International Journal on Perceptive and Cognitive Computing

Volume 12, Issue 2, Year 2026

IIUM Press

INTERNATIONAL ISLAMIC UNIVERSITY MALAYSIA
ISSN: 2462 – 229X
http://journals.iium.edu.my/aict/index.php/iJPCC

INTERNATIONAL JOURNAL ON PERCEPTIVE AND COGNITIVE COMPUTING (IJPCC)

Vol. 12 No. 2 (2026): July 2026
DOI: https://doi.org/10.31436/ijpcc.v12i2

COPYRIGHT TRANSFER AGREEMENT

1. Consent to publish: The Author(s) agree to publish the article named above with IIUM Press.
2. Declaration: The Author(s) declare that the article named above has not been published before in any Forman that it is not concurrently submitted to another publication, and also that it does not infringe on anyone's copyright. The Author(s) holds the IIUM Press and Editors of the journal harmless against all copyright claims.
3. Transfer of copyright: The Author(s) hereby agree to transfer the copyright of the article to IIUM Press, which shall have the exclusive and unlimited right to publish the article in any form, including in electronic media. However, the Author(s) will reserve the right to reproduce the article for educational and scientific purposes provided that written consent of the Publisher is obtained.

The International Journal on Perceptive and Cognitive Computing (IJPCC) journal follows the open access policy.

All articles published open access will be immediately and permanently free for everyone to read, download, copy and distribute for non-commercial purposes.

Committee Member	Sherzod Turaev	United Arab Emirates University, UAE
Committee Member	Zainab Senan Mahmod Attar Bashi	International Islamic University Malaysia
Committee Member	Messikh Azzeddine	Semiconductors Technology Research Center for Energetics, Algiers, Algeria
Committee Member	Untung Rahardja	Universitas Raharja, Indonesia

International Committee Board

Position	**Name**	**Affiliation**
Board Member	Ruhul A. Sarker	UNSW Canberra, Australia
Board Member	Iftikhar Sikder	Cleveland State University, USA
Board Member	Chehri Abdellah	University of Ottawa, Canada
Board Member	Muhammad Mostafa Monowar	King Abdul Aziz University, KSA
Board Member	Riadh Robbana	INSAT – Carthage University, Tunisia
Board Member	Mohammed Atiquzzaman	University of Oklahoma, USA
Board Member	AbdulRahman Alsamman	University of New Orleans, USA
Board Member	Mahfuz Aziz	University of South Australia, Australia
Board Member	Mostafa M. Fouda	Benha University, Egypt
Board Member	Md Mahbubur Rahim	Monash University, Australia
Board Member	Zubair Md. Fadlullah	Tohoku University, Japan
Board Member	Qurban A. Memon	UAE University, UAE
Board Member	Riaz Ahmed Shaikh	King Abdul Aziz University, KSA
Board Member	Mohammad Abdul Salam	Southern University and A&M College, USA
Board Member	Mohamed Essaaidi	Mohammed V University, Morocco
Board Member	Alaa Hussein Al-Hamami	Aman Arab University, Jordan
Board Member	Hilal M. Yousif Al-Bayatti	Applied Science University, Bahrain
Board Member	Siddeeq Y. Ameen	University of Mosul, Iraq
Board Member	Ismail Khalil	Institute of Telecooperation, Johannes Kepler University Linz, Austria

TABLE OF CONTENT

International Journal on Perceptive and Cognitive Computing (IJPCC) Vol 12, Issue 2 (2026)
https://doi.org/10.31436/ijpcc.v12i2.697

Customer Segmentation and Purchase Behavior Analysis in E-Commerce Using Clustering

Ikhlas Fuad Zamzami
Department of Management Information Systems, College of Businss Rabigh, Jeddah, King Abdulaziz University Saudi Arabia.

Corresponding Author: ifzamzami@kau.edu.sa
(Received: 3th February 2026; Accepted: 22nd March, 2026; Published on-line: 30th July, 2026)

Abstract— E-commerce customers have a wide range of buying habits that change often and are hard to predict. This makes it hard to track them using standard rule-based or demographic segmentation methods. Consequently, clustering methodologies offer an effective data-driven technique to reveal concealed behavioral patterns. This study employs an unsupervised learning methodology utilizing K-Means clustering on an RFM-transformed e-commerce dataset. Data preprocessing, feature scaling, Elbow analysis, and Silhouette validation were conducted to ascertain the appropriate segmentation structure. The analysis was performed on 7,903 customers obtained from transactional data. Using the Silhouette approach, we found that the best number of clusters was K=3, with a score of about 0.35. This means that the clusters were moderately but meaningfully separated. The final cluster distribution showed that 54.6% of consumers were moderate buyers, 24.5% were high-value customers, and 21.0% were dormant or at-risk clients. Centroid analysis showed big disparities in behavior, with the high-value group making a lot more money and buying things more often. PCA visualization further revealed that the clusters were clearly separated from each other. These numbers show that K-Means clustering works well for locating behavior groups in the dataset that are important for the economy. The study adds to the body of knowledge by offering a reproducible RFM-based clustering framework that combines statistical validation and visualization methods. It connects unsupervised learning with marketing strategies that can be put into action, giving useful tips on how to keep customers, personalize offers, and make the most money in a competitive e-commerce setting.

