability. However, the equitable distribution of these technological benefits and their broader social implications remain critical challenges. This research investigates how AI can be leveraged to optimize resource management while addressing the digital divide and fostering social transformation.

Amaravati, the capital of the state of Andhra Pradesh, a greenfield smart city in India, provides a unique case study for understanding the challenges and opportunities associated with AI implementation. By examining Amaravati's journey, this research offers insights into the role of AI and digital infrastructure in shaping sustainable, inclusive smart cities.

1. The objectives of this study are to:
   1. Examine the role of AI in optimizing resource management in smart cities.
2. Assess the impact of AI on the digital divide and social equity.
3. Explore Amaravati's potential as a model for AI-driven urban development.

KEY RESEARCH QUESTIONS

Investigating an AI based Framework drawing from case study of Amaravati: The framework is to cover critical aspects such as equity, ethics, socio-economic impacts, urban design, and environmental sustainability while addressing the core research topics as below.

1. **Bridging the Digital Divide**: How can AI-driven solutions ensure equitable access to digital infrastructure and improve digital literacy, particularly among marginalized groups in a smart city?
2. **Social and Economic Impacts**: How can AI-powered initiatives in smart cities be leveraged to create new economic opportunities while minimizing job displacement and reducing social inequality?
3. **Privacy and Security**: What best practices can be adopted to enhance data privacy, security, and citizen trust in the use of AI in smart cities?
4. **Urban Planning and Citizen Engagement**: How can AI-powered tools optimize urban planning for sustainability and inclusivity while actively engaging citizens in decision-making processes?
5. **Environmental Monitoring and Mitigation**: How can AI be employed to monitor and mitigate the environmental impacts of urban development in smart cities while improving energy efficiency?
6. **Ethical AI** and **Ethical Governance & Regulations**: What strategies can be implemented to ensure AI systems used in smart cities are transparent, accountable, and free from data bias to prevent perpetuating social inequalities?

What ethical guidelines and regulations are necessary to govern the development and deployment of AI technologies in smart cities?

LITERATURE REVIEW

Emerging Smart Cities: As urban populations grow and technological advancements accelerate, smart cities are becoming essential to address complex urban challenges. The Literature highlights their potential to enhance resource efficiency, sustainability, and quality of life through digital solutions. [1], [8], [11] & [12].

Digital Infrastructure and Artificial Intelligence: Smart city infrastructure increasingly integrates advanced AI applications for managing resources such as energy, water, waste, and transportation. However, challenges remain in scalability, integration, and ethical considerations. [4-6], [9] & [13].

Digital Divide and Social Transformation: While AI has the potential to optimize urban systems, it can exacerbate digital inequality, particularly among marginalized communities. Research underscores the need for strategies to bridge this divide and ensure equitable access to digital services. [2], [3} & [14]

Amaravati Capital City: Amaravati's development as a greenfield smart city highlights five key themes—amenities,

technology, people and skills, knowledge and innovation, and governance. As a model for sustainable urbanization, Amaravati offers valuable insights into AI-driven solutions for resource management and social equity. [7] & [10]

Concept of Smart Cities

Smart cities are urban areas that leverage advanced digital technologies, including artificial intelligence (AI), the Internet of Things (IoT), and big data, to enhance the quality of life for citizens, optimize resource management, and promote sustainable development. They aim to address urban challenges through data-driven solutions, fostering innovation and improving governance.

The Characteristics of Smart Cities include Technology-Driven Infrastructure, Data-Driven Decision-Making, Sustainability, Citizen-Centric Services, Connectivity, Economic Growth, Resilience and Participatory Governance.

The Models of Smart Cities include Technology-Centric, Citizen-Centric, Green City, Collaborative Governance, Economic Development and Hybrid.

Amaravati Smart Capital City aligns with multiple models and exhibits a range of characteristics based on its vision and development plans. It largely conforms to the Hybrid Model, combining technology-centric, green city, and collaborative governance approaches. It embodies characteristics such as sustainability, participatory governance, and connectivity while aiming to establish itself as a smart and sustainable urban centre.

Methodology

A mixed-methods approach will be employed to comprehensively address the research questions. This approach combines qualitative and quantitative data collection and analysis to provide a holistic understanding of the topic.

Data Collection:

1. Qualitative Data: Semi-structured interviews, focus group discussions, and document analysis.
2. Quantitative Data: Surveys and secondary data sources such as government reports and IoT sensor data.

Data Analysis:

1. Thematic analysis for qualitative data.
2. Descriptive and regression analysis for quantitative data.
3. Triangulation to integrate qualitative and quantitative findings

Results and Discussion (Anticipated)

The research is expected to:

1. Demonstrate how AI can optimize resource management, improving efficiency in energy, water, waste, and transportation systems.
2. Highlight AI's potential to bridge the digital divide through targeted interventions and inclusive policies.
3. Provide actionable insights into the socio-economic and cultural implications of AI adoption in Amaravati.

Conclusion and Future Work

This study aims to contribute to the discourse on smart cities and AI by providing a comprehensive analysis of Amaravati's journey. The findings will inform strategies for leveraging AI to create sustainable and inclusive urban environments.

Future Research Directions:

1. Long-term impact assessments of AI-driven urban initiatives.
2. Exploration of ethical considerations, including privacy and bias.
3. Development of policy frameworks for equitable AI deployment in smart cities.
4. Promotion of international collaborations to address global challenges.

Acknowledgements

I thank my primary supervisor Professor Dharmendra Sharma, who provided support & guidance in articulatmy Research Topic & Research Questions and in preparation of my Research Proposal.

References

[1] J.S. Gracias and G.S. Parnell, "Smart Cities – A Structured Literature Review", Smart Cities 2023 6, 1719-1743 https://www.mdpi.com/2624-6511/6/4/80

[2] Imran. A, "Factors Influencing Digital Inequality: A Scoping Review", https://www.researchgate.net/publication/367653084_Factors_Influencing_Digital_Inequality_A_Scoping_Review/link/63d9e4a264fc86063800bbb5/download?_tp=eyJjb250ZXh0Ijp7ImZpcnN0UGFnZSI6InB1YmxpY2F0aW9uIiwicGFnZSI6InB1YmxpY2F0aW9uIn19

[3] S. Shin, D. Kim, and S.A. Chun, "Digital Divide in Advanced Smart City Innovations", Sustainability 2021, 13, 4076 https://www.mdpi.com/2071-1050/13/7/4076

[4] H.M.K.K.M.B. Herath and M. Mittal, "Adoption of artificial intelligence in smart cities: A comprehensive review", International Journal of Information Management Data Insights 2022 https://www.sciencedirect.com/science/article/pii/S2667096822000192

[5] B. Choudhuri, P.R. Srivastava, S. Gupta, A. Kumar and S. Bag, "Determinants of Smart Digital Infrastructure Diffusion for Urban Public Services", in 2021 Journal of Global Information Management Volume 29 Issue 6

[6] S.S. Tippanavar and S.D. Yashwanth, "Unleashing the IoT Revolution in India : Trends, Advantages, Applications and Strategic Importance", Journal of ISMAC, December 2023, https://www.researchgate.net/publication/376313943_Unleashing_the_IoT_Revolution_in_India_Trends_Advantages_Applications_and_Strategic_Importance#:~:text=The%20technology's

[7] P. Singh, "Towards a New Paradigm of a Smart India: The case of Amaravati City in India's "Singapore" in the Making",

https://www.researchgate.net/publication/347945600_Towards_a_New_Paradigm_of_a_Smart_India_The_Case_of_Amaravati_City_in_India's_Singapore_in_the_Making
[8] Ministry of Housing and Urban Affairs, Government of India, "DATASMART CITIES", [online], Available: https://dsc.smartcities.gov.in/
[9] NITI Aayog, "National Strategy For Artificial Intelligence" - #AIFORALL, June 2018 [online], Available: https://www.niti.gov.in/sites/default/files/2023-03/National-Strategy-for-Artificial-Intelligence.pdf
[10] APCRDA, Municipal Administration & Urban Development Department, Government of Andhra Pradesh, India, "White Paper on Amaravati", July 2024 [online, Available: https://crda.ap.gov.in/crda_norifications/NOT07099565/01~White%20Paper%2003072024.pdf
[11] A. Bernard-Sasges and G. Bowen, "Global Cities Index 2024", Oxford Economics 2024, https://static.poder360.com.br/2024/05/oxford-economics-cidades-mai2024.pdf
[12] T. Singh, A. Solanki, S.K. Sharma, A. Nayyar and A. Paul, "A Decade Review on Smart Cities: Paradigms, Challenges and Opportunities", 2022 https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9801818
[13] R. Wolniak and K. Stecula, "Artificial Intelligence in Smart Cities – Applications, Barriers and Future Directions; A Review", Smart Cities 2024 https://www.mdpi.com/2624-6511/7/3/57
[14] S. Lythreatis, S.K. Singh and A. El-Kassar, "The digital divide: A review and future research agenda", Technological Forecasting & Social Change 2022 https://www.sciencedirect.com/science/article/abs/pii/S0040162521007903

A Proposal for Enhancing Socratic Questioning in Mental Health Therapy

1st Xiaomeng Wang
Faculty of Science & Technology
University of Canberra
Canberra, Australia
cherry.wang@canberra.edu.au

2nd Dharmendra Sharma
Faculty of Science & Technology
University of Canberra
Canberra, Australia
dharmendra.sharma@canberra.edu.au

3rd Dinesh Kumar
Faculty of Science & Technology
University of Canberra
Canberra, Australia
dinesh.kumar@canberra.edu.au

Mental health disorders are a growing global concern, affecting approximately 970 million people worldwide [1]. These conditions, which include depression, anxiety, and stress-related disorders, can significantly impact an individual's quality of life and productivity, leading to societal challenges. Psychological interventions, such as Cognitive Behavioral Therapy (CBT), have been shown to be effective in helping individuals identify and alter negative thought patterns and behaviors that contribute to mental health challenges [2], [3]. However, many people face barriers to accessing or engaging with these treatments, including high costs, stigma, and a limited availability of mental health professionals [4]. As a result, there is an urgent need for scalable and accessible solutions to improve mental health care.

Recent research has explored the potential of Large Language Models (LLMs) in supporting Cognitive Behavioral Therapy (CBT) [4]–[6]. One promising direction is the use of LLM agents as virtual therapists that engage clients in therapeutic conversations, often incorporating psychological methods such as Socratic questioning [7], [8]. Socratic questioning, a fundamental method in CBT, encourages clients to critically examine their beliefs and thought patterns through open-ended questions [9]. While much of the existing research has focused on generating the content of Socratic questions [10], predicting Socratic question strategies [7], and evaluating the overall effectiveness of conversations including Socratic questions using tools like the Socratic Dialogue Rating Scale & Coding Manual [11] and the Cognitive Therapy Rating Scale [12], little attention has been paid to the quality of individual Socratic questions. Specifically, the way these questions are phrased and delivered—encompassing aspects such as neutrality, empathy, and tone—can significantly influence therapeutic outcomes [13]. The effectiveness of Socratic questioning hinges not only on its content but also on the subtleties of its delivery. As such, focusing on the linguistic attributes of Socratic questions could lead to more effective interventions, especially in contexts where AI-driven solutions are employed to support mental health care.

To address this gap, we propose the following approach: (1) We will develop a theoretical framework for evaluating the quality of individual Socratic questions in mental health counseling, with a focus on the linguistic attributes that affect how these questions are phrased and delivered. This framework will be grounded in psychological literature and informed by expert insights from clinical psychologists, ensuring its relevance and applicability in therapeutic contexts. (2) A dataset will be collected, consisting of scenarios and paired Socratic questions, each labeled as either high quality or low quality. Annotations based on key linguistic attributes, such as neutrality and empathy, will be provided by annotators with psychological expertise. (3) Automatic evaluation models will be developed to classify the quality of Socratic questions, leveraging these linguistic attributes. (4) We will explore the potential of LLMs to generate enhanced versions of low-quality Socratic questions. The proposed approach will be validated through both automatic metrics and human evaluation to assess the accuracy of quality detection and the effectiveness of the enhanced questions.

The proposed work can be further utilized in conversational therapy to enhance the quality of AI-generated questions, ultimately improving the effectiveness of virtual therapists in mental health care.

References

[1] WHO. (2022) World mental health report: Tranforming mental health for all. Accessed on January 20, 2024. [Online]. Available: https://www.who.int/teams/mental-health-and-substance-use/world-mental-health-report

[2] S. G. Hofmann, A. Asnaani, I. J. Vonk, A. T. Sawyer, and A. Fang, "The efficacy of cognitive behavioral therapy: A review of meta-analyses," *Cognitive therapy and research*, vol. 36, pp. 427–440, 2012.

[3] A. T. Beck, *Cognitive therapy and the emotional disorders*. Penguin, 1979.

[4] A. Sharma, K. Rushton, I. W. Lin, D. Wadden, K. G. Lucas, A. S. Miner, T. Nguyen, and T. Althoff, "Cognitive reframing of negative thoughts through human-language model interaction," *arXiv preprint arXiv:2305.02466*, 2023.

[5] M. Maddela, M. Ung, J. Xu, A. Madotto, H. Foran, and Y.-L. Boureau, "Training models to generate, recognize, and reframe unhelpful thoughts," *arXiv preprint arXiv:2307.02768*, 2023.

[6] X. Wang, D. Sharma, and D. Kumar, "Cognitive reframing via large language models for enhanced linguistic attributes," in *The Second Tiny Papers Track at ICLR 2024*, 2024.

[7] S. Lee, S. Kim, M. Kim, D. Kang, D. Yang, H. Kim, M. Kang, D. Jung, M. H. Kim, S. Lee, K.-M. Chung, Y. Yu, D. Lee, and J. Yeo, "Cactus: Towards psychological counseling conversations using cognitive behavioral theory," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 14 245–14 274. [Online]. Available: https://aclanthology.org/2024.findings-emnlp.832/

[8] M. Xiao, Q. Xie, Z. Kuang, Z. Liu, K. Yang, M. Peng, W. Han, and J. Huang, "HealMe: Harnessing cognitive reframing in large language models for psychotherapy," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 1707–1725. [Online]. Available: https://aclanthology.org/2024.acl-long.93/

[9] C. A. Padesky, "Socratic questioning: Changing minds or guiding discovery," in *A keynote address delivered at the European Congress of Behavioural and Cognitive Therapies, London*, vol. 24, 1993.

[10] B. H. Ang, S. D. Gollapalli, and S. K. Ng, "Socratic question generation: A novel dataset, models, and evaluation," in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 2023, pp. 147–165.

[11] C. Padsesky, "Socratic dialogue rating scale and coding manual," 2020.

[12] G. A. Aarons, E. A. Miller, A. E. Green, J. A. Perrott, and R. Bradway, "Adaptation happens: a qualitative case study of implementation of the incredible years evidence-based parent training programme in a residential substance abuse treatment programme," *Journal of Children's Services*, vol. 7, no. 4, pp. 233–245, 2012.

[13] M. Neenan and S. Palmer, *Cognitive behavioural coaching in practice*. Routledge, 2012.

Optimising Machine Learning Models for Weather Forecasting: A Case Study on Temperature Prediction

Quang Anh Nguyen
University of Canberra
Canberra, Australia

Thanh Dat Nguyen
University of Canberra
Canberra, Australia

***Abstract*— This report details the implementation and optimisation of machine learning models to predict temperature using weather data. Three key algorithms—Linear Regression, Decision Tree, and K-Nearest Neighbors—were evaluated using cross-validation. Hyperparameter tuning using GridSearch is used to choose the best performance parameters. The model was further validated on unseen data, and a standalone version was created for future predictions. Ethical and privacy considerations, including data bias and model transparency, were addressed. The final Decision Tree model demonstrates the effectiveness of ensemble learning for predictive modelling.**

Keywords*— Machine Learning, Temperature Prediction, Weather Data, Linear Regression, Decision Tree, K-Nearest Neighbors, Cross-Validation, Hyperparameter Tuning, GridSearchCV (key words***)**

I. Introduction

Accurate predictions of temperature based on weather data are crucial for several industries, including agriculture, energy, and disaster management. Machine learning models can help predict future temperatures based on past weather data, leading to better decision-making processes. The primary goal of this project is to evaluate and improve different machine learning models to predict temperature using weather data. By fine-tuning algorithms and leveraging ensemble techniques, we aim to develop a high-performing model that generalizes well to unseen data. This report covers the end-to-end process, from data preparation and algorithm selection to optimisation and final predictions.

II. Problem Definition and Dataset

A. Problem to be Solved

The project aims to predict temperature based on weather data features such as humidity, wind speed, and pressure. The goal of this project is to develop a machine-learning model that can accurately predict weather conditions by learning from historical data available on Kaggle. This is a regression problem, where the target variable is the temperature, and the input features include various meteorological factors [1]. In today's world, accurate temperature forecasting is crucial for various purposes, such as farming and managing natural disasters. This research uses machine learning to explore how weather factors, for instance, humidity and wind, affect temperature, aiming to develop a reliable and practical approach to weather prediction. Even though weather forecasting has improved, it is still challenging to predict the accurate temperature. The goal is to provide more accurate forecasts, which helps with predicting natural disaster or planning for cultivation and agriculture.

B. Dataset Description

Firstly, for this project, the weather dataset was selected by the research team from Kaggle. This dataset is called Weather in Szeged 2006-2016. This dataset includes information on historical weather observations from 2006 to 2016 in Szeged, a city in Hungary. This data will be used by the research team to build machine learning models used for weather forecasting purposes.

Formatted Date: timestamp of the observation. (2016-09-09 19:00:00.000 +0200).

Summary: overall weather condition at the time of observation (Overcast, Partly Cloudy).

Precip Type: The type of precipitation (rain, snow).

Temperature (C): actual temperature in degrees Celsius.

Apparent Temperature (C): perceived temperature in degrees Celsius.

Humidity: humidity as a percentage.

Wind Speed (km/h): wind speed in kilometres per hour.

Wind Bearing (degrees): The direction from which the wind is coming, in degrees.

Visibility (km): The visibility distance in kilometres.

Loud Cover: cloud cover.

Pressure (millibars): Atmospheric pressure in millibars.

Daily Summary: summary of the weather conditions for the day (Partly cloudy starting in the morning).

A total of 96,453 rows of data were used, with an 80-20 train-test split. Pre-processing included scaling the features using MinMaxScaler.

.

III. Approach and Implementation

A. Choosing Algorithms

- Linear Regression: Provides a baseline due to its simplicity and ability to model linear relationships.

- Decision Tree: A non-linear algorithm that provides flexibility in handling complex patterns in the data.
- K-Nearest Neighbors (KNN): Relies on the proximity of data points to make predictions and can capture local patterns effectively.

B. Dataset Pre-processing

Proper data preprocessing is essential for improving model performance. This step also includes handling missing values. However, in this specific dataset, there is no missing value so the research team decide to forward to further step. Before training, the dataset was pre-processed in these steps:

- **Scaling**: The features were scaled using MinMaxScaler to bring all features into the same range. This step help to reduce the risk of over-fitting, as the range is calculated in the same range.

- **Train-Test Split**: To validate the models, the dataset will be split into training and test sets. The training set will be used to train the models, while the test set will be reserved for evaluating the models' performance on unseen data. The data was split into 80% training and 20% testing to evaluate model generalisation [2]. It seems that 80-20 train-test split is a common practice in machine learning because it provides sufficient training data and enough data to evaluate the performance of the machine learning model.

C. Data Visualisation

A line plot was used to identify key features highly correlated with temperature. Visualiations, including histograms and scatter plots, helped analyse the distribution of the data.

IV. Evaluation of Algorithms

A. Data Split and Cross-Validation

The dataset was split into training (80%) and test (20%) sets. We employed cross-validation on all models to avoid overfitting and ensure evaluation, with MSE as the evaluation metric.

B. Model Results

Table 1: Model Performance Metrics

Model	R-squared Score	Mean Squared Error	Mean Absolute Error
LR	0.999216	0.907180	7.745977
KNN	0.991453	0.111735	0.216499
DTM	0.999944	0.005200	0.019767

As the result seen above, MAE of DTM (0.019) < MAE of KNN (0.216). This is a big difference, so we will choose DTM as the main algorithm for this assignment.

Figure 1: Overfitting Investigation

```
dtm = DecisionTreeRegressor(random_state=5)
dtm_scores = cross_val_score(dtm, X_train, y_train, cv=5, scoring='neg_mean_squared_error')
dtm_scores
Executed at 2024.10.29 14:05:56 in 1s 236ms
```

```
array([-0.00723608, -0.00605308, -0.00467673, -0.00677309, -0.00728073])
```

Table 2: Mean Square Error

Fold	Mean Squared Error (Negative)
1	-0.007236
2	-0.006053
3	-0.004677
4	-0.006773
5	-0.007281

These small negative MSE values indicate that the Decision Tree Regressor model is performing well on the test dataset when performing cross-validation.

V. Improving Results

To improve the results, Hyperparameter Tuning using GridSearch to identify the best parameters for each learning model [3]. For the Decision Tree Regressor, a parameter grid for random_state and splitter is used to find the best-performing parameters. As a result, the best parameters for Decision Tree Regressor are:

- **random_state = 2**
- **splitter = 'best'**

For K-Nearest Neighbors (KNN): A range of values for n_neighbors is tested to find the best neighborhood size. The best parameter found is:

- **n_neighbors = 4**

Finally, this was applied to the KNN model.

VI. Classification Report and Model Selection

A. Finalising the Best Model:

Figure 2: Scores for decision Tree model

```
Decision Tree Regressor - MSE: 0.004371660551744533 MAE: 0.019448908241612245
K-Nearest Neighbors Regressor - MSE: 0.07080084920941193 MAE: 0.1622878256412029
```

Based on the results, the Decision Tree Regressor was chosen as the best-performing model. The classification report compared its performance against the other models using MSE and MAE. Decision Tree Regressor performed better than the K-Nearest Neighbors Regressor, with an MSE: 0.0043716605517445330 and MAE: 0.019448908241612245. Based on the above results, it is easy to see that the DTM model is better than KNN based on the two indexes MSE and MAE.

B. Visualisation of Results

We plotted Actual vs. Predicted values for the Decision Tree model on test sets. The residual plot showed random distribution, indicating that the model captured most of the data's underlying patterns.

FIGURE 3: ACTUAL AND PREDICTED TEMPERATURE

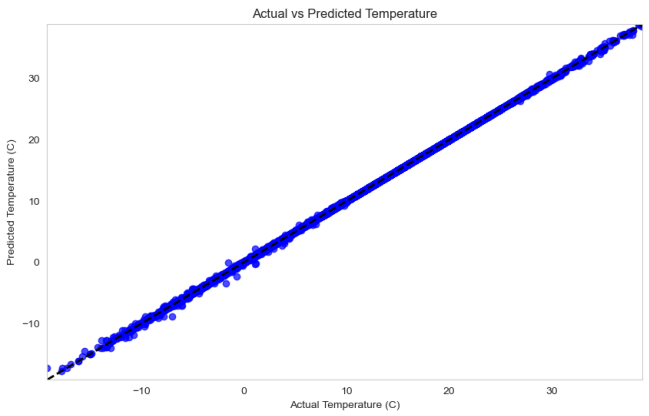


TABLE 3: PREDICTED DATA AND SAMPLE INPUT DATA

SAMPLE INPUT DATA	PREDICT DATA
7.8 ℃	9.34 ℃
10.4 ℃	10. 4℃
12.8 ℃	12.8℃

VII. PREDICTIONS ON UNSEEN DATA

The Decision Tree model was used to predict outcomes for unseen data, which was held back from the original test set. The predictions were compared to the actual outcomes. These predictions illustrate a similarity between predicted and actual temperatures.

FIGURE 4: DATA PREDICTION RESULTS

```
# Make predictions for the unseen data
y_pred_unseen = best_model.predict(unseenData_scaled)

# Display the prediction results
for i, pred in enumerate(y_pred_unseen):
    print(f'Prediction for sample {i + 1}: Temperature = {pred:.2f} °C')
```

Executed at 2024.10.29 14:08:17 in 5ms

```
Prediction for sample 1: Temperature = 9.34 °C
Prediction for sample 2: Temperature = 10.40 °C
Prediction for sample 3: Temperature = 12.80 °C
```

.

FIGURE 5: NEW INPUT DATA FOR PREDICTION

```
unseenData = [
    [7.8, 0.93, 10.34, 1010],
    [10.4, 0.7, 12.12, 1020],
    [12.8, 0.6, 14.12, 1030]
]
```

VIII.. PREDICTIONS ON UNSEEN DATA

A standalone version of the Decision Tree model was created by training it on the entire dataset (training and test sets combined).

FIGURE 6: DATA PREDICTION RESULTS

```
Mean Absolute Error (MAE): 0.00000
Mean Squared Error (MSE): 0.00000
R-squared (R²): 1.00000
```

The above results show that the trained model has a very impressive performance.

IX. SAVING AND FUTURE DEVELOPMENT

A. Model Development

The saved model can be updated with new data as it becomes available. New observations can be added to the training set, and the model can be retrained to improve its predictive power. This process is combining the new data with the existing training dataset.

B. Use Case

The saved model can be loaded for use in real-world applications, predicting temperatures based on incoming weather data. The model was saved using joblib library. This library helps to save this model to a file. By saving the trained model, it can be shared, archived, or used later in the future for weather forecasting.

X. RESEARCH OUTCOME

The study demonstrated that using Hyperparameter Tuning with GridSearch to identify the optimal parameters for each learning model resulted in more accurate temperature predictions.

XI. ETHICAL AND PRIVACY CONCERNS

A. Ethical Considerations

Bias: The model must be regularly audited to avoid bias toward certain weather conditions or geographic locations.

Accuracy in Decision-Making: Incorrect predictions in applications like disaster management could have severe consequences

B. Privacy Concerns

Although the weather data did not contain personal information, future applications might require location-based data, which should comply with privacy regulations (for example: GDPR)

XII. CONCLUSION

This report demonstrates the successful implementation and optimization of machine learning models for predictive modeling using weather data. By comparing various models and applying hyperparameter

tuning and ensemble methods, the Decision Tree model was identified as the best-performing algorithm. The model was saved and is ready for use in future predictions, with continuous updates possible as new data is collected.

XIII. Application

Urban Planning: This model can assist the government in preparing for floods, heatwaves, and other extreme weather

Agriculture: Supports farmers to decide when it is suitable for planting and watering crops.

Airline Operations: Enhances flight safety and reduces delays by forecasting storms and turbulence.

Disaster Management: Offers early warnings for cyclones and other severe weather. This AI model can also apply in mobile phone apps, which will give the users an SOS message.

XIV. Future Work

In the future, it can be considered to explore deep learning models such as LSTM for time-series weather predictions [4].
Moreover, implementing more advanced stacking techniques for further improving performance will be benificial. A possible future extension of this work could be to improve the model to predict specific weather events, whether it will rain tomorrow or if a cyclone is coming.

References

[1] B. Bochenek and Z. Ustrnul, "Machine Learning in Weather Prediction and Climate Analyses—Applications and Perspectives," *Atmosphere,* vol. 13, no. 2, p. 180, 2022, doi: 10.3390/atmos13020180.

[2] A. Rácz, D. Bajusz, and K. Héberger, "Consistency of QSAR models: Correct split of training and test sets, ranking of models and performance parameters," *SAR and QSAR in environmental research,* vol. 26, no. 7-9, pp. 683-700, 2015, doi: 10.1080/1062936X.2015.1084647.

[3] G. S and S. Brindha, "Hyperparameters Optimization using Gridsearch Cross Validation Method for machine learning models in Predicting Diabetes Mellitus Risk," 2022: IEEE, pp. 1-4, doi: 10.1109/IC3IOT53935.2022.9768005.

[4] M. M. Patil, P. M. Rekha, A. Solanki, A. Nayyar, and B. Qureshi, "Big data analytics using swarm-based long short-term memory for temperature forecasting," *Computers, materials & continua,* vol. 71, no. 2, pp. 2347-2361, 2022, doi: 10.32604/cmc.2022.021447.

Energy Optimisation in modular Data Centers using Machine Learning Techniques

Jyotsna Gaur
Faculty of Science and Technology
University of Canberra
Canberra, Australia
jyotsna.gaur@canberra.edu.au

Dharmendra Sharma
Faculty of Science and Technology
University of Canberra
Canberra, Australia
Dharmendra.Sharma@canberra.edu.au

***Abstract*—One of the most complex challenges of running and managing a Data Center is optimising the energy needed to run the facility. Each of the Data Center components such as Power subsystems, Uninterruptible power supplies (UPS), Backup generators, Ventilation and cooling equipment, Fire suppression systems and Building security systems consume a large amount of energy. Many aspects of Data Center technology require high amounts of energy consumption and this needs efforts in order to be managed efficiently. This paper describes design science research and a conceptual framework for building a machine learning model to predict and optimise energy consumption in a Data Center facility. Firstly, energy simulation will be carried out on different modular data center arrangements to get a time series output for sensible cooling load. Then this time-series based cooling load output will be used for data modelling. The model will be based on developing a ARIMA, Convolutional Neural Network (CNN) and LSTM model to predict the sensible cooling load of a modular data center for different time intervals. The paper also explores computational fluid dynamics and its simulation which can be used in gathering the micro-climatic data for the Data Center.**

Keywords—Energy optimisation in Data Center, Machine Learning in Data Center, Power Usage Effectiveness (PUE), Energy Load, Green IT, Power efficiency, Data Center Cooling Simulation.

I. INTRODUCTION

A Data Center is a physical location that stores computing machines and their related hardware equipment. It contains the computing infrastructure that IT (Information Technology) systems require, such as servers, data storage drives, and network equipment. It is the physical facility that stores any company's digital data. [1] Previously every company invested in and managed their own data center facility. Over the years, the size and power requirements of computers has reduced. But the demand for these services-e.g., computing, storage, networking with minimum latency; has risen exponentially. This has led to an increase in the rapid expansion of Data Center infrastructure and operational costs as well as energy consumption. Data centres contribute around 0.3 per cent to overall carbon emissions, while the ICT sector accounts for more than 2 per cent of global emissions. [2] This percentage is growing exponentially as the years go by. Rising energy costs and environmental responsibility has put tremendous pressure on the Data Center industry to improve its operational efficiency.

This paper gently introduces the planned research design ideas for optimisation of energy usage in Data Centers. This research will lead to a better understanding and management of Data Centers. It will help bring down the running energy costs of their cooling infrastructure. Investing in this research will lead to better optimized Data Centers. The stakeholders can be localized Data Center companies which function locally in Australia or the global technology players like Google and Amazon which have their Data Centers spread across various regions in the world. The infographic below describes certain characteristics of a modern data center.