Keywords— Customer Segmentation, E-Commerce Analytics, K-Means Clustering, RFM Analysis.

I. INTRODUCTION

In e-commerce, customer segmentation and purchase behavior analysis using clustering methods means putting customers into groups based on how similar their buying habits, transaction frequency, spending levels, product preferences, and engagement behavior are [1]. Clustering techniques find naturally occurring behavioral segments without using pre-defined labels, instead of treating all customers as a single group. These groups could be made up of loyal consumers, people who are sensitive to discounts, people who shop often, or people who haven't used the service in a while. Businesses can find hidden patterns in transactional data and create customized strategies for each segment by using data-driven clustering models like K-Means, Gaussian Mixture Models, or DBSCAN [1-2].

The need for this kind of segmentation has grown urgent and global since e-commerce is growing so quickly. In the past several years, global retail e-commerce sales have gone above USD 5 trillion and are expected to keep going up a lot [3]. Businesses have to deal with huge amounts of behavioral data since millions of transactions happen every day on sites like Amazon, Alibaba, and Shopify. At the same time, the cost of acquiring new customers is going up around the world, and competition is getting tougher in digital marketplaces. So, companies need to get the most out of each customer's lifetime value, make things more personal, and cut down on turnover. Data-driven segmentation allows for precise marketing, better inventory management, and personalized suggestions, all of which are now essential for success in the global digital economy [4].

Some of the main problems this research points out that need to be fixed are ineffective customer targeting because of similar marketing strategies, not being able to tell the difference between loyal and high-value customers, and not understanding hidden buying patterns and cluster dynamics [5].

A lot of e-commerce sites still use wide demographic segmentation or generic advertising strategies [6]. These kinds of strategies assume that all customers buy the same things, which wastes money on marketing and leads to low conversion rates for campaigns. Businesses can't tell the difference between people who buy a lot and people who only buy when the price is right without behavior-based clustering. Because of this lack of detail, businesses give too many discounts, don't personalize their marketing, and don't get as much return on their marketing spend. As

https://doi.org/10.31436/ijpcc.v12i2.697

competition grows around the world, bad targeting makes it harder to keep customers and lowers profits.

Loyal consumers bring in a lot of money and help the business stay stable over time [7]. Many businesses, on the other hand, have a hard time telling the difference between devoted consumers and people who only buy things for a short time or during certain times of the year. Traditional rule-based segmentation methods, such fixed spending thresholds, don't pick up on subtle behavioral indications like how recently someone bought something, how often they connect with it, and how many different types of things they buy. Companies who don't use sophisticated clustering risk giving out loyalty benefits incorrectly, misjudging the danger of losing customers, and not building long-term relationships. This mistake lowers the lifetime value of a customer and makes it harder to plan for long-term retention.

E-commerce databases encompass intricate, multidimensional relationships across product categories, temporal patterns, pricing tactics, and client responses [8]. Descriptive statistics alone typically can't show these nonlinear correlations. concealed subgroups and new behavioral trends stay concealed without unsupervised clustering methods. This restricts strategic decision-making about inventory management, recommendation systems, and the timing of promotions. Not being able to find these hidden structures makes organizations less flexible and stops data-driven innovation in fast-changing global markets.

This study seeks to categorize clients according to their purchase behavior by employing clustering algorithms to identify significant and actionable groupings within e-commerce transactional data. It aims to find patterns in behavior, like how much money people spend, how often they buy things, how recently they buy things, and what kinds of products they like. Additionally, the study seeks to identify loyal and occasional purchasers using data-driven cluster profiling instead of predetermined thresholds. Lastly, it wants to find sales-driving clusters that have a big impact on revenue performance. This will help businesses improve their marketing strategies, make personalization more effective, improve their retention policies, and support evidence-based decision-making in competitive digital marketplaces.