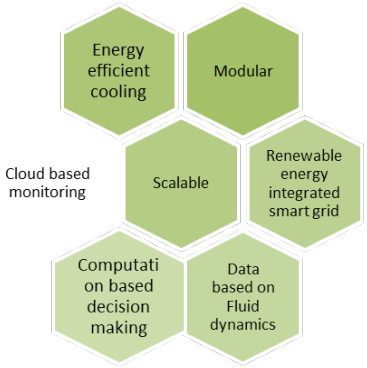


Figure 1: Smart Data Center features, source: Author

A crucial aspect of a smart data center is its modularity. Modular data centers offer various arrangements to cater to different computing or storage needs and preferences. The following are some possible modular arrangements being used in the industry:

i. All-in-One Module: This module includes all essential components, such as IT equipment, power and cooling systems, and security features, in a single enclosure. It's a self-contained unit that can be easily deployed and expanded as needed. Its size is typically in length 5m by width 3m by height 2.5m.

ii. IT Module and Power/Cooling Module: Separate modules for IT equipment and power/cooling systems. This allows for more flexibility in scaling IT capacity independently of power and cooling infrastructure. Each separate module has a typical size of 4m by 2.5m by 2.5m.

iii. Containerized Modules: Data center components are housed in standardized containers, providing a modular and portable solution. These containers can be stacked or arranged side by side to scale the data center

capacity. A 20ft container has a size of 6.1m by 2.4m by 2.6m.

iv. Scalable Rack Pods: The data center is organized into scalable rack pods. Each pod contains a set number of server racks along with associated power and cooling infrastructure. Additional pods can be added to scale capacity. A pod has the dimensions 4m by 4m by 2.5m.

v. Row-Based Modules: Data center components are organized in rows, with each row containing a set number of server racks, power distribution units, and cooling systems. This modular approach allows for incremental expansion. A typical row size is 10m by 4m by 2.5m.

vi. Function-Specific Modules: Modules are designed for specific functions, such as computing, storage, or networking. This allows organizations to customize the data center based on their specific needs. A typical storage module has the size 6m by 3m by 3m.

vii. Edge Data Center Modules: Modular setups designed specifically for edge computing, allowing organizations to process data closer to the source. These modules are often smaller and more self-contained. Each module is 3m by 2m by 2.5m.

viii. Prefabricated Modules: Components of the data center are prefabricated off-site and assembled on-site. This approach can save time and resources during the construction phase. The prefabricated modules have the dimensions of 8m by 4m by 3m.

The choice of a modular arrangement depends on factors such as scalability requirements, space constraints, mobility needs, and the specific goals of the organization deploying the data center.

II. Literature Review of Energy Efficiency Incrementing Optimization Methods

A. *Thermal optimisation*

Thermal optimization of the data centers is important to maintain the temperature, which leads to optimum energy conservation. In their paper, more than 10 data centers have been studied by Ni and Bai [1] by considering the energy consumption of the cooling systems. It has been observed by them that more than 50% of the data centers were not operating efficiently, leading to energy wastage.

This wastage can be mitigated by developing passive and active strategies like humidity control, temperature control, optimisation of airflow and economic cycles. Some of these energy saving techniques have been compared by Oro et al [4]. They found that integrating the energy grid with recovery of waste heat through an absorption-refrigeration system or through direct power generation and indirect power generation made a huge difference in energy usage.

Some passive cooling techniques for data centers have been studied by Daraghmeh and Wang [2]. These techniques included applications of the air, water and heat pipe which were incorporated with other operations like solar systems, absorptions, evaporative cooling, and geothermal cooling.

B. *Machine Learning and AI based optimisation*

Multiple technologies like network optimizations, servers and processors have been proposed to improve the efficiency of the data centers. The usage of all these technologies can generate historical databases that can later be analyzed and modelled for making future predictions.

The most notable work is where a neural network methodology has been proposed by Gao [5] for energy optimization of the data centers. As a result of multiple configurations and non-linearity, it is difficult to optimise the energy efficiency of the data centers, which is where a neural network seems to be the most useful model. The model gets trained from the actual operational data for modelling the performance of the data center and to predict the Power Usage Effectiveness (PUE) accurately. A PUE of 1.1 has been obtained with an error of 0.4%. This work has been tested and validated in the Google Data centers and it has been seen that AI techniques are effective ways to optimise the performance and improve energy efficiency.

The table below summarizes some of the machine learning techniques that have been researched in the past for Data Centres:

TABLE 1: COMPARISON OF VARIOUS ML ALGORITHMS FOR DATA CENTERS, SOURCE: KUMAR ET AL [3]

S. No.	Type of Algorithm used	Comments	Author
1	ANN	Power Usage effectiveness is calculated. It is validated in Google Data centers	Gao [5]
2	Naïve Bayes	Implemented on rack mounted servers on data center testbed	Marco et al [6]
3	Multivariable Linear Regression	FIESTA IoT and Resl DC testbed were used for the implementation. ANN and SVM were considered as a future scope	Smpokos et al [7]
4	Reinforcement learning	Airflow regulation was performed. It has a high computation time.	Lazic et al [10]
5	ANN	The effect of the cooling parameters was not considered.	Yu et al [9]
6	Deep Deterministic Policy Gradient	It has been used for both solving simple optimisation and complicated optimisations like the data center cooling	Lillicarp et al [8]

The above machine learning techniques and their selection vary on a case-by-case basis. But there is a strong presence of neural networks while using data that is organic in nature, and this is true for energy output data. The variables can be based on different micro-climatic factors in a data center. These can be the ambient temperature, humidity, water consumption, fan speeds, radiation, etc. Analysing the microclimate of a data center premise and estimating its cooling effectiveness is not a novel but definitely a niche aspect of research related to energy optimisation. The proposed study will analyse the climatic factors and different energy outputs to evaluate the

most effective variables contributing to cooling of data centers.

III. METHODOLOGY

The methodology followed for conducting the experiment was based on a typical schematic flow of processes involving methods used for pattern recognition and machine learning.

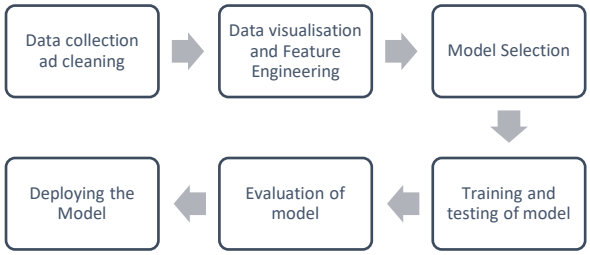


Figure 2:Steps followed for project design. Source: Author

The research will take a quantitative approach towards research design. The study will have a quantitative approach because it is mostly dealing with quantitative variables and the data is mostly numeric. In the quantitative approach, there will be a machine learning model based on neural networks that will be built. The model will have input variables pertaining to micro-climatic factors of Data Center premises. The research method will be based on Model Building using machine learning.

Based on previous research paper readings, models were selected to model the data. The evaluation was done based on metrics such as accuracy, RMSE and the best model was decided.

IV. EXPERIMENTS

A. Energy simulation on a Modular arrangement

For the purpose of energy simulation, first the building information models were created using Rhino3D and EnergyPlus software. Each type of modular data center was created into a 3D model. Next, its energy simulation was carried out using the weather epw files available on ladybug tools and epw map. These files contained weather conditions of the specific location. In our case, the location was Canberra, and the weather conditions were modelled accordingly. This simulation resulted in a time series cooling load data which then became the base dataset for running machine learning algorithms on it. The following table gives us an energy consumption estimate found out through energy simulations for different types of modular arrangements:

TABLE 2: SENSIBLE COOLING LOAD FOR DIFFERENT CONFIGURATIONS

S.No.	Modular configuration	Dimensions (LxWxH)(in metres)	Sensible Cooling Load (in Watts)
1	All-in-One Module	5x3x2.5	58773.36 W
2	IT module and cooling module	4x2.5x2.5 each	52345.45 W
3	Containerised Modules (20ft standard shipping container)	6.1x2.4x2.6	59230.34 W
4	Scalable Rack Pods	4x4x2.5 pod	51098.23 W
5	Row based modules	10x4x2.5 row	62342.45W
6	Function specific modules	6x3x3 storage	59456.34 W
7	Edge data center modules	3x2x2.5	49773.36 W
8	Pre fabricated modules	8x4x3	60392.47W

B. Dataset

The selected Dataset is a set of observations sourced from the energy simulation of the 3d models of the data centers using EnergyPlus.

This data has been gathered for different data center modular arrangements. The simulation result of each type generates a resultant csv file. The simulation data has been gathered using preconditioned epw weather file of the given location. The sensor data is gathered at city level. In our case the epw file of Canberra is selected to get sensor data.

Characteristics of the dataset: It is composed of text data, numeric values, not hugely dimensional, and timestamps. There is missing data for heat load and other variables which are explainable since the selected building is a data center which has majorly electrical, lighting and cooling load. It is a typical time series data with 147 observations and 5 variables as explained in the preprocessing section. The simulation data is observed for the hottest day of the year for 24hrs at an interval of 10 minutes. The question from the dataset is - Can we predict the most energy efficient modular arrangement for a Data Center covering a defined area?

C. Data Pre-processing

The data was arranged and processed for the purpose of visualisation and modelling. The resulting csv file had 16 variables. The variables with missing data were removed to achieve 5 variables. They were- time stamp, Sensible cooling load(W), cooling mass flow(kg/s), cooling zone temperature (degree celsius) and cooling zone relative humidity (%).

For the purpose of modelling and visualization, the data was pre-processed in different ways. For visualisation the cleaned data with 5 variables was kept analysing the trend of sensible cooling load and cool mass flow. For the purpose of modelling, only the timestamp and sensible cooling load fields were kept.

D. Time series data visualisation

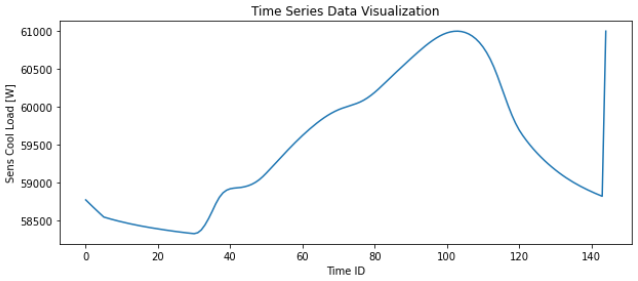


Figure 3: Time Series data visualisation for observed Sensible Cooling Load, source: Author

The time series data has the value for sensible cooling load as it varies at every 10-minute interval throughout a particular summer day in Canberra. The highest load requirement is observed from 4-6pm when the heat in the surrounding regions is maximum.

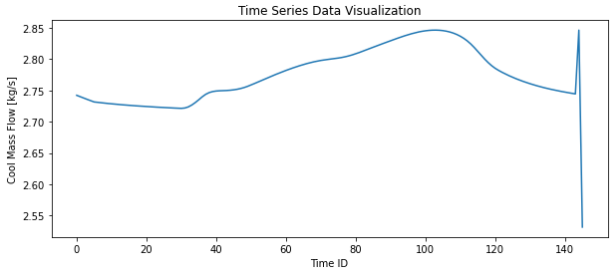


Figure 4:Time Series data visualisation for observed Cool Mass Flow, source: Author

The above graph shows the variation for the Cool Mass Flow throughout a single day. The term "cool mass flow" refers to the rate at which coolant or cooling fluid flows through a system, often measured in kilograms per second (kg/s). In the context of a data center, it is likely related to the cooling system used to dissipate heat generated by IT equipment.

In a data center, cool mass flow may be associated with the flow rate of a cooling fluid (such as chilled water or refrigerant) in the cooling system. The cooling system is responsible for maintaining an optimal temperature within the data center to prevent IT equipment from overheating.

The specific value for cool mass flow in a data center can vary widely depending on the design of the cooling system, the size of the data center, and the heat load generated by the IT equipment. It is typically determined based on the cooling requirements to dissipate the heat produced by servers and other hardware.

E. Prediction of Cooling Load using Machine Learning

Since the time series data was showing peaks and troughs and non-stationarity, it was differenced to create a more stable time series data. The resulting dataset was used to run machine learning algorithms on it. At first, ARIMA model was used which is a typical statistical model for time series. The resulting error was very high and a deep neural network was applied on it. Lastly, LSTM model (Long Short-Term Memory) was applied on the data using TensorFlow libraries. In all the models, we were trying to predict the sensible cooling load of the data center at different time intervals. This type of regression analysis resulted in the following errors as depicted in the next section.

F. Evaluation of Model

The three models which were used to model the all-in-one modular data center's cooling load were: ARIMA, CNN and LSTM. The root mean square error values of each are given below:

TABLE 3: RESULTS FROM PREDICTION OF COOLING LOAD USING DIFFERENT MODELS

Model type	RMSE
ARIMA	13.38
CNN	0.0362
LSTM	0.0082

G. Experimental Results

The resulting modelling shows that LSTM model prediction of sensible cooling load of data centers shows the least error. This implies that such a model can be used to predict the cooling load for a modular data center.
During the energy simulation, it was found that the type of modular configuration of data center that requires the least energy to compensate for the cooling load is an Edge data center module.

V. CONCLUSION AND FUTURE WORK

The research indicates to better prediction of sensible cooling load of Data Centers through the use of LSTM modelling. This would help in energy optimization of the cooling infrastructure of Data Centers. This insight is helpful to both the theory and practice of designing cooling infrastructures of Data Centers. In terms of theory, a policy framework can be devised tabulating the best practices and construction guidelines which can be further researched on. In terms of practice, the ML model can be run for various data sets of various Data Centers.

Furthermore, to build on this work- a tool may be developed with a GUI interface where the modularity configuration a data center can be added and its energy usage and sensible cooling load along with the predicted cooling load can be shown as an output.

Limitations: In the above steps there is a provision of possible variables, but they are based on the simulation data from BIM software which is EnergyPlus. Using different softwares and diffferent weather files may generate different results.

A. Ethical Issues

Data Centers store a massive amount of sensitive and confidential information. Energy optimization research could inadvertently access and expose this data. Researchers need to ensure that data security protocols are in place to protect the privacy of the data.

Energy optimization research aims to reduce the amount of energy consumed by Data Centers. However, some methods of reducing energy consumption, such as the use of renewable energy sources, can have unintended environmental impacts. For example, the production and disposal of solar panels can have negative environmental effects. Researchers need to weigh the environmental impact of their methods against the benefits they provide.

Energy optimization research may inadvertently introduce bias into the systems it aims to optimize. For example, if the

research is conducted on a dataset that is not diverse, the resulting energy optimization system may not work well for all users. Researchers need to be aware of potential biases and work to eliminate them.

Energy optimization research can benefit large Data Centers with significant financial resources to invest in energy-efficient technologies. However, smaller Data Centers and organizations may not have the same resources, putting them at a disadvantage. Researchers need to consider the impact of their methods on all organizations, including those with fewer resources.

Energy optimization research can have significant impacts on the operation and management of Data Centers. Therefore, it is essential to ensure that the research is transparent and that Data Center operators understand how the methods work. This transparency can help to build trust and ensure that the research is used responsibly.

B. Research Quality

In energy optimization research, it is essential to ensure that the methods used to optimize energy consumption are valid and reliable. For example, if the research measures the effectiveness of a specific technology, it is crucial to ensure that the technology is accurately and appropriately evaluated. It is essential to ensure that the findings are generalizable across different data centers and that the methods used can be replicated in other settings.

It is crucial to ensure that the data collected for the study is representative of the data center population being studied. For example, if the study only focuses on large data centers, the findings may not be applicable to smaller data centers. This finding here is limited to smaller modular data centers only. It may not be able to give accurate results for larger data center establishments which consume power and energy at a much larger and expansive scale.

REFERENCES

[1] J. Ni and X. Bai, "A review of air conditioning energy performance in data centers," Renew. Sustain. Energy Rev., vol. 67, pp. 625–640, Jan. 2017.

[2] H. M. Daraghmeh and C.-C. Wang, "A review of current status of free cooling in datacenters," Appl. Therm. Eng., vol. 114, pp. 1224–1239, Mar. 2017.

[3] Kumar, R., Khatri, S.K. and Diván, M.J., 2020, July. "Effect of cooling systems on the energy efficiency of data centers: Machine learning optimisation" In *2020 International Conference on Computational Performance Evaluation (ComPE)* (pp. 596-600). IEEE.

[4] E. Oró, V. Depoorter, A. Garcia, and J. Salom, "Energy efficiency and renewable energy integration in data centres. Strategies and modelling review," Renew. Sustain. Energy Rev., vol. 42, pp. 429–445, Feb. 2015.

[5] J. Gao, "Machine Learning Applications for Data Center Optimization," Google white Pap., pp. 1–13, 2013.

[6] V. S. Marco, Z. Wang, and B. Porter, "Real-Time Power Cycling in Video on Demand Data Centres Using Online Bayesian Prediction," in 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS), 2017, pp. 2125–2130.