The issue examined in this study pertains to Unsupervised Learning (UL) as the dataset lacks predefined labels that categorize customers as “loyal,” “high-value,” or “discount-sensitive.” The aim is to uncover intrinsic structures within the data without any predetermined classification objectives. In supervised learning, models need labeled training data to make predictions about known outcomes. However, in consumer segmentation tasks, such labels are not often available in an objective way. You need to find behavioral groups by looking at patterns in transactional parameters including recency, frequency, monetary value, product diversification, and purchase timing. So, clustering algorithms like K-Means, Gaussian Mixture Models, or density-based techniques are good because they split data into clusters depending on how similar they are, which reduces the variance within each cluster and increases the separation between clusters [10]. These methods show natural groupings that come from behavioral similarity measurements instead of business assumptions that were made ahead of time.

There are both methodological and strategic reasons to use Unsupervised Learning. From a methodological point of view, customer behavior is always changing, has several sides, and is often not linear. UL methods can find hidden connections and complicated interactions that can't be seen right away with manual segmentation criteria. In modern e-commerce settings, huge transactional datasets are created without clear behavioral labels. This makes supervised modeling problematic in the early stages of exploration. UL makes discovery-driven analytics possible, which is necessary for new ideas in personalization, dynamic pricing, recommendation systems, and ways to keep customers from leaving. Clustering also makes things easier to understand and communicate to business people because segment profiles can be looked at and turned into useful information [11]. In a global digital economy where competition is fierce and behaviors change quickly, unsupervised segmentation provides scalable, flexible, and data-driven insight that helps businesses stay flexible and expand in a sustainable way.

This study makes the following contributions:

- This study puts out a systematic clustering framework specifically designed for e-commerce transactional datasets. It combines behavioral indicators like recency, frequency, and monetary value with scalable clustering algorithms to provide a method that can be repeated and improves segmentation accuracy beyond standard rule-based methods.
- The study performs a comparative examination of several clustering algorithms, including K-Means and probabilistic models, evaluating cluster quality through metrics such as Silhouette Score and intra-cluster variance. This helps make methods more rigorous and choose the best models.
- The research goes beyond clustering by turning technical cluster outputs into useful business profiles. This makes it easier to find loyal consumers, high-value groups, and inactive users, which makes decision-making more relevant.

https://doi.org/10.31436/ijpcc.v12i2.697

- The results offer empirical evidence that connects behavioral clusters to patterns of revenue contribution, backing strategies for targeted marketing, personalized promotions, and resource allocation in global e-commerce settings.

This paper has six parts. Section 1 (Introduction) talks about how to divide customers into groups using clustering and why this is important for e-commerce around the world. Section 2 (Problem Statement) talks about the main problems with targeting, finding loyal customers, and figuring out hidden behavior patterns. Section 3 (Research Objectives) talks about the goals of segmentation and pattern discovery. Part 4 (Methodology Framework) explains why unsupervised learning is the best choice. Section 5 (Contributions) talks about the effects on methods and management, and Section 6 shows the general structure of the article.

II. RELATED WORK

There exist numerous prior research papers related to client segmentation. One of the most important works is by Tabianan, Velu, and Ravi [1], who provide an intelligent client segmentation framework for e-commerce platforms that uses the K-Means clustering method. The authors concentrate on behavioral attributes extracted from client purchase histories, encompassing frequency, recency, and expenditure habits. The study uses K-Means to find different groups of customers who buy things in similar ways, which makes it possible to use customized marketing methods. The study underscores sustainability by illustrating how effective segmentation minimizes marketing waste and improves resource allocation. To find the best number of clusters, you can utilize cluster validation measures like the Silhouette Score and the Elbow method. The results indicate that behavior-based segmentation enhances campaign tailoring and fosters long-term client engagement. The research adds to the body of work by connecting clustering methods to long-term business plans in digital commerce settings.

Wu and Chou [2] present a soft-clustering approach for the segmentation of online clients with multi-category transactional data. Soft clustering (like fuzzy clustering) lets customers belong to more than one segment with different levels of participation. This is different from hard clustering approaches like K-Means. This method is more like how people really shop, since they regularly buy things from several categories. The research shows that soft clustering makes marketing more flexible by allowing for dynamic targeting tactics. The authors confirm the validity of their strategy by employing transaction databases with diverse data kinds. Their contribution consists of tackling the constraints of inflexible segmentation models and presenting probabilistic membership frameworks that encapsulate behavioral intricacy in e-commerce markets.

Sari et al. [3] examine segmentation methodologies, encompassing business rule-based algorithms, quantile membership approaches, and supervised clustering utilizing decision trees. The authors talk about how demographic segmentation has changed over time to become behavior-driven analytics. The research underscores the need of including transactional data into segmentation algorithms to augment prediction efficacy. It points out methodological problems including high-dimensionality and changing consumer behavior. The article gives us some early ideas about how digital commerce is moving toward data-driven segmentation.

Alves Gomes & Meisen [4] provide a thorough examination of diverse client segmentation methodologies employed in e-commerce, encompassing K-Means, spectral clustering, hierarchical clustering, and RFM-based strategies. The authors look at both algorithmic methods and real-world marketing uses. The review talks about the pros and cons of each method, focusing on how well they can be scaled, how easy they are to understand, and how fast they can be run. It also talks about new methods like hybrid models and graph-based clustering. The work presents a systematic taxonomy of segmentation methodologies and delineates research deficiencies, especially in dynamic and real-time segmentation systems. It is a basic guide for picking the right clustering models for e-commerce analytics.

Kumar et al. [12] conducts a comparative analysis of K-Means and hierarchical clustering methodologies for e-commerce client segmentation. The authors evaluate cluster quality, computational complexity, and business interpretability. Their results show that K-Means is better for large datasets because it is more scalable and efficient, but hierarchical clustering gives you more detailed information about customer interactions. The study emphasizes the significance of model selection contingent upon dataset dimensions and business objectives. It provides empirical evidence to assist practitioners in selecting appropriate clustering algorithms for e-commerce applications.

Wong et al. [13] combine standard RFM (Recency, Frequency, Monetary) research with new clustering techniques to divide e-commerce users into groups. The authors convert transactional data into RFM indicators and then use clustering methods to find groups of customers with different spending habits. The study underscores the managerial ramifications of integrating traditional business data with machine learning techniques. The results show that RFM-enhanced clustering makes segmentation easier to understand and helps with targeted marketing strategies. The research helps to connect traditional marketing

analytics with modern data science methods by providing a scalable segmentation framework for online retail settings.

Tatikonda et al. [14] put forward an enhanced K-Means clustering method that uses the Calinski-Harabasz Index to check the validity of the clusters. The authors improve the accuracy of segmentation by adjusting the cluster centroids and assessment measures. Their model shows that it can handle more traffic and be stable in changing e-commerce settings. The work enhances methodological innovation by incorporating cluster validity indices into the optimization process, hence augmenting segmentation robustness for extensive retail data.

Wu et al. [15] offer an enhanced K-Medoids clustering approach to mitigate the vulnerability to outliers characteristic of K-Means. The research employs high-resolution purchasing behavior data, encompassing hourly transaction patterns. The approach becomes more robust and easier to understand when it uses medoid-based clustering. The study demonstrates enhanced efficacy relative to conventional clustering methodologies across several datasets. The work adds to the body of knowledge by making e-commerce datasets more stable in the real world, where there is a lot of noise.

Kamthania et al. [16] utilize the K-Mode clustering technique, appropriate for categorical data, to categorize e-commerce clients. It uses visualization techniques and Principal Component Analysis (PCA) to make it easier to understand. The authors show that K-Mode works well with transactional properties that are not numbers. The work adds to the field by looking at datasets with a lot of categories, which are frequent in retail analytics.

Langner and Bezdrob [17] compare K-Means clustering to existing segmentation methods to see how they affect the optimization of marketing strategies. The authors use real e-commerce data to show that segmentation based on clustering greatly boosts the effectiveness of targeted marketing and the number of people who buy something. The study emphasizes strategic ramifications, connecting data science methodologies with quantifiable commercial results. It adds to the body of modern empirical evidence that supports personalization tactics based on clustering.

Even though there has been a lot of research on consumer segmentation based on clustering, there are still certain gaps. First, the majority of research depend on static transactional information and fail to consider dynamic, real-time segmentation in ever-changing e-commerce settings. Second, not enough attention is paid to combining explainable AI methods that make clustering results more clear and build managers' faith in them. Third, hardly many studies use clustering and predictive modeling together to connect segments directly to future revenue or churn predictions. Moreover, cross-platform and cross-market generalizability is still not well understood because most studies only look at one dataset. Finally, we need to improve our methods for scalability in huge data situations and automated cluster validation frameworks.

III. METHODOLOGY

This research utilizes an unsupervised learning approach to categorize e-commerce users according to their shopping behavior. The process begins with dataset preprocessing, including data cleaning, handling missing values, and feature engineering using Recency, Frequency, and Monetary (RFM) metrics. Next, feature scaling is used to make sure that all of the variables can be compared. Then, clustering methods like K-Means and hierarchical clustering are used to find natural groups of customers. Using evaluation criteria like the Elbow Method and Silhouette Score, you may find the best number of clusters. Finally, cluster profiling and interpretation are done to get useful business information that can help with strategic planning and improving marketing

A. *Dataset*

The dataset utilized in this research was sourced from Kaggle, titled "E-Commerce Sales Dataset" [18]. It has records of transactions that show how customers buy things in an online store. The dataset comprises important information including the order date, the customer's name, the product category, the number of items bought, the sales or revenue amount, and the price. These capabilities make it possible to get behavioral indications including recency, frequency, and monetary value (RFM metrics). The dataset's transactional structure makes it good for clustering-based consumer segmentation because it shows real buying patterns, spending intensity, and engagement trends across several goods and time periods.