[7] G. Smpokos, M. A. Elshatshat, A. Lioumpas, and I. Iliopoulos, "On the Energy Consumption Forecasting of Data Centers Based on Weather Conditions: Remote Sensing and Machine Learning Approach," 2018.

[8] T. P. Lillicrap et al., "Continuous control with deep reinforcement learning," Mach. Learn., Sep. 2015.

[9] W. Yu, Z. Wang, Y. Xue, L. Guo, and L. Xu, "A combined neural and genetic algorithm model for data center temperature control," CEUR Workshop Proc., vol. 2252, pp. 58–69, 2018.

[10] N. Lazic et al., "Data center cooling using model-predictive control," Adv. Neural Inf. Process. Syst., vol. 2018-Decem, no. NeurIPS, pp. 3814–3823, 2018.

[11] Hamilton, J., 2006. "Architecture for modular data centers." arXiv preprint cs/0612110.

[12] Hongyu Zhu et al, "Future data center energy-conservation and emission-reduction technologies in the context of smart and low-carbon city construction", Sustainable Cities and Society, Volume 89, 2023,104322, ISSN 2210-6707 https://doi.org/10.1016/j.scs.2022.104322.

[13] Xiaotong Shao et al, "A review of energy efficiency evaluation metrics for data centers", Energy and Buildings, Volume 271,2022, ISSN 0378-7788, https://doi.org/10.1016/j.enbuild.2022.112308.

Customer Purchase Prediction: A Comparison of Machine Learning and Ensemble Techniques

Name: Tamzid Ibrahim
ID: *u3265713*
University of Canberra
Address: Canberra, Australia

Name: Fariha Nuzhat Majumdar
ID: u3272137
University of Canberra
Address: Canberra, Australia

***Abstract*—With the rapid growth of e-commerce, predicting customer purchase intentions has become crucial for enhancing marketing strategies. This study systematically compares traditional machine learning models, such as Logistic Regression and Support Vector Machines, with advanced ensemble techniques, including LightGBM and CatBoost, to assess their performance in predicting purchase behavior on e-commerce platforms. Using the Online Shoppers Purchasing Intention Dataset from the UCI Machine Learning Repository, we optimized model parameters, employed balancing techniques, and evaluated key metrics across multiple data partitions to achieve robust results. Ensemble models, particularly CatBoost and LightGBM demonstrated superior performance, achieving high average accuracies of 84.15% and 84.0% respectively, with strong recall values in capturing purchase intentions. But LightGBM turned out to be more superior after Hyper Parameter tuning. The study's results underscore the effectiveness of ensemble techniques in handling imbalanced datasets and predicting consumer behavior with accuracy and precision, providing valuable insights for real-time application in personalized marketing. This research highlights the promising potential of machine learning in driving data-informed strategies within the e-commerce industry.**

Keywords— E-commerce purchase prediction, Machine learning in e-commerce, Customer purchase intent prediction, Ensemble learning in marketing, Consumer behavior modeling, Imbalanced data handling, Predictive analytics in e-commerce, Clickstream data analysis

I. INTRODUCTION

There has been a continual rise in mobile phone users worldwide. In 2022, there were 8.58 billion mobile subscribers globally, surpassing the global population of 7.95 billion [1]. By the end of 2023, nearly 70 percent of the world's population were smartphone users, with subscriptions reaching 7 billion, projected to approach 8 billion by 2028 [2]. With the increase in smartphone usage, e-commerce adoption has also surged, a trend that became more pronounced after the COVID-19 pandemic. Lockdowns accelerated the shift to digital sales channels, fundamentally reshaping consumer behaviour and driving the growth of e-commerce platforms [3, 4]. The global e-commerce market is projected to reach US$4.117 trillion in 2024 [5], with the United States and China being the two largest markets, expected to generate US$1.223 trillion and US$1.469 trillion, respectively [6, 7]. Furthermore, global retail e-commerce sales are anticipated to surpass US$8 trillion by 2027 [8].

E-commerce refers to the buying and selling of goods and services online, offering consumers the convenience of shopping from anywhere with an internet connection. By 2025, it is expected that nearly a quarter of all global sales will be conducted online [3].

In this context, purchase intention refers to the likelihood or inclination of a consumer to buy a specific product or service online [9]. Understanding this intent is crucial for businesses, as it provides insights into consumer behaviour, which is not always immediately visible. By analysing purchase intention, companies can enhance customer experiences through targeted advertisements, personalized recommendations, and custom offers, ultimately increasing the likelihood of purchases and driving revenue [10]. To support effective marketing decisions, data-driven methods such as machine learning (ML) are used to extract valuable insights from large datasets, enabling informed marketing strategies [11]. Machine learning has been applied successfully in marketing areas like recommender systems, sentiment analysis, personalization, and optimization [12]. But it's potential to predict online purchase conversion of consumers remains underexplored [13].

Understanding consumer behaviours is essential for successful marketing strategies, and machine learning provides powerful tools to make accurate predictions [14, 15]. By accurately predicting behaviours like purchase intent, businesses can enhance marketing efforts, improving return on investment (ROI) through targeted advertisements [16]. Although The ability to analyse customer journeys through clickstream data offers new opportunities for businesses to improve engagement and sales [17, 18]. However, further research is needed to fully understand which machine learning models work best in this context.

Machine learning (machine learning) techniques have proven effective in predicting customer behaviour by utilizing large datasets, predictive models such as gradient boosting, random forests and deep learning have been applied to customer purchase prediction with significant success [11, 19]. In non-contractual settings, gradient tree boosting has been shown to achieve high accuracy in predicting future purchases [20]. To address the challenge of imbalanced datasets factorization machine models have been effectively used, this also resulted in improved accuracy of the models [21]. It was also shown that deep learning models further enhanced performance, often outperforming traditional machine learning methods in predicting customer behaviour [19]. Additionally, ML techniques have been used for dynamic pricing strategies, enabling real-time prediction of consumer behaviour based on pricing changes [22].

Despite these advancements, there remain challenges in accurately predicting online purchase behaviour. Much of the existing research focuses on post-session purchase prediction, this limits the opportunities for real-time user engagement. Early prediction frameworks have been introduced to address this issue by anticipating purchase intent during active browsing sessions, enabling real-time marketing intervention [23]. However, there are

limitations in some studies when it comes to determining which machine learning techniques offer the best performance. For example, research using Multi-Layer Perceptron (MLP) Artificial Neural Networks (ANN) to predict customer quality in e-commerce social networks has shown potential, but it does not adequately explain which machine learning models are most effective [24]. In a related study, the performance of five machine learning algorithms was compared in the context of text sentiment analysis, yet similar studies focusing on online consumer behaviours remain scarce [25]. But there is a gap which persists in understanding which machine learning models perform best across different e-commerce context, with the need for further exploration [11, 26].

While machine learning models have been widely adopted for predicting customer purchase behavior, there is still a need for a systematic comparison of traditional machine learning algorithms against advanced ensemble techniques specifically within e-commerce applications. This study addresses this gap by investigating the predictive effectiveness of both traditional models (e.g., Logistic Regression, Naïve Bayes, Support Vector Machines) and advanced ensemble methods (e.g., XGBoost, LightGBM, AdaBoost) for forecasting online purchase decisions. Additionally, this research explores data sampling methods to handle class imbalance—a common challenge in purchase prediction that can impact model performance. By evaluating these models across various metrics and sampling strategies, this study aims to enhance prediction accuracy and provide actionable insights.

This study seeks to address the following key research questions:

a. What is the best way to deal with unbalanced data in cases like this?
b. What evaluation metrics should be used to effectively assess the performance of these models in predicting purchase behavior?
c. Which machine learning models "Traditional or Ensemble" offer the highest accuracy in predicting whether a session will lead to a purchase?
d. Which specific models perform the best?

The project utilizes the Online Shoppers Purchasing Intention Dataset from the UCI Machine Learning Repository [27]. The dataset contains 12,330 sessions, capturing various features of user interactions on e-commerce platforms. These features include the number of different types of pages visited, the total time spent on each page category, bounce rates, exit rates, and whether a transaction was completed (which serves as the target variable) [27]. This dataset provides a strong foundation for comparing the performance of predictive models, as it encompasses a wide range of user behaviours that are crucial for predicting online purchase intentions [28].

The increasing shift of consumer activity to online shopping platforms presents both a challenge and an opportunity for businesses aiming to accurately anticipate purchase intentions. With vast amounts of consumer data available, understanding customer purchase behavior has become crucial to enable personalized marketing strategies and improve conversion rates. Yet, predicting purchase behavior is complex due to factors like varying user preferences, session characteristics, and the high imbalance between purchasing and non-purchasing sessions. Although machine learning has been widely used in predicting purchase behavior, prior studies often focus on single techniques or post-session predictions, overlooking the potential of ensemble learning in early-stage prediction scenarios. Our study seeks to address this gap by providing a comprehensive evaluation of traditional machine learning methods alongside advanced ensemble techniques. Additionally, our work explores data sampling methods to effectively manage class imbalance, thus contributing to more accurate and actionable predictions for e-commerce platforms.

II. Literature Review

1) Related Works

Using machine learning to predict consumer behaviour has become a very critical area of research particularly in marketing specifically in the e-commerce field. The diversity of techniques which have been previously explored demonstrates the complexity of understanding consumer behaviour for creating better marketing strategies.

The role of machine learning has been widely acknowledged, and the use of Machine Learning has increased [11], research indicates that supervised and unsupervised learning techniques have become integral for better targeting, personalized recommendations, and predicting purchase behaviour [11]. Both supervised and unsupervised learning techniques have proven effective in processing vast amounts of data, uncovering hidden patterns, and aiding marketing decision-making processes [11]. ML's integration into marketing strategies has resulted in significant improvements in revenue growth and customer retention by enabling businesses to better understand and anticipate consumer needs [11].

One of the critical challenges in predicting purchases by customers was to predict consumer purchase intention in a non-contractual setting, where customer loyalty is not guaranteed [20]. In these specific setting (non-contractual), gradient boosting models have proven to be comparatively better, achieving good accuracy rates (89%), and an AUC of 0.95 on large datasets [20].

When the prices of the products are dynamically adjusted on e-commerce platforms based on real-time data, it is called dynamic pricing strategies, dynamic pricing models have proven to significantly influence purchase decisions [22]. Machine learning frameworks which integrate dynamic pricing strategies have proven to show better model accuracy, highlighting a pivotal role which pricing plays in shaping customer behaviour and optimizing conversion rates [22]. In tandem, machine learning frameworks for early purchase prediction enable real-time marketing interventions during consumer browsing sessions, which provides opportunities for businesses to increase conversions by offering personalized recommendations or discounts [23].

Feature selection and optimization play a vital role in building effective models for predicting customer purchases [29]. In particular, the use of two-stage feature selection and under-sampling techniques have been particularly useful for handling the issue of imbalanced data, this addresses a

frequent issue in large-scale e-commerce datasets [29]. These techniques of feature selection ensures that models are more efficient while processing and analysing substantial volumes of data [29].

Addressing imbalanced data has always been a significant hurdle in purchase prediction, however factorization machine while combined with ensemble learning have demonstrated better accuracy compared to other models [21]. Furthermore, the application of XGBoost has been particularly effective when combined with data-balancing techniques such as SMOTE, which addresses the issue of imbalanced datasets by generating synthetic data points to balance class distribution [13, 16].

Comparative studies of machine learning algorithms have consistently highlighted that ensemble methods like XGboost and AdaBoost, often outperform individual classifiers by combining multiple model strengths [26]. These models reduce the overfitting and improve prediction accuracy [26]. Hybrid models like KnnSgd, blends K-Nearest Neighbours with stochastic gradient descent, have been particularly effective, with studies reporting accuracies of upto 92.42% in predicting customer purchase [26].

Deep learning models demonstrated considerable advantages over traditional machine learning techniques [19]. Neural networks have proved to consistently outperforms traditional machine learning methods like decision trees and support vector machines, specifically when applied to large and complex datasets [19]. This performance is attributed to deep learning's ability to continuously learn from complex data patterns and better capture non-linear relationships between variables [19].

In addition to these methods, neural networks have been employed to predict customer behavior in e-commerce social networks [24]. While multi-layer perceptrons (MLP) have shown promise in this area, there remains ambiguity regarding which machine learning models provide the most effective results for purchase prediction in such social network contexts [24].

Machine learning models that analyze customer preferences in real-time have proven to be valuable in tailoring marketing strategies and improving conversion rates, further emphasizing the need for precise and timely predictions in e-commerce [15].

Machine learning models have also been applied to clickstream data. Click stream data captures the detailed information about customer interactions on websites [30]. By analyzing clickstream data, models can successfully predict anonymous customer purchase intentions [30]. models like logistic regression, random forests, and deep learning perform well in predicting purchase intentions, especially when combined with factors such as product trendiness and popularity [30].

Despite extensive research on using machine learning for predicting customer purchase behaviour, there remains a significant gap in comparing traditional models with advanced ensemble techniques specifically for e-commerce platforms. While neural networks and deep learning methods have demonstrated high accuracy, they require substantial computational resources, making them less practical for businesses with limited resources. This study addresses this gap by focusing on ensemble techniques such as XGBoost, LightGBM, and CatBoost, which offer competitive predictive performance while being more resource efficient.

Analyzing click-stream data allows businesses to understand the browsing patterns, product preferences, and the journey leading to purchase or abandonment [30, 31]. However, effectively leveraging click-stream data for purchase prediction remains a challenge due to the complex, non-linear nature of online customer interactions [30, 31].

By leveraging the Online Shoppers Purchasing Intention Dataset, which is a type of click-stream data capturing detailed user interactions on e-commerce platforms, this research aims to provide a comprehensive comparison of ensemble techniques (dataset er reference).

the following sections detail the methodology, tools, and techniques used to explore the effectiveness of these models in predicting online purchase behaviour.

2) Dataset Used

For this project, we utilized the Online Shoppers Purchasing Intention Dataset, which is publicly available from the UCI Machine Learning Repository [27]. The dataset contains a total of 12,330 sessions, representing various users who browsed a particular e-commerce website over a period of one year. Each session corresponds to a unique user, ensuring that the dataset captures diverse user behaviours across different times of the year. One of the notable features of the dataset is its completeness; there are no missing values, and the dataset has already been pre-processed, requiring minimal feature engineering.

The dataset includes 18 attributes, out of which 10 are numerical, and the remaining 8 are categorical. The numerical features provide detailed information about user interactions with various types of pages, such as Administrative, Informational, and Product-related pages. Similarly, the dataset also includes session-specific attributes like Bounce Rates, Exit Rates, and Page Values, which are all measured using Google Analytics.

Notable attributes include: Administrative, Informational, Product-related: Different categories of pages visited; Administrative Duration, Informational Duration, Product-related Duration: Time spent on these specific types of pages; Bounce Rates and Exit Rates: Behavioral metrics measured by Google Analytics, reflecting the user's interaction intensity; Revenue: The target variable indicating whether a transaction was completed (True) or not (False). Out of 12,330 sessions, only 1,908 sessions ended in a purchase, making it an imbalanced dataset.

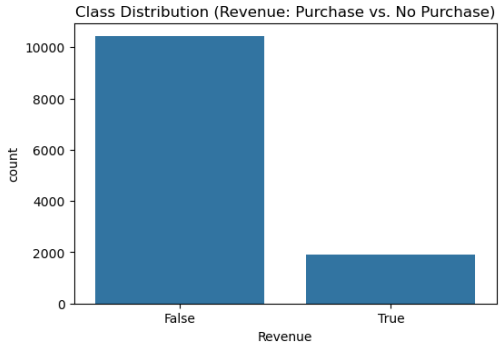


Figure 1: Class Imbalance Analysis

The dataset suffers from class imbalance, as demonstrated in Figure 1, where approximately 85% of the sessions ended without a purchase. This imbalance poses a challenge for predictive modeling, as the models might be biased toward predicting the majority class (no purchase). Dealing with this imbalance effectively is critical for obtaining accurate predictions in purchase behavior models.