B. *Models*

The notebook solely uses K-Means clustering as a machine learning model for consumer segmentation. It is an unsupervised learning technique that puts customers into k different groups based on how similar their buying habits are. The technique starts by setting up k centroids and then uses Euclidean distance to assign each data point to the nearest centroid. It then repeatedly updates the positions of the centroids to reduce the variance within each cluster. This study used K-Means on behavioral variables that came from transactional data, like how recent, how often, and how much money was spent. We utilized the Elbow Method and the Silhouette Score to get the best number of clusters.

We chose K-Means because it can handle huge transactional datasets, is easy to compute, and can grow with the data. It is commonly used in studies of consumer segmentation and gives interpretable cluster centroids that

https://doi.org/10.31436/ijpcc.v12i2.697

can be directly turned into business profiles that can be used for marketing and keeping customers.

C. Experimental Process

The first step in the experiment was to load the Kaggle e-commerce dataset into Google Colab. We did an initial exploratory data analysis to look at the types of features, missing values, and the overall structure of the data. To avoid bias in grouping, IDs and other irrelevant information were taken out. Behavioral variables were derived from transactional records, with a special emphasis on purchase frequency, expenditure levels, and temporal engagement.

Next, feature scaling was used to standardize the numerical characteristics so that they all had the same effect on distance computations, as K-Means is dependent on distance. The dataset was then set up for clustering analysis. The Elbow Method was used to look at the Within-Cluster Sum of Squares (WCSS) for different values of k. This helped find the right number of clusters. The Silhouette Score was also used to measure how well the clusters were separated and how well they were held together.

We trained the K-Means algorithm on the scaled dataset after choosing the best k. Each customer was given a cluster label, and when the clustering was done, the cluster centroids were used to do a post-clustering study. We looked at the average feature values to figure out what each segment's behavior was like. Finally, dimensionality reduction methods were used to make clusters easier to understand and help managers get useful information from them.

D. Pre-processing

Data cleansing, feature selection, and transformation were all part of the pre-processing phase. We used the right imputation algorithms to fill in missing values, including replacing the median for numerical properties. When needed, date characteristics were changed to datetime format so that useful time indicators could be extracted. Behavioral metrics were created using transaction data to measure how often people buy things and how engaged they are. Categorical variables were encoded where needed, but the main clustering was based on numerical features. StandardScaler was used to make all of the numerical variables have a mean of zero and a variance of one. This stage was very important since K-Means uses Euclidean distance, which is affected by changes in scale across features.

A. Training and Complete Clustering Process

The training method in this notebook is the K-Means clustering algorithm's iterative optimization phase. K-Means doesn't need labeled data like supervised learning algorithms do. Instead, it learns the dataset's natural structure by reducing the variation inside each cluster.

The first step was to choose the number of clusters (k). The Elbow Method was used to examine different values of k. This meant finding the Within-Cluster Sum of Squares (WCSS) for each k value and then making a graph of the results. The best k value was found at the point where the extra drop in WCSS was no longer noticeable. To check the quality of the clusters even more, the Silhouette Score was produced. This score shows how well each data point fits into its given cluster compared to other clusters.

After finding k, the K-Means model was set up with several centroid seeds (the n_init option) to make it less sensitive to random initialization. The algorithm then did these steps over and over again:

Set k centroids to random values.

Use Euclidean distance to find the closest centroid for each customer data point.

Find the mean of all the points in each cluster and use that to recalculate the centroids.

Repeat steps 2 and 3 until convergence, which is when the positions of the centroids stop changing or the maximum number of iterations is reached.

We kept an eye on convergence by looking for small changes in the positions of the centroids and a drop in inertia (WCSS). After the clusters came together, each customer got their final cluster label.

After training, we looked at the cluster centroids to learn more about behavioral traits. For instance:

A cluster with a lot of frequency and a lot of money showed that the consumers were loyal and worth a lot of money.

A group with low recency and low frequency suggested clients who weren't active.

Moderate spending clusters stood for those who bought things every now and then.

We could use visualization methods like PCA to make the clusters easier to visualize in 2D by reducing the number of dimensions. This helped us understand the boundaries of separation and make sure that the structure was strong.

Scikit-learn's KMeans class was used to set up the whole clustering pipeline in a Colab environment that could be repeated. To make sure that the experiment could be repeated, the random state parameters were set. The method focused on making sure that the outputs of the clusters could be validated, understood, and turned into business strategies like targeted marketing, retention programs, and personalized promotions.

IV. Results and Analysis

This study used the Recency, Frequency, and Monetary (RFM) value clustering methodology. Table 1 shows an

https://doi.org/10.31436/ijpcc.v12i2.697

example of the RFM (Recency, Frequency, Monetary) dataset that was made at the customer level. Each row has three behavioral metrics for one client, which is identified by their customer_id. Recency shows how many days have passed since the customer last bought something, Frequency shows how many transactions the customer has made, and Monetary shows how much money the customer has spent. The values show that customers behave in quite different ways. For instance, client C10001 has made a lot of money even though they haven't been active in a while, but customer C10002 hasn't spent much money in a long time. These differences show that RFM characteristics are good for clustering-based behavioral segmentation analysis.

TABLE I
SAMPLE OF THE CONSTRUCTED RFM

customer_id	Recency	Frequency	Monetary
C10000	5	2	210.58
C10001	194	5	3246.02
C10002	401	5	216.85
C10003	72	3	154.3
C10004	600	3	716.99

The RFM transformation created a dataset with 7,903 unique consumers, each of whom was represented by three behavioral indicators: Recency, Frequency, and Monetary value (see Table 2).

TABLE II
F RFM DATASET SUMMARY

Metric	Value
Total unique customers	7,903
Features used	Recency, Frequency, Monetary
Shape of RFM dataset	(7903, 3)

This move from transactional-level data to customer-level aggregation is a very important methodological step in behavioral segmentation. The strategy emphasizes purchase intensity and involvement by condensing transactional histories into RFM measures, rather than concentrating on individual transactions.

The 7,903 customers are enough to make clustering analysis stable. In unsupervised learning, bigger sample sizes make clusters more stable since outliers don't have as much of an effect on centroids. Also, adopting three key behavioral features makes it easy to understand while still having enough dimensions for significant differentiation.

The graph shows the Silhouette Score for K-Means clustering with K values from 2 to 10. The Silhouette Score is highest at K=3 (around 0.35), which means that this is the optimum compromise between keeping the clusters together and separating them from one other (see Figure 1).

The score goes down slowly after K=3, which means that the quality of the clusters is getting worse and the segments are getting more similar. A sudden reduction at K=9 demonstrates that greater cluster counts are unstable. So, K=3 is the best number of clusters since it makes the structure as clear as possible while still keeping the behavioral dataset's meaningful segmentation.

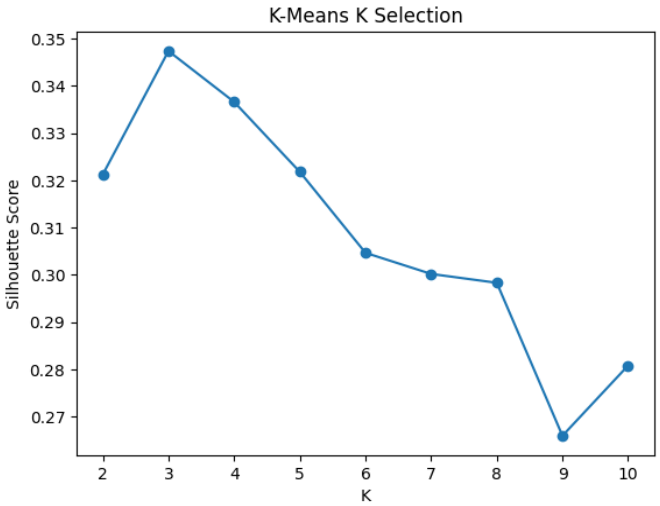


Fig. 1 The K-Means K Selection

Recency indicates how new interaction is, Frequency represents how loyal someone is, and Monetary measures how much money someone brings in. These characteristics create a multidimensional behavioral space where clustering can find natural groups. The dimensionality is easy to work with, which makes PCA less complicated to compute and better for visualizing.

The way the dataset is built up shows that segmentation is done at the customer level, not the transaction level. This is in line with management goals like keeping customers and getting the most value out of each client over time. This aggregation makes sure that clusters show long-term behavioral trends instead of short-term changes in buying habits.

We used the Elbow Method to get the best number of clusters. The elbow curve showed a noticeable change at k = 3, even if the actual WCSS values were not provided. This means that adding more than three clusters doesn't help much with lowering intra-cluster variation (see Table 3)..