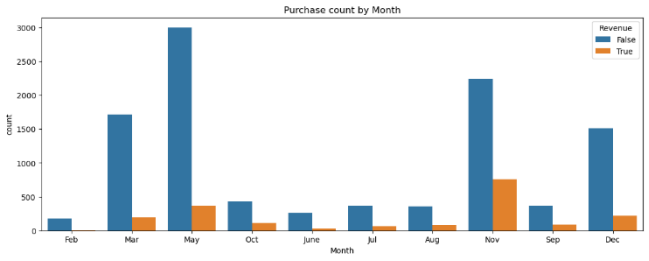


Figure 2: Monthly Purchase Distribution

In Figure 2, the purchase count distribution by month shows a significant variation. Most purchases were made in May and November, likely due to special events such as sales or holidays. This trend emphasizes the impact of seasonality on purchase behaviors, which can provide valuable insights for marketing strategies.

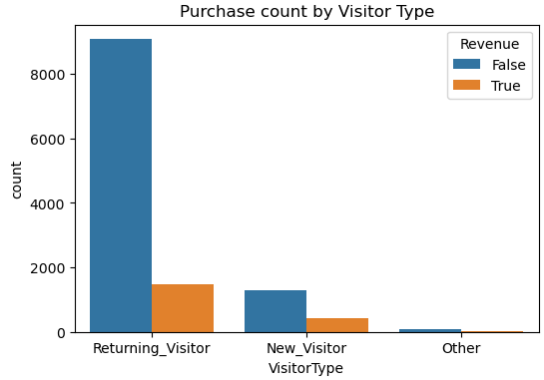


Figure 3: Purchase count by visitors

Most purchases came from Returning Visitors, while New Visitors showed a significantly lower conversion rate. This highlights the importance of customer retention and suggests that returning visitors are more likely to convert into actual buyers.

This dataset offers a rich source of clickstream data, capturing user interactions with various pages and features of the e-commerce platform. The dataset was selected for this project due to its comprehensive nature and suitability for addressing the core objectives of this research. First, the dataset offers a well-balanced combination of numerical and categorical features, providing a detailed picture of user interactions on an e-commerce platform. These features, such as session duration on different page categories, bounce rates, and exit rates, offer valuable insights into user behaviour, which is critical for predicting purchase intentions.

3) ML Classifiers Used

Logistic Regression: This algorithm classifies data by finding a linear decision boundary between classes. While linear models may appear restrictive in low-dimensional spaces, they gain considerable predictive power in high-dimensional data settings, making Logistic Regression highly effective for linearly separable data [32].

Naïve Bayes: Based on Bayes' theorem, Naïve Bayes is a probabilistic classification algorithm. Despite its simplicity, it is widely used and effective, particularly for high-dimensional datasets and cases where independence between features can be assumed [33].

Support Vector Machine (SVM): A supervised learning algorithm that identifies the optimal hyperplane to maximize the margin between classes in classification tasks. SVM is known for its effectiveness in both linear and non-linear classification through the use of kernel functions [34].

Decision Trees: This supervised algorithm builds a model by recursively splitting the data into subsets based on feature values, creating a tree structure of decisions leading to a final classification. It is intuitive and performs well in capturing complex patterns [35].

LightGBM: Designed for both classification and regression tasks, LightGBM employs a leaf-wise tree growth strategy, allowing it to be fast, efficient, and scalable on large datasets. Its unique growth approach makes it highly effective for complex datasets [36].

XGBoost: A tree-based model similar to LightGBM but using a level-wise growth strategy, which results in more balanced trees. While generally slower than LightGBM, XGBoost is robust in handling varied data types and is highly accurate [26].

CatBoost: Developed by Yandex, CatBoost is a gradient boosting algorithm known for its ability to handle categorical features directly. It employs ordered boosting to reduce overfitting and provides unbiased gradient estimation, making it particularly advantageous for imbalanced datasets [37].

AdaBoost: This ensemble technique combines multiple weak learners, adjusting their weights based on performance to form a robust classifier. By focusing on misclassified instances, AdaBoost creates a strong model with improved accuracy [26].

III. Methodology

This research explores how machine learning algorithms can predict purchase intentions in an e-commerce setting by analyzing user session data. The methodology includes data preprocessing, feature selection, model training, evaluation, and optimization, to ensure reliable and robust model performance across a range of data splits and evaluation measures.

Data Preprocessing: After the data was loaded into the program using the pandas library, an exploratory data analysis (EDA) was performed. This involved analyzing the data through descriptive statistics and creating visual representations like bar charts for categorical data and histograms for numerical data. Following this, the dataset was examined for any missing values. The class imbalance was fixed using under sampling. Categorical data was then transformed into numerical values using LabelEncoder, and numerical data was standardized using the StandardScaler for consistency.

Feature Selection: Analyzing high-dimensional data is a difficult job for ML and datamining researchers [28]. Feature selection effectively solves this issue by removing

redundant data, greatly reducing the time needed for processing, improving accuracy, and helps in deeper understanding of the dataset [38]. Shrawan et al. [28] applied five selected features chosen through the wrapper method and a greedy search approach to develop their prediction model. This selection was motivated by the need to **reduce dimensionality** while keeping the most important consumer characteristics for predicting online purchase behavior. The feature selection of this project has been done based on the accuracy of the models. The balanced dataset was tested against all the classifiers (ML models: Logistic Regression, Naïve Bayes, SVM, Decision Tree, and Ensemble models: AdaBoost, LightGBM, XGBoost, CatBoost) and the accuracies of the models have been calculated. Then feature selection was carried out using Recursive Feature Elimination (RFE) with CatBoost as a wrapper model, chosen for its high accuracy across different splits.

Implementation of ML Models and Ensemble Techniques: The experiment was carried out by using multiple classifiers: Logistic Regression, Naïve Bayes, Support Vector Machines (SVM), and Decision Trees, as well as ensemble methods such as AdaBoost, LightGBM, XGBoost, and CatBoost. Each model was trained and evaluated on four data partitions—50-50, 66-34, 80-20, and 10-fold cross-validation—to assess generalization and mitigate biases introduced by a specific data split [39].

Model Evaluation: The performance of all the models was evaluated using accuracy, precision, recall, F1 score, and AUC. These metrics provided a thorough insight into how well the models could correctly classify customer purchase intentions. Accuracy measured the model's overall correctness, precision reflected the model's consistency in making correct predictions, while recall evaluated how well the model identified true positives. The F1 score harmonized the trade-off between precision and recall, and AUC measured the model's effectiveness across all classification thresholds [40].

Optimization and Hyperparameter Tuning: For further refinement, hyperparameter tuning was conducted on the top two performing models- CatBoost and LightGBM using GridSearchCV across different parameters- including learning_rate, tree depth, L2 regularization, and number of boosting iterations and evaluated against the performance metrics- accuracy, precision, recall, f1 score, AUC to determine the best performing model.

Visualization and Final Model Deployment: The top model was visualized using ROC curves, Precision-Recall curves, and confusion matrices, showcasing the model's performance across different splits. Finally, the optimized and refined model was saved for later use, allowing for predictions on new data, continuous development with additional data, and easy integration into practical applications for predicting online purchase intentions.

IV. Reuslts and Analysis

The analysis of the results in this research shows a thorough comparison of several machine learning and ensemble techniques for predicting customer purchase decisions. Shrawan et al. [28] used multiple train-test partitions (50-50, 66-34, 80-20, and 10-fold cross-validation) to evaluate the model performance comprehensively. It allowed the testing of the model's robustness across different data splits and avoid over-reliance on a specific partition.

This research also split the training and testing dataset 4 times and compared the accuracy of all the ML models and ensemble methods in the balanced dataset. In the first split, 50% of the balanced dataset has been used for training and the rest 50% for testing purpose. In the second split, 66% of the balanced dataset has been used for training and the rest 34% has been used for testing. In the third split, 80% of the balanced dataset has been used for training and the rest 20% has been used for testing purpose. Lastly, 10 fold cross-validation has been applied.

Model	50-50 Split Accuracy (%)	66-34 Split Accuracy (%)	80-20 Split Accuracy (%)	10-Fold Cross-Validation Accuracy (%)
Logistic Regression	82.29	82.43	82.85	81.84
SVM	82.81	82.90	83.51	81.42
Naïve Bayes	74.90	74.81	74.74	75.34
Decision Tree	79.87	79.20	80.24	76.21
AdaBoost	84.85	84.44	84.82	80.77
LightGBM	84.75	83.90	84.69	82.65
XGBoost	84.64	83.05	85.08	81.87
CatBoost	**85.12**	**85.36**	**85.60**	**82.44**

Table 1: Accuracies (%) of the models across different splits on full balanced dataset

Table 1 shows the accuracies of all the models across all train-test partitions (50-50, 66-34, 80-20, and 10-fold cross-validation) on the full preprocessed and balanced data.

1) Selecting top 2 models based on performance metrics

After selecting the top 5 features of the dataset, all the models were again trained on the balanced dataset with selected features. To evaluate the model performance comprehensively multiple train-test partitions (50-50, 66-34, 80-20, and 10-fold cross-validation) have been used. The models' performances were analysed using key metrics- accuracy, precision, recall, F1 score, and ROC AUC. Based on the results of the performance metrics, top 2 models have been selected for optimization. Table 3, 4, 5, 6 depict the accuracies, precisions, recalls, and f1 scores respectively of all the models trained with balanced dataset with selected features across all the train-test partitions (50-50, 66-34, 80-20, and 10-fold cross-validation).

Model	50-50 Split Accuracy (%)	66-34 Split Accuracy (%)	80-20 Split Accuracy (%)	10-Fold Cross-Validation Accuracy (%)
Logistic Regression	82.34	82.36	82.59	81.87
SVM	81.92	82.20	83.51	81.29
Naïve Bayes	78.51	78.27	79.45	77.91
Decision Tree	79.51	78.51	80.37	77.46
AdaBoost	84.33	84.75	85.73	81.18
LightGBM	83.65	84.05	84.82	80.92
XGBoost	83.12	83.20	84.69	79.98
CatBoost	85.43	84.90	84.82	81.47

Table 2: Accuracies (%) of the models across different splits on balanced dataset with selected features

A thorough examination of Table 2 shows that based on accuracy, the top 2 models are CatBoost and LightGBM with average accuracy of 84.15% and 84% respectively.

Model	50-50 Split Precision (%)	66-34 Split Precision (%)	80-20 Split Precision (%)	10-Fold Cross-Validation Precision (%)
Logistic Regression	87.79	87.66	88.16	87.25
SVM	87.95	88.32	89.62	87.55
Naïve Bayes	85.36	85.46	87.21	84.42
Decision Tree	80.28	78.72	79.53	78.89
AdaBoost	87.37	87.41	87.89	86.64
LightGBM	83.33	83.51	83.59	82.96
XGBoost	83.37	82.41	82.71	81.71
CatBoost	86.37	85.08	83.76	83.87

Table 3: Precisions (%) of the models across different splits on balanced dataset with selected features

Based on the precision scores which are presented in Table 3, the top 2 best performing models are SVM and Logistic Regression with average precision of 88.36% and 87.72% respectively.

Model	50-50 Split Recall (%)	66-34 Split Recall (%)	80-20 Split Recall (%)	10-Fold Cross-Validation Recall (%)
Logistic Regression	75.21	74.61	74.87	74.63
SVM	74.06	73.51	75.40	72.96
Naïve Bayes	68.93	67.24	68.52	68.45
Decision Tree	78.35	77.12	81.22	75.00
AdaBoost	80.33	80.56	82.54	73.74
LightGBM	84.21	84.17	86.24	77.83
XGBoost	82.85	83.70	87.30	77.25
CatBoost	84.21	84.01	85.98	77.94

Table 4: Recall scores (%) of the models across different splits on balanced dataset with selected features

According to the recall scores shown in Table 5, the top 2 best-performing models are LightGBM and CatBoost, with average recall values of 83.11% and 83.03%, respectively.

The 2 highest performing models are CatBoost and LightGBM as indicated by the f1-scores in Table 5 with average f1-scores of 83.87% and 83.20%, respectively.
The Figure.1 below shows the ROC curves with (AUC scores) of all the models across all train-test splits on the balanced dataset with selected features. The **ROC curve (Receiver Operating Characteristic Curve)** is a graphical tool for evaluating the performance of a binary classification model at various threshold values.

Model	50-50 Split F1-score (%)	66-34 Split F1-score (%)	80-20 Split F1-score (%)	10-Fold Cross-Validation F1-score (%)
Logistic Regression	81.01	80.61	80.97	80.45
SVM	80.41	80.24	81.90	79.59
Naïve Bayes	76.27	75.26	76.74	75.60
Decision Tree	79.30	77.91	80.37	76.89
AdaBoost	83.71	83.85	85.13	79.67
LightGBM	83.77	83.84	84.90	80.31
XGBoost	83.11	83.05	84.94	79.42
CatBoost	85.28	84.54	84.86	80.79

Table 5: F1-scores (%) of the models across different splits on balanced dataset with selected features

It helps in visualizing the trade-off between the **True Positive Rate (TPR)** and the **False Positive Rate (FPR)** at different categorization thresholds [39]. In the ROC curve the x-axis represents the False Positive Rate, and the y-axis represents the True Positive Rate. The **AUC (Area Under the ROC Curve)** represents a single numerical value that evaluates the overall performance of a binary classifier. It calculates the area under the ROC curve and provides an overview of the classifier's ability to differentiate between the two classes [39]. An AUC score of 1.0 represents perfect classifier that correctly differentiates between all positive and negative cases.

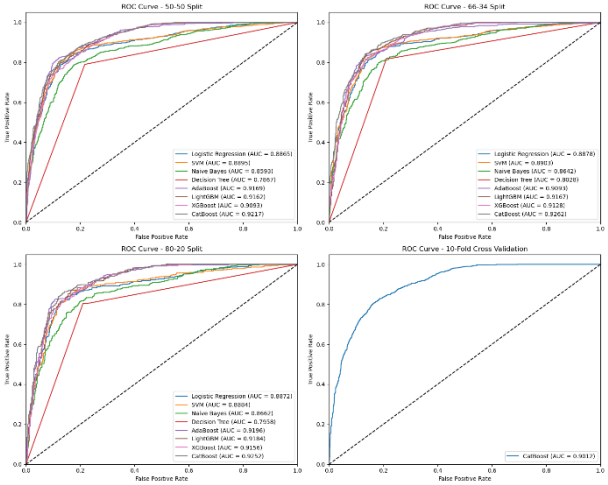


Figure 4: ROC Curves Across all 4 partitions

From the ROC curves in all splits shown in Figure 4, the best two performing models are CatBoost and AdaBoost.
A detailed analysis on the performance metrics tables-accuracy, precision, recall, f1-score, and ROC AUC scores show that, CatBoost has consistent high performance across accuracy, recall, F1, and ROC AUC, and LightGBM shows highest recall score which ensures capturing potential buyers and higher scores in other performance metrics. Thus, after evaluation. The top 2 best performing models are CatBoost and LightGBM which are selected for further analysis for improving results.

2) Top 2 models based on cross validation & overfitting

To generalize the models and to prevent bias and overfitting, all the models have been cross validated and the test results have been analyzed. **CatBoost** had the highest cross-validation accuracy (85.4%), followed by **LightGBM** (84.8%). Other models demonstrated varying performance with Support Vector Machines (SVM) and Logistic Regression reaching scores of 0.820 and 0.823, respectively, and Naïve Bayes scoring the lowest at 0.757.
Test accuracies were evaluated across 50-50, 66-34, and 80-20 splits. **CatBoost** consistently performed well across all splits, with accuracies around 84.6-84.7. **LightGBM** also achieved strong test results, with its highest test accuracy being 84.9 in the 80-20 split.

Most models showed slight overfitting, with the accuracy achieved through cross-validation being just a bit greater than that on the test set. LightGBM and CatBoost were identified as the top-performing models, showing superior performance in both cross-validation and test results with consistently high accuracy across different data splits and minimal overfitting. Figures and Tables.

Model	50-50 Split	66-34 Split	80-20 Split
Logistic Regression	0.009	0.010	0.010
SVM	0.009	0.008	0.002
Naïve Bayes	0.010	0.022	0.008
Decision Tree	~0	~0	~0
AdaBoost	0.010	0.017	0.005
LightGBM	0.010	0.013	~0
XGBoost	0.006	0.001	0.007
CatBoost	0.008	0.008	0.007

Table 6: Overfitting Analysis (Difference Between CV and Test Accuracies)

3) Improving results of the top 2 models by hyperparameter tuning and determining the best model

Standard settings for hyperparameters cannot ensure the best performance of machine learning methods [41]. To maximize the effectiveness and generalizability of machine learning (ML) models, hyperparameter adjustment is crucial [42]. Tuning is particularly important for gradient boosting models such as CatBoost and LightGBM, since their performance is significantly affected by factors such as learning rate, tree depth, and number of estimators. The most used method for hyperparameter tuning is GridSearchCV.

Hyperparameter tuning and model evaluation have been performed on CatBoost and LightGBM models across different train-test splits and a 10-fold cross-validation. The parameters chosen for the tuning are: L2 regularization (chosen values: 3, 5), depth of tree (chosen values: 4, 6), learning rate (chosen values: 0.1, 0.05), number of boosting iterations (chosen values: 100, 300). Moderate levels of regularization have been chosen because low level of regularization could allow overfitting and high level of regularization could lead to underfitting [43]. Moderate tree depths ranging from 4 to 6 enable the model to identify significant interactions among features without getting overly complex [43]. By using shallow trees, the model generalizes better, and reducing the maximum depth also speeds up training [44]. A smaller learning rate of 0.05 can lead to more conservative updates, whereas 0.1 is faster and can achieve good generalization [45]. By selecting learning rates that are not too big, the model makes updates more carefully, which aids in preventing overfitting [46]. Training time and model accuracy are balanced by the number of trees. Accuracy increases with more trees, but overfitting and training time also increases. By testing both 100 and 300, the modeler can find a balance between performance and efficiency [43].

GridSearchCV has been applied to find the best hyperparameters. The dataset X and y are split into training and testing sets based on the split ratio (50-50, 66-34, and 80-20). The training set is further split into training and validation sets which is later used to predict unseen data by the best model found by tuning the hyper parameters. Evaluation of the models based on key metrics such as accuracy, precision, recall, F1 score, and ROC AUC has been performed and the model performances were compared across different train-test splits and cross-validation to choose the best-performing model configuration.

The analysis shows that LightGBM and CatBoost performed equally in terms of accuracy in the 50-50 split (84.3%). LightGBM has slightly higher accuracy in the 80-20 split (86.6% vs 86%) and in the 10-fold cross-validation (87.2% vs 86.3%). In terms of precision, CatBoost has a higher precision in the 50-50 split and slightly higher in 10-fold CV, whereas LightGBM has better precision in the 66-34, 80-20

Evaluation	CatBoost Test Accuracy (%)	CatBoost Precision (%)	CatBoost Recall (%)	CatBoost F1-score (%)	CatBoost AUC (%)	LightGBM Test Accuracy (%)	LightGBM Precision (%)	LightGBM Recall (%)	LightGBM F1-score (%)	LightGBM AUC (%)
50-50 Split	84.3	85.4	82.7	84.0	91.8	84.3	82.5	86.4	84.4	84.3
66-34 Split	83.4	83.8	82.9	83.3	91.5	86.0	84.0	88.0	85.9	86.0
80-20 Split	86.0	85.9	86.1	86.0	91.7	86.6	86.9	85.5	86.2	86.6
10-Fold CV	86.3	87.3	85.0	86.1	94.0	87.2	87.2	87.2	87.2	95.0

Table 7: Comparison of CatBoost and LightGBM Performance (%) After Hyperparameter Tuning Across Different Train-Test Splits and 10-Fold Cross-Validation

splits. LightGBM performed much better in recall scores in the 50-50, 66-34 splits and 10 fold cv, where it achieved higher recall scores compared to CatBoost. Across all splits both models had very similar F1-score, but LightGBM again slightly outperforms CatBoost in all cases.

In terms of ROC AUC, CatBoost scored better in the single-split evaluations (50-50, 66-34, and 80-20 splits) where it surpassed LightGBM. Nonetheless, in the 10-fold cross-validation, LightGBM secured the top ROC AUC score (95.0%), surpassing CatBoost's score of 94.0%. After hyperparameter tuning both models showed remarkable performance. But in most splits, LightGBM scored marginally better than CatBoost in terms of test accuracy and recall, but CatBoost performed better in terms of ROC AUC for individual splits. Overall, LightGBM performed best after hyperparameter tuning especially in 10-fold cross-validation, when it achieved the highest ROC AUC. So, the best model after hyperparameter tuning is **LightGBM**.

4) Visualising test cases agianst training cases

Model performance of LightGBM has been visualized by plotting confusion matrix, ROC AUC curves, and Precision-Recall curves of the test cases against training cases.

A confusion matrix is a tool for assessing the performance of a classification model in machine learning and statistics to summarize how well a classification model performs. The matrix provides insight into the number of correct and incorrect predictions, categorizing them based on actual and predicted classes [47]. In a binary classification problem, the confusion matrix is a 2x2 table where [47, 48] **True Positive (TP)** represents the number of correctly predicted positive cases, **True Negative (TN)** is the number of correctly predicted negative cases, **False Positive (FP)** represents the number of negative cases incorrectly predicted as positive, and lastly, **False Negative (FN)** represents the number of positive cases incorrectly predicted as negative.

From the confusion matrix we can see the counts of true positives, true negatives, false positives, and false negatives. For better visualization we showed this as a Heatmap. Figure 5 displays the confusion matrices of training and testing sets across 50-50, 66-34, 80-20 splits and 10-fold CV.

The analysis of the confusion metrices in Figure: 5 show that the optimized model LightGBM demonstrates strong and robust performance throughout various train-test splits, especially in **10-fold cross-validation**. Overall, in the 50-50 split, the model performs reasonably well but shows some

generalization issues with higher false positives and false negatives on the test set. The split 66-34 exhibits better generalization, as it has lower false positive and false negative rates on the test set, which suggests improved performance when more data is used for training. In the 80-20 Split, the model performs quite well with fewer false positives and negatives on the test set, which indicates that the model benefits from more training data and generalizes well on the remaining test data. The confusion matrix of LightGBM in 10-fold cross validation shows excellent generalization with very low false positives and false negatives, indicating that the model can generalize well to unseen data.

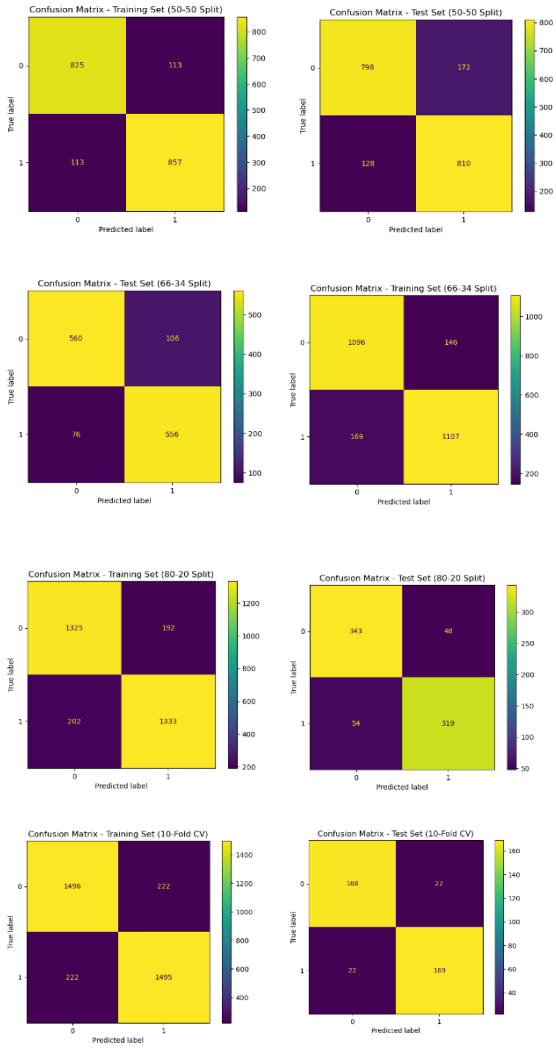


Figure 5: Confusion Matrix

The ROC curves of LightGBM for all splits (50-50, 66-34, 80-20, and 10 fold cv) are shown in Figure 6. Overall, The ROC curves for every split demonstrate exceptional results, with AUC scores constantly exceeding 0.90 for both the training and evaluation datasets. 10-fold cross-validation yields the most accurate generalization, achieving the highest test AUC (0.96), indicating a robust approach for evaluating the model.

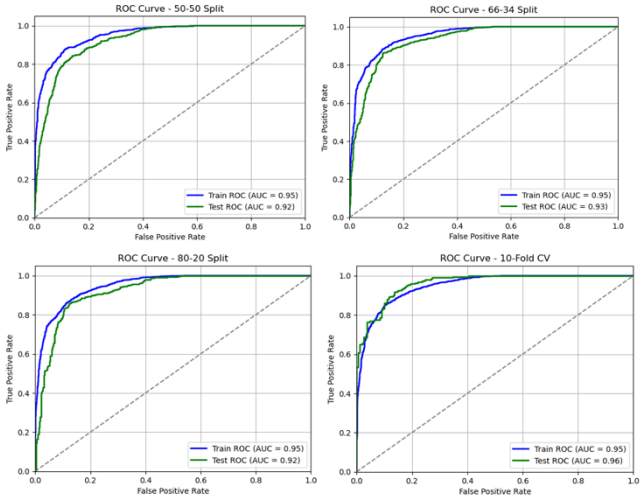


Figure 6: ROC curves of LightGBM for training and testing sets across all splits

The Precision-Recall (PR) curves of LightGBM model for training and testing sets in all splits are shown in Figure.7.

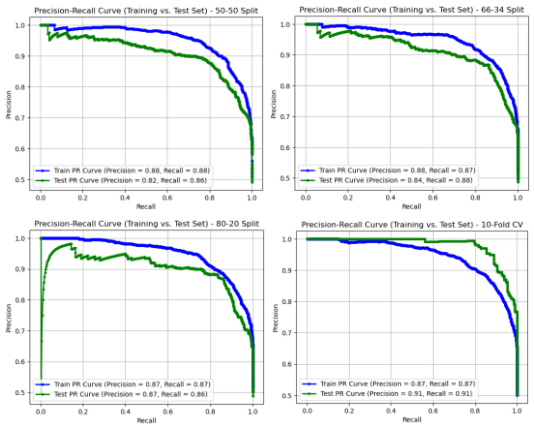


Figure 7: Precision-Recall Curves of LightGBM for training and testing sets across all splits

Precision-Recall curves are especially useful when models need to be evaluated on datasets with imbalanced classes, as they emphasize performance balance between precision and recall [45]. The PR curves in 50-50 split demonstrate a significant drop in precision on the test set when compared to the training test indicating a minor case of overfitting. The test curve is closer to the training curve in 66-34, and 80-20 splits, which suggests better generalization. Precision and recall are relatively consistent, indicating a more even performance when faced with unseen data. In the 10-fold cross validation, the precision and recall values are almost the same demonstrating best generalization.

5) Predictions on sample unseen cases using the optimised model

The optimized model is further evaluated on a validation set to ensure that the model is not overfitting to the test set. In 50-50, 66-34, and 80-20 split, the validation set is 10%, 6.8%, and 4% of the original data. Figure. 8 shows ROC curves for

the optimized LightGBM classifier evaluated on test sets and validation sets across different train-test splits (50-50, 66-34, and 80-20 splits). And the classification reports for the model's test and validation sets are shown in Table. 9.

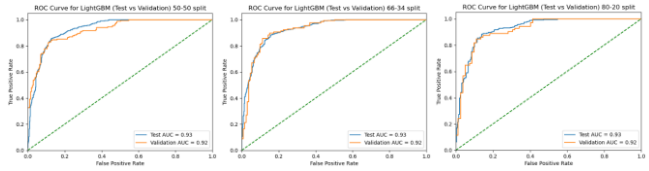


Figure 8: ROC curves for optimized LightGBM classifier evaluated on test and validation sets across different train-test splits (50-50, 66-34, and 80-20 splits)

In every split, the AUC scores for the test and validation are nearly identical (Test AUC $\approx$ 0.93, Validation AUC $\approx$ 0.92), which suggests that the model generalizes well to unseen data. It also proves that the model is not overfitting. The small difference in the AUCs between the test and validation sets (0.01 difference) indicates that the validation data serves as a reliable predictor of the model's performance on unseen data.

Split	Set	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
50-50	Test	85.98	85.68	86.13	85.90	92.78
	Validation	84.56	83.15	84.53	83.84	91.81
66-34	Test	85.84	84.82	86.34	85.57	92.74
	Validation	86.15	83.21	89.76	86.36	92.46
80-20	Test	86.42	86.38	86.09	86.24	92.71
	Validation	86.27	83.78	87.32	85.52	91.96

Table 8: Performance Metrics Comparison of Test and Validation sets for optimized LightGBM model Across Train-Test Splits (50-50, 66-34, 80-20)

From Table. 9, the test set shows slightly better accuracy, precision, recall, and F1 score than the validation set in 50-50 split. The AUC is higher for the test set as well. The validation set performs better in terms of accuracy, recall, and F1 score in the 66-34 split. The AUC score is also similar in both test and validation set. This might be more suited for applications where high recall is necessary. In the 80-20 split, the test and validation sets show very similar results, with the test set having slightly better precision (86.38%) and the validation set having better recall (87.32%).

V. Ethical and Privacy Considerations:

This study prioritizes ethical and privacy considerations, particularly as it involves consumer data that inherently contains sensitive personal and behavioral information. The dataset used is sourced from a publicly available repository, which helps mitigate direct privacy concerns related to data ownership and informed consent. However, issues of data privacy and ethical responsibility remain crucial.

There is a potential risk of algorithmic bias. The class imbalance in the dataset—where non-purchase sessions vastly outnumber purchase sessions—can lead to bias in model predictions, inadvertently impacting certain user groups if deployed without adjustments. Addressing this bias through proper evaluation and ethical algorithm design is essential to prevent skewed results that could misinform business strategies or unfairly target certain consumer profiles.

VI. Limitation and Future Work

Limitations: While this study provides a robust analysis of various machine learning and ensemble techniques for predicting customer purchase intentions, it does have some limitations. Firstly, the dataset used was highly imbalanced, and while balancing techniques were applied, we did not implement synthetic oversampling methods such as SMOTE, which may have provided additional insight into handling minority classes more effectively. Consequently, the performance of the models, especially in capturing positive purchase instances, might be limited by the inherent data distribution. Another limitation is the dependency on specific model parameters and the selected features, which, although optimized, may not generalize as well in different e-commerce settings or datasets with more complex user behaviors.

Additionally, the models used, such as CatBoost and LightGBM, are computationally intensive, especially with large datasets, and may not be feasible for real-time applications without significant computational resources. This computational requirement can limit the scalability of the solution for online, real-time consumer behavior prediction across multiple e-commerce platforms. Finally, the study's reliance on clickstream data limits the scope to only online purchase intentions, and the results may not fully apply to offline retail contexts or other non-clickstream data sources.

Future Work: To address the identified limitations, future research could explore the use of advanced oversampling techniques, such as SMOTE or adaptive synthetic sampling, to improve model performance on imbalanced datasets further. Expanding the model scope to include more diverse datasets from different e-commerce environments would allow a more comprehensive evaluation of the models' generalizability.

Furthermore, integrating additional features, such as demographic information, seasonal trends, or customer sentiment analysis from textual reviews, could enhance the predictive accuracy and relevance of the models. Real-time implementation of these models could be explored using streaming data technologies and lightweight model architectures, such as gradient-based models with feature engineering optimizations, to support efficient and scalable deployment in production environments.