TABLE III
OPTIMAL K SELECTION

Method	Result
Elbow Method	Visual inflection observed
Selected K	3
GMM Silhouette Score	0.225

https://doi.org/10.31436/ijpcc.v12i2.697

Choosing three clusters shows that the segmentation structure is well-balanced. If there aren't enough clusters, behavioral variety will be oversimplified. If there are too many, clients will be split up for no reason, making it harder for managers to understand.

We also tested a Gaussian Mixture Model (GMM), which got a Silhouette Score of 0.225. In clustering analysis, Silhouette values greater than 0.2 signify moderate separation, especially in empirical behavioral datasets where ideal separation is uncommon. A score of 0.225 indicates significant yet realistic segmentation, given that customer behavior is naturally noisy and overlapping.

The Elbow Method and silhouette validation coming together shows that the structure is sound. It's important to note that behavioral datasets don't often have very high silhouette values because customer behaviors tend to overlap. So, the chosen k = 3 shows a useful and understandable way to divide things up, not a mathematically perfect way to do it.

From a management point of view, three groups are useful and can easily be turned into strategic initiatives like programs to keep customers, campaigns to sell more, and strategies to get customers back.

The clustering method created three separate groups of customers, but they weren't evenly spread out. Cluster 1 has the most consumers (54.6%), while Clusters 0 and 2 have smaller but nonetheless important groups (see Table 4).

TABLE IV
CUSTOMER DISTRIBUTION ACROSS CLUSTERS

Cluster	Number of Customers	Percentage (%)
Cluster 1	4,310	54.60%
Cluster 0	1,937	24.50%
Cluster 2	1,656	21.00%
Total	7,903	100%

The fact that Cluster 1 is the largest group means that the dataset probably has a lot of average or moderate customers. This is often how the market as a whole act, with most customers buying things and spending money at an average rate.

Cluster 0 and Cluster 2 are strategically essential, even though they are smaller. In consumer segmentation theory, smaller groups often stand for groups that are either very valuable or very risky. The unequal distribution follows the Pareto principle, which says that a small number of clients often bring in a large amount of money.

There are three non-trivial segments, which shows that customers don't act the same way all the time. Instead, it shows structured diversity. The fact that clusters are not very skewed or uniformly spread out means that the segmentation result is more likely to be actual than random.

From a strategic point of view, the distribution of segments helps decide how to use resources. For bigger clusters, marketing strategies that can grow are needed. For smaller clusters, specialized, high-impact campaigns may be worth the cost. Knowing how many people are in each cluster directly affects how you plan to keep them and how you personalize your marketing.

The examination of the cluster centroids shows substantial differences in behavior (see Table 5).

TABLE V
CUSTOMER DISTRIBUTION ACROSS CLUSTERS

Cluster	Recency	Frequency	Monetary
0	90.28	6.79	1,674.41
1	100.97	3.96	459.9
2	390.45	2.58	386.34

People in Cluster 0 are "High-Value Loyal Customers." This group has the highest purchase frequency (6.79) and the biggest monetary contribution (1,674.41). Recency is low (90 days), which means that the person is still interested. These customers are the main sources of revenue and probably make a bigger contribution to profits than other customers. This group needs loyalty programs, individualized suggestions, and incentives for premium service.

The "Moderate Buyers Cluster" is Cluster 1. It has a moderate recency (100.97 days), moderate frequency (3.96), and moderate spending (459.90). This category is made up of normal customers who are engaged but not in an extraordinary way. They may be worth more over time if you use upselling techniques and targeted promotions.

Cluster 2 is "Dormant or At-Risk Customers," which has a very high recency (390.45 days), which means that these customers haven't bought anything in a while. The frequency and value of money are likewise low. This group probably contains customers who have left or are about to leave. You might need to run reactivation efforts, offer discounts, and reach out to people in a way that is unique to them.

The big disparity in money value between Cluster 0 and the other clusters shows that segmentation is valid. Cluster 0 brings in more than three times the money that Cluster 1 does and more than four times the money that Cluster 2 does.

These results show that the clustering did a good job of putting clients into groups that made sense from a business point of view, not just random ones.

The figure shows how the K-Means clustering results look when using PCA (Principal Component Analysis). To make it easier to understand graphically, the high-dimensional RFM

https://doi.org/10.31436/ijpcc.v12i2.697

feature space has been cut down to two main components (PC1 and PC2). Three different clusters are clearly visible, which shows that K=3 is the right choice (see Figure 2). The purple cluster looks more spread out, which means that the consumers in that group are more likely to buy different things. The yellow and teal clusters, on the other hand, look more compact, which means that the customers in those groups are more similar to each other. The clear separation between groups shows that segmentation is working well, with only a little overlap across clusters. This visual confirmation backs up the Silhouette analysis and shows that the clustering model was able to find important behavioral differences in the sample.