Lastly, a comparative study of deep learning models alongside the traditional and ensemble techniques evaluated here would provide insights into whether neural networks could yield further improvements in accuracy, particularly for datasets with complex behavioral patterns or higher volumes of real-time data.

VII. Conclusion

In this study, data imbalance was addressed through balancing techniques such as resampling to ensure the model

could adequately capture the minority class (purchase events). For e-commerce applications, methods like SMOTE (Synthetic Minority Over-sampling Technique) are suggested to further improve results by generating synthetic examples for the minority class, an area intended for future exploration.

Accuracy alone proved insufficient for evaluating performance on imbalanced datasets. Consequently, a comprehensive range of metrics, including Precision, Recall, F1-score, and ROC AUC, was utilized to assess each model. Recall and F1-score were particularly essential, as they offered insight into how well the models captured actual purchase events without excessive bias toward the majority class.

A comparison between traditional and ensemble models indicated that ensemble techniques, specifically LightGBM and CatBoost, provided superior accuracy and recall over traditional machine learning models like Logistic Regression and Naïve Bayes. The ensemble techniques excelled in capturing complex, non-linear patterns within the data, demonstrating robustness in handling the intricacies of purchase behavior prediction.

LightGBM and CatBoost emerged as the top-performing models, achieving the highest accuracy and recall scores. These models consistently outperformed others across multiple metrics and data partitions, demonstrating their suitability for purchase intent prediction in e-commerce contexts.

Following hyperparameter tuning, both models demonstrated strong performance. LightGBM achieved slightly higher test accuracy and recall across most data splits, while CatBoost excelled in ROC AUC for individual splits. Overall, LightGBM emerged as the top-performing model after tuning, especially in 10-fold cross-validation, where it reached the highest ROC AUC score. Thus, LightGBM was identified as the best model post-tuning

In conclusion, ensemble methods, particularly LightGBM offer substantial advantages for predicting customer purchase behavior in imbalanced datasets typical of e-commerce. While balancing techniques and a range of evaluation metrics contributed to model robustness, future research could extend into advanced sampling methods, integrating additional data features, and implementing real-time applications. This study underscores the importance of selecting appropriate models and evaluation strategies, paving the way for more effective, data-driven e-commerce strategies.

VIII. References:

[1] F. Richter. "There are more mobile phones than people in the world " World Economic Forum https://www.weforum.org/agenda/2023/04/charted-there-are-more-phones-than-people-in-the-world/#:~:text=According%20to%20the%20International%20Telecommunication,billion%20halfway%20through%20the%20year. (accessed 13 October, 2024).

[2] Statista. "Smartphones - statistics & facts." Statista. https://www.statista.com/topics/840/smartphones/#topicOverview (accessed 13 October, 2024).

[3] McKinsey&Company. "What is e-commerce?" McKinsey & Company https://www.mckinsey.com/business-functions/growth-marketing-and-sales/our-insights/what-is-e-commerce (accessed 9th, October, 2024).

[4] P. Ahluwalia and M. I. Merhi, "Understanding country level adoption of e-commerce: A theoretical model including technological, institutional, and cultural factors," *Journal of Global Information Management (JGIM)*, vol. 28, no. 1, pp. 1-22, 2020.

[5] Statista. "eCommerce - Worldwide." Statista. https://www.statista.com/outlook/emo/ecommerce/worldwide (accessed 26 August, 2024, 2024).

[6] Statista. "eCommerce - United States." Statista. https://www.statista.com/outlook/emo/ecommerce/united-states (accessed August 26, 2024, 2024).

[7] Statista. "eCommerce - China." Statista. https://www.statista.com/outlook/emo/ecommerce/china (accessed August 26, 2024, 2024).

[8] Statista. "Retail e-commerce sales worldwide from 2014 to 2027 (in billion U.S. dollars)." Statista https://www.statista.com/statistics/379046/worldwide-retail-e-commerce-sales/ (accessed 13 October, 2024).

[9] M.-Y. Chen and C.-I. Teng, "A comprehensive model of the effects of online store image on purchase intention in an e-commerce environment," *Electronic Commerce Research*, vol. 13, pp. 1-23, 2013.

[10] Y. K. Dwivedi *et al.*, "Setting the future of digital and social media marketing research: Perspectives and research propositions," *International journal of information management*, vol. 59, p. 102168, 2021.

[11] E. W. Ngai and Y. Wu, "Machine learning in marketing: A literature review, conceptual framework, and research agenda," *Journal of Business Research*, vol. 145, pp. 35-48, 2022.

[12] R. E. Bawack, S. F. Wamba, K. D. A. Carillo, and S. Akter, "Artificial intelligence in E-Commerce: a bibliometric study and literature review," *Electronic markets*, vol. 32, no. 1, pp. 297-338, 2022.

[13] J. Lee, O. Jung, Y. Lee, O. Kim, and C. Park, "A comparison and interpretation of machine learning algorithm for the prediction of online purchase conversion," *Journal of Theoretical and Applied Electronic Commerce Research*, vol. 16, no. 5, pp. 1472-1491, 2021.

[14] D. Cui and D. Curry, "Prediction in marketing using the support vector machine," *Marketing Science*, vol. 24, no. 4, pp. 595-615, 2005.

[15] A. Politz and W. E. Deming, "On the necessity to present consumer preferences as predictions," *Journal of Marketing*, vol. 18, no. 1, pp. 1-5, 1953.

[16] A. Lambrecht and C. Tucker, "When does retargeting work? Information specificity in online advertising," *Journal of Marketing research*, vol. 50, no. 5, pp. 561-576, 2013.

[17] B. Libai *et al.*, "Brave new world? On AI and the management of customer relationships," *Journal of Interactive Marketing*, vol. 51, no. 1, pp. 44-56, 2020.

[18] S. Gupta, A. Leszkiewicz, V. Kumar, T. Bijmolt, and D. Potapov, "Digital analytics: Modeling for insights and new methods," *Journal of Interactive Marketing*, vol. 51, no. 1, pp. 26-43, 2020.

[19] N. Chaudhuri, G. Gupta, V. Vamsi, and I. Bose, "On the platform but will they buy? Predicting customers' purchase behavior using deep learning," *Decision Support Systems*, vol. 149, p. 113622, 2021.

[20] A. Martínez, C. Schmuck, S. Pereverzyev Jr, C. Pirker, and M. Haltmeier, "A machine learning framework for customer purchase prediction in the non-contractual setting," *European Journal of Operational Research*, vol. 281, no. 3, pp. 588-596, 2020.

[21] S.-x. Chen, X.-k. Wang, H.-y. Zhang, and J.-q. Wang, "Customer purchase prediction from the perspective of imbalanced data: A machine learning framework based on factorization machine," *Expert Systems with Applications*, vol. 173, p. 114756, 2021.

[22] R. Gupta and C. Pathak, "A machine learning framework for predicting purchase by online customers based on dynamic pricing," *Procedia Computer Science*, vol. 36, pp. 599-605, 2014.

[23] R. Esmeli, M. Bader-El-Den, and H. Abdullahi, "Towards early purchase intention prediction in online session based retailing systems," *Electronic Markets*, vol. 31, no. 3, pp. 697-715, 2021.

[24] M. T. Ballestar, P. Grau-Carles, and J. Sainz, "Predicting customer quality in e-commerce social networks: a machine learning approach," *Review of Managerial Science*, vol. 13, pp. 589-603, 2019.

[25] J. Hartmann, J. Huppertz, C. Schamp, and M. Heitmann, "Comparing automated text classification methods," *International Journal of Research in Marketing*, vol. 36, no. 1, pp. 20-38, 2019.

[26] G. Chaubey, P. R. Gavhane, D. Bisen, and S. K. Arjaria, "Customer purchasing behavior prediction using machine learning classification techniques," *Journal of Ambient Intelligence and Humanized Computing*, vol. 14, no. 12, pp. 16133-16157, 2023.

[27] O. Sakar, S. Polat, M. Katircioglu, and Y. Kastro, "Online shoppers purchasing intention dataset," *UCI machine learning Repository*, vol. 10, p. C5F88Q, 2018.

[28] S. K. Trivedi, P. Patra, P. R. Srivastava, J. Z. Zhang, and L. J. Zheng, "What prompts consumers to purchase online? A machine learning approach," *Electronic Commerce Research*, pp. 1-37, 2022.

[29] D. Theng and K. K. Bhoyar, "Feature selection techniques for machine learning: a survey of more than two decades of research," *Knowledge and Information Systems*, vol. 66, no. 3, pp. 1575-1637, 2024.

[30] Z. Wen, W. Lin, and H. Liu, "Machine-learning-based approach for anonymous online customer purchase intentions using clickstream data," *Systems*, vol. 11, no. 5, p. 255, 2023.

[31] S.-C. Necula, "Exploring the impact of time spent reading product information on e-commerce websites: A machine learning approach to analyze consumer behavior," *Behavioral Sciences*, vol. 13, no. 6, p. 439, 2023.

[32] G. Chaubey, D. Bisen, S. Arjaria, and V. Yadav, "Thyroid disease prediction using machine learning approaches," *National Academy Science Letters*, vol. 44, no. 3, pp. 233-238, 2021.

[33] M. J. Sánchez-Franco, A. Navarro-García, and F. J. Rondán-Cataluña, "A naive Bayes strategy for classifying customer satisfaction: A study based on online reviews of hospitality services," *Journal of Business Research*, vol. 101, pp. 499-506, 2019.

[34] J. Nalepa and M. Kawulok, "Selecting training sets for support vector machines: a review," *Artificial Intelligence Review*, vol. 52, no. 2, pp. 857-900, 2019.

[35] L. Rokach and O. Maimon, "Decision trees," *Data mining and knowledge discovery handbook*, pp. 165-192, 2005.

[36] A. Alsubayhin, M. S. Ramzan, and B. Alzahrani, "Crime Prediction Model using Three Classification Techniques: Random Forest, Logistic Regression, and LightGBM," *International Journal of Advanced Computer Science & Applications*, vol. 15, no. 1, 2024.

[37] X. Dou, "Online purchase behavior prediction and analysis using ensemble learning," in *2020 IEEE 5th International conference on cloud computing and big data analytics (ICCCBDA)*, 2020: IEEE, pp. 532-536.

[38] J. Cai, J. Luo, S. Wang, and S. Yang, "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp. 70-79, 2018.

[39] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," *Morgan Kaufman Publishing*, 1995.

[40] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *arXiv preprint arXiv:2010.16061*, 2020.

[41] P. Schratz, J. Muenchow, E. Iturritxa, J. Richter, and A. Brenning, "Hyperparameter tuning and performance assessment of statistical and machine-learning algorithms using spatial data," *Ecological Modelling*, vol. 406, pp. 109-120, 2019.

[42] J. A Ilemobayo *et al.*, "Hyperparameter Tuning in Machine Learning: A Comprehensive Review," *Journal of Engineering Research and Reports*, vol. 26, no. 6, pp. 388-395, 2024.

[43] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009.

[44] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*. Springer, 2013.

[45] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189-1232, 2001.

[46] W. I. D. Mining, "Data mining: Concepts and techniques," *Morgan Kaufinann*, vol. 10, no. 559-569, p. 4, 2006.

[47] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning* (no. 4). Springer, 2006.

[48] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 233-240.

Student Performance Prediction in Australian Higher Education

Dehai Lu
University Of Canberra
Canberra, Australia
LUDEHAI1999@gmail.com

***Abstract*— This study aims to predict academic performance among university students in Australia using machine learning models, specifically Random Forest (RF) and Deep Neural Networks (DNN). With a focus on student outcomes such as Excellent, Good, Satisfactory, Needs Improvement, and Poor, the research addresses the increasing concern over student dropout rates and subpar grades. Building on prior research, our models leverage a dataset of over 100,000 records containing academic, personal, and socio-economic variables. To improve the reliability of predictions, we address class imbalance by employing Random Over-Sampling (ROS) and Synthetic Minority Oversampling Technique (SMOTE). The results will offer insights into key factors influencing student success and provide a basis for early intervention strategies aimed at improving educational outcomes.**

Keywords*— *Student Performance Prediction, Random Forest, Deep Neural Networks, Imbalance Data Problem, Machine learning, Educational Data

I. INTRODUCTION

In recent years, the academic performance of university students in Australia has raised significant concerns, particularly with nearly 20% of students dropping out in 2016 [1]. This alarming trend, coupled with average grades consistently falling short of expectations, underscores the urgent need for predictive interventions to identify at-risk students and improve their academic outcomes [2].

Previous research has demonstrated the effectiveness of various machine learning models in predicting student performance and dropout rates across different contexts. For instance, Sergi Rovira successfully applied models such as Logistic Regression, Naive Bayes, Support Vector Machine (SVM), Random Forest, and Adaptive Boosting to predict student dropout in Barcelona, achieving an impressive 80% accuracy [2]. Similarly, Agathe Merceron utilized decision tree models to forecast Jamaican university student performance with 81% accuracy [3], while Beaulac's implementation of a Random Forest model reached a remarkable 91% accuracy in Canada [4]. Additionally, Aya Nabila's study highlighted the effectiveness of Deep Neural Networks (DNN), achieving a 91% accuracy rate in predicting outcomes for public university students [5].

Building on these successful applications, our study aims to employ both Random Forest (RF) [6]and Deep Neural Network (DNN) [7] models to predict academic performance among university students in Australia. Recognizing the importance of addressing class imbalance, particularly in the "Excellent" performance category, we will incorporate Random Over-Sampling (ROS) and the Synthetic Minority Oversampling Technique (SMOTE) [8] into our methodology.

These techniques will enhance the reliability of our predictive models by ensuring a more balanced representation of all performance categories. By leveraging these advanced modelling techniques, we aspire to provide valuable insights and contribute to effective interventions that support student success in higher education.

II. METHODOLOGY

A. Research Design

This study follows a quantitative research approach, as the relevant theories on student performance are established prior to the investigation. The aim is to empirically test these theories by formulating hypotheses about the effects of various personal, academic, socio-economic, and lifestyle factors on student performance. Through comprehensive data collection and statistical analysis, this design will help identify the relationships between these variables and the academic outcomes of university students in Australia.

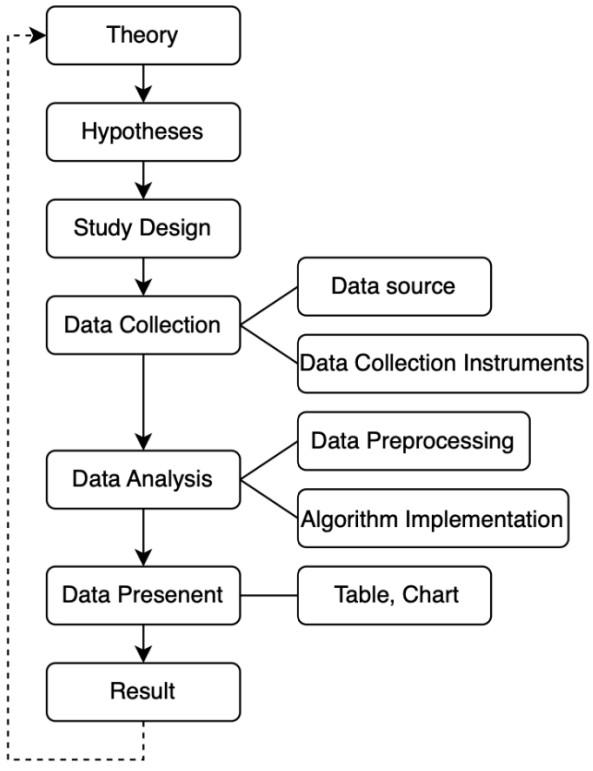


Fig. 1. Research Design

This research focuses on student performance (Excellent, Good, Satisfactory, Needs Improvement, Poor) as the target variable. It will be evaluated through a combination of metrics such as project and assignment scores, midterm and final exam results, peer reviews, core course averages, extracurricular activity participation, and peer evaluations.