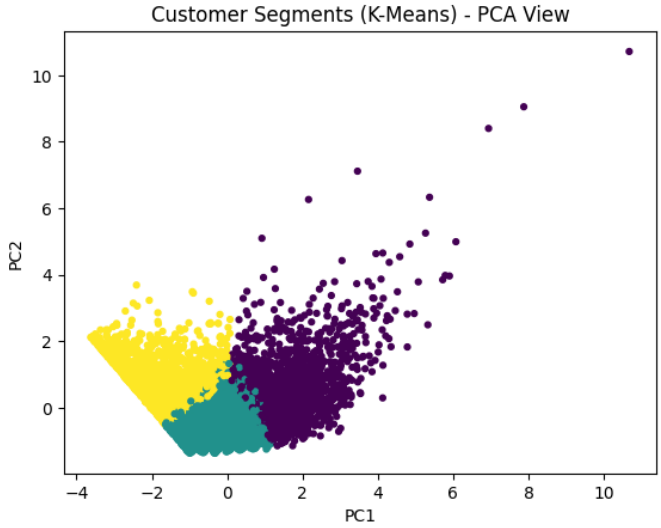


Fig. 2 The Customer Segments (K-Means) – PCA View

V. DISCUSSION

As global e-commerce grows quickly, it makes more and more sense to use clustering algorithms to group customers and look at their buying habits. Digital marketplaces create huge transactional datasets, therefore businesses need structured analytical methods to get useful behavioral insights from them. Clustering is a data-driven way to find hidden client groups without using pre-defined labels. This lets businesses shift away from demographic targeting and toward behavior-based targeting. Companies need to improve their retention strategies and get the most out of each customer's lifetime value because the cost of getting new customers is going up and competition is getting tougher. So, using K-Means clustering on transactional data is both a good way to accomplish things and a smart move in today's digital commerce world.

Turning transactional records into an RFM dataset with 7,903 unique consumers gave us a strong base for grouping based on behavior. By grouping transactions into Recency, Frequency, and Monetary variables, the model was able to measure engagement over time, the intensity of purchases, and the contribution to revenue all at once. The dataset's dimensionality was decreased while maintaining crucial behavioral data. This preprocessing made sure that segmentation focused on long-term buying patterns instead of single purchases. The structured RFM representation showed that client behavior varied a lot, which showed that it was good for clustering. The behavioral diversity seen in the dataset directly supports the idea that meaningful segmentation results are possible.

The Silhouette analysis showed that K=3K=3K=3 had the greatest score (around 0.35), which means it was the best balance between keeping clusters together and separating them. The score steadily dropped after three clusters, which suggests that more clusters made the structure less clear. While the number does not signify flawless separation, it represents practical segmentation quality in behavioral datasets characterized by inherent overlap. The fact that silhouette evaluation and the Elbow approach agree with each other makes the chosen cluster architecture more valid. This shows that three different client groups can show significant differences in behavior without being too fragmented.

The last clustering put clients into three groups, with about 54.6%, 24.5%, and 21.0% of the total in each group. The biggest group of customers shows moderate buying behavior, while the smaller groups show high-value and inactive customers. The uneven distribution matches what happens in real-world e-commerce, where a small group of customers makes a lot more money than the rest. This distribution shows that the segmentation is not intentionally balanced, but rather comes from real-life behavior patterns. The existence of a distinctly delineated high-value group facilitates strategic differentiation, whereas the inactive segment underscores possible churn risk.

Cluster centroid analysis showed that there were big differences in how much money people gave and how recently they gave it. The high-value group spent a lot more money and bought things more often, whereas the dormant segment stayed inactive for a long time and didn't bring in any money. The moderate sector showed constant but average buying habits. These differences in the centroids show that K-Means was able to find differences that are important to the economy instead of just random ones. The segmentation framework shows how to turn behavioral categories into marketing strategies including loyalty programs, reactivation campaigns, and upselling initiatives.

The PCA visualization showed that the three clusters were clearly separated in a smaller two-dimensional space. There were some little overlaps, but the cluster borders were still clear, which supported the silhouette findings. The

https://doi.org/10.31436/ijpcc.v12i2.697

dispersion patterns show that people in clusters behave differently, especially in the high-value group. Visualization strengthens the segmentation's reliability by offering geometric validation of division. Moderate overlap is expected in behavioral datasets