To achieve the objective, the study will explore relationships between several key factors:

- Personal Factors: Age, gender, university, and major, which may impact a student's integration within the university environment.
- Academic Factors: Previous GPA and entrance exam scores, which serve as important indicators of academic readiness and engagement.
- Socio-Economic Factors: Family income and financial aid, which can significantly influence a student's ability to succeed academically [9].
- Lifestyle Factors: Health condition, mental health status, diet quality, and study habits (such as hours spent studying per week), all of which play a role in academic performance [10].

The research will follow these steps:

1) Define Problem and Objectives: The goal is to create a highly accurate model to predict student performance using a diverse range of features.

2) Data Collection: Data will be sourced from Kaggle and other relevant databases.

3) Data Preprocessing:

- Data Cleaning: Handle missing data, remove duplicates, and address outliers to ensure data quality.
- Categorical Encoding: Convert categorical features into numerical form (using label encoding).
- Addressing Class Imbalance: Techniques like Random Oversampling (ROS) or SMOTE will be used to handle imbalanced classes, particularly the "Excellent" category.
- Normalization/Standardization: Ensure that features are on a comparable scale to avoid bias from dominant features.

4) Data Splitting: The dataset will be divided into training and testing subsets for model training, hyperparameter tuning, and performance evaluation.

5) Model Training: Train Random Forest (RF) and Deep Neural Network (DNN) models on the prepared training data.

6) Model Evaluation:

- K-fold Cross-Validation: To enhance model robustness, we will use K-fold cross-validation. This ensures consistent performance across different data splits by dividing the data into K subsets and rotating the validation and training sets.
- Evaluation Metrics: The performance will be assessed using accuracy, precision, recall, F1-score, and classification error. While accuracy is widely used, the F1-score will be particularly insightful for handling imbalanced datasets as it balances precision and recall across different classes.

7) Optimization: Fine-tuning the models based on evaluation results to maximize predictive accuracy.

B. Planned Sample Strategy

As this research is data-driven, the source and quality of the dataset are of utmost importance. For this study, we will use the *Australian Student Performance Dataset*, which is publicly available on Kaggle [11]. This dataset comprises 100,256 records and includes a variety of features related to student performance in higher education institutions across Australia, such as academic metrics, personal characteristics, and socio-economic factors.

The dataset from Kaggle will serve as the primary source for our analysis. Since it is a publicly accessible repository, there is no need to engage directly with individuals or organizations for data collection, simplifying the data acquisition process. The dataset provides comprehensive information necessary to investigate the relationships between various influencing factors and student outcomes.

Given the likelihood of an imbalanced distribution in student performance categories (e.g., fewer records in the "Excellent" category Fig.5. will show that), we will employ techniques such as Random Oversampling (ROS) and Synthetic Minority Over-sampling Technique (SMOTE) to address class imbalance. These methods will help ensure that the predictive models perform well across all categories by improving representation of minority classes.

C. Planned Data Collection Methods

As this research is data-driven, the source and quality of the dataset are critical. For this study, we will utilize the Australian Student Performance Dataset, which is publicly available on Kaggle. The dataset contains 100,256 records, including features related to student performance across Australian higher education institutions, such as academic metrics, personal characteristics, and socio-economic factors.

To incorporate the data into the project, the dataset will be downloaded as a zip file from Kaggle, then extracted and saved as a CSV file within the Python project folder. This CSV file will serve as the primary input for data analysis.

Rather than focusing on complex algorithmic design at this stage, we will use 'pandas' to correctly read the CSV file, ensuring that the dataset is properly loaded. We will test this by outputting the first few rows to confirm the data's accuracy and by checking the total number of records to verify its completeness. This will ensure that the dataset is ready for further analysis without any issues in the initial data import.

D. Planned Variables

In this study, we plan to collect both quantitative and categorical data related to various factors that influence student performance. The data will be sourced from the Australian Student Performance Dataset and will include the following:

Quantitative Data: These are numerical values that can be measured and analysed statistically.

1) Age: Represents the student's age.

2) GPA: The Grade Point Average is a key indicator of academic performance.

3) Entrance Exam Score: Scores from entrance exams, showing academic readiness.

4) Family Income: The financial background of the student's family.

5) Hours of Study per Week: The time students dedicate to studying, which directly impacts performance.

Categorical Data: These are categorical variables representing non-numerical information.

1) Gender: Student's gender categorized as female or male.

2) University ID: Identifies the university the student attends (universityA, universityB, universityC).

3) Major: The academic field of study (EE, ME, CS, CE, BA).

4) Financial Aid: Indicates whether a student receives financial support (0 = No, 1 = Yes).

5) Health Condition: Self-reported health status (Excellent, Good, Fair, Poor).

6) Mental Health Status: Student's mental health condition, categorized similarly to health condition (Excellent, Good, Fair, Poor).

7) Diet Quality: Assesses the quality of the student's diet (Excellent, Good, Fair, Poor).

8) Performance: Our project target variable (Excellent, Good, Satisfactory, Needs Improvement, Poor)

F. Planned Data Analysis Methods

In this study, we will adopt a mixed-methods approach, utilizing both quantitative and categorical data analysis techniques. Quantitative Data Analysis: This component will focus on statistical techniques to analyse numerical data. The methods include:

- Descriptive Statistics: To summarize the data, we will calculate metrics such as mean, median, mode, and standard deviation, which provide a clear overview of student performance across different categories.
- Inferential Statistics: We will employ methods like regression analysis and hypothesis testing to draw conclusions and make predictions based on the sample data.
- Machine Learning Algorithms: We will utilize models like Random Forest and Deep Neural Networks to predict student performance based on various features, allowing for complex relationships between variables to be explored.

Categorical Data Analysis: This aspect will involve analysing categorical data through:

- Frequency Analysis: To determine the distribution of different performance categories among students (e.g., how many students fall into each performance category).
- Cross-Tabulation: This will help us explore relationships between categorical variables, such as comparing performance across different majors or universities.

Data Preprocessing: Preprocessing is essential to ensure data quality and suitability for analysis. The steps will include:

Data Cleaning: This involves handling missing values, removing duplicates, and addressing outliers, which ensures the integrity of the dataset and prevents skewed results.

Categorical Encoding: This is crucial for converting categorical variables into a numerical format suitable for machine learning algorithms. This process simplifies input for the algorithms, preserves ordinal relationships when relevant, and ensures compatibility with statistical methods. Ultimately, it enhances the modelling of relationships between various features and student performance.

Feature	Category	Encoded Value
Gender	Female	0
	Male	1
University ID	UniversityA	0
	UniversityB	1
	UniversityC	2
Major	EE	0
	ME	1
	CS	2
	CE	3
	BA	4
Financial Aid	No	0
	Yes	1
Health Condition	Excellent	0
	Good	1
	Fair	2
	Poor	3
Mental Health Status	Excellent	0
	Good	1
	Fair	2
	Poor	3
Diet Quality	Excellent	0
	Good	1
	Fair	2
	Poor	3
Performance	Excellent	0
	Good	1
	Satisfactory	2
	Needs Improvement	3
	Poor	4

Fig. 2. Categorical Encoding Table

Addressing Class Imbalance: We will implement techniques like Random Oversampling (ROS) or SMOTE (Synthetic Minority Over-sampling Technique) to manage class imbalance, especially for categories with fewer instances, such as "Excellent."

Normalization/Standardization: We will ensure that all numerical features are on a similar scale, preventing any single feature from disproportionately influencing the results.

After these steps, we will move forward with implementing the machine learning algorithms and evaluating model performance to ensure the effectiveness of our predictions.

Random Forest (RF): We will train a Random Forest model, a robust ensemble method that leverages multiple decision trees to improve prediction accuracy. It will help capture complex relationships between the input features and student performance outcomes. The hyperparameters, such as the number of trees and the depth of each tree, will be tuned using grid search or random search for optimal performance.

Deep Neural Networks (DNN): Additionally, a Deep Neural Network will be implemented with multiple hidden layers to capture intricate patterns in the dataset. Hyperparameters like the number of layers, learning rate, activation functions, and neurons in each layer will be adjusted during experimentation to find the most suitable architecture.

Once the models are trained, their performance will be evaluated using several key metrics:

1) Accuracy: A basic measure of how often the model makes correct predictions.
2) Precision and Recall: These metrics will be particularly important for the imbalanced dataset, focusing on how well the models identify the "Excellent" and other smaller categories.
3) F1-Score: A balanced measure that combines both precision and recall, especially useful for imbalanced data.
4) Confusion Matrix: To visualize the performance of the models and identify misclassification patterns.

To ensure robustness, K-fold cross-validation will be employed during model training. This approach allows us to validate the models on different subsets of the data, minimizing the impact of random variations and overfitting.

G. Plan for Data and Outcomes Presentation

For presenting our data and outcomes, we will utilize a range of Python tools and techniques to ensure comprehensive and informative analysis.

Descriptive Statistics for Quantitative Data: We will employ the 'pandas' library to calculate descriptive statistics that summarize the quantitative data. Key metrics such as mean, median, mode, and standard deviation will be computed to provide insights into the central tendency and variability of variables like GPA and entrance exam scores. This statistical analysis will help in understanding the overall distribution of the data.

Descriptive Statistics:

	GPA	Entrance Exam Score	Hours of Study per Week
count	100256.000000	100256.000000	100256.000000
mean	2.497590	74.429052	19.528337
std	0.864179	14.408504	11.557678
min	1.000000	50.000000	0.000000
25%	1.750000	62.000000	9.000000
50%	2.500000	74.000000	20.000000
75%	3.240000	87.000000	30.000000
max	4.000000	99.000000	39.000000

Fig. 3. Descriptive Statistics

Frequency Analysis for Categorical Data: For the categorical variables, we will conduct frequency analysis to assess the distribution of categories such as performance. Using 'pandas', we will count the occurrences of each category and display these counts in a clear and organized manner.

Performance Category Counts:

Performance	
Satisfactory	29968
Needs Improvement	25223
Good	20093
Poor	14908
Excellent	10064

Fig. 4. Performance Category Counts

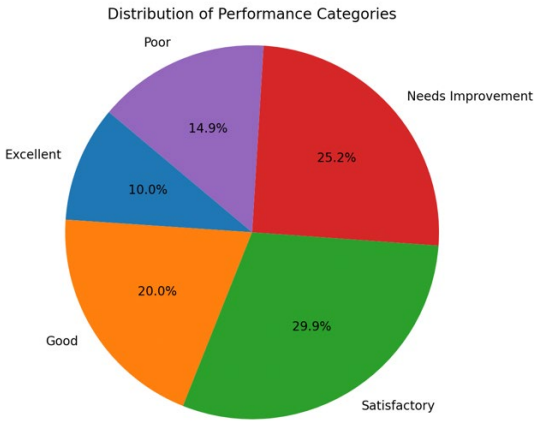


Fig. 5. Pie Chart for Performance Category Counts

After analysing the performance data, it will become evident that the count for the "Excellent" category is significantly lower, underscoring the necessity of using techniques like Random Oversampling (ROS) and SMOTE to address class imbalance.

Cross-Tabulation: We will create cross-tabulations using 'pandas' to explore relationships between different categorical variables, such as the relationship between gender and performance level. This technique allows us to observe how different categories interact with one another, providing deeper insights into the data.

Cross-Tabulation of Gender and Performance:

Performance	Excellent	Good	Needs Improvement	Poor	Satisfactory
Gender					
F	5084	10001	12400	7464	14945
M	4980	10092	12823	7444	15023

Fig. 6. Cross-Tabulation of Gender and Performance

Classification Reports: We will generate classification reports using 'scikit-learn', which will include metrics such as precision, recall, and F1 score for each performance category. This report will provide a detailed overview of the model's performance, allowing us to identify strengths and weaknesses in our predictive capabilities.

Visualization Tools: To enhance our data presentation, we will utilize 'matplotlib' and 'seaborn' for visualizations. We will create bar charts to compare the frequency of different performance categories, allowing for easy visual comparison of the distribution. Additionally, we will visualize descriptive statistics through histograms and box plots to illustrate the spread and central tendencies of quantitative variables.

By combining these Python tools and techniques, we aim to provide a robust analysis of both quantitative and categorical data, facilitating a clear understanding of the factors influencing student performance and their interrelationships.

III. Ethical Consideration

In conducting this study, we are committed to ensuring the ethical treatment of all participants. The following ethical concerns will be addressed:

A. Gender Equality and Non-Discrimination

We will ensure that gender does not introduce bias in the study. All participants, regardless of gender, will be treated equally. The study aims to explore factors affecting student performance without any discrimination.

B. Privacy Protection

Participant privacy will be strictly protected. Personal data such as family income, health status, and mental health will be anonymized and securely stored. Only authorized personnel will have access to the data.

C. Informed Consent

All participants will be fully informed about the purpose of the study, the data collection process, and their right to withdraw at any time. Participation is voluntary, and the data collected will be used only for research purposes.

D. Data Usage and Publication

Data collected will only be used for research purposes and reported in aggregate form. We will ensure that findings are presented transparently and without bias, and no individual's data will be identifiable. Importantly, the results of this study will not be shared with professors and will not affect students' academic evaluations or final grades.

IV. Conclusion

In conclusion, this study aims to shed light on the multifaceted influences on student performance by employing a robust quantitative research design. By leveraging the Australian Student Performance Dataset, we will systematically explore relationships among various factors through detailed data analysis and machine learning techniques. The planned methodological steps, including data preprocessing, categorical encoding, and the application of Random Oversampling (ROS) and SMOTE for handling class imbalance, are integral to enhancing the accuracy of our predictive models.

The anticipated outcomes will not only provide insights into the key determinants of student success but also contribute valuable knowledge for educational stakeholders seeking to improve academic outcomes in higher education settings. The findings will be presented using a combination of statistical measures and visualizations, ensuring clarity and comprehensibility for both researchers and practitioners.

V. References

[1] G. Bowrey, "DEA analysis of Performance Efficiency in Australian Universities using Student Success Rates against Attrition, Retention and Student to Staff Ratios," *Journal of New Business Ideas & Trends,* vol. 18, no. 1, pp. 21-27, 2020.

[2] E. P. I. Sergi Rovira, "Data-driven system to predict academic grades and dropout," *PLOS ONE,* vol. 12, no. 2, 2017.

[3] M. K. P. Agathe Merceron, "Predicting Student Academic Performance at Degree Level: A Case Study," *I.J. Intelligent Systems and Applications,* vol. 01, pp. 49-61, 2015.

[4] C. R. J. Beaulac, "Predicting University Students' Academic Success and Major Using Random Forests.," *Res High Educ ,* vol. 60, pp. 1048-1064, 2019.

[5] M. S. a. A. A.-E. A. Nabil, "Prediction of Students' Academic Performance Based on Courses' Grades Using Deep Neural Networks," *IEEE Access,* vol. 9, pp. 140731-140746, 2021.

[6] G. S. E. A. r. f. g. t. Biau, "A random forest guided tour," *TEST,* vol. 25, pp. 197-227, 2016.

[7] I. K. Vora D, "A Survey of Inferences from Deep Learning Algorithms," *Journal of Engineering and Applied Sciences,* vol. 12, pp. 9467-9472, 2017.

[8] H. W. K. A. N. a. B. S. B Santoso, "Synthetic Over Sampling Methods for Handling Class Imbalanced Problems : A Review," *IOP Conference Series: Earth and Environmental Science,* vol. 58, p. 012031, 2016.

[9] S. N. F. N. Musarat Azhar, "IMPACT OF PARENTAL EDUCATION AND SOCIO-ECONOMIC STATUS ON ACADEMIC ACHIEVEMENTS OF UNIVERSITY STUDENTS," *European Journal of Psychological Research,* vol. 1, p. 1, 2024.

[10] T. P. F. T. W. e. a. Zając, "Student mental health and dropout from higher education: an analysis of Australian administrative data," *High Educ ,* vol. 87, p. 325–343, 2024.

[11] Kaggle, "Australian_Student_PerformanceData (ASPD24)," 2024. [Online].Available:https://www.kaggle.com/datasets/nasirayub2/australian-student-performancedata-aspd24?resource=download. [Accessed 2024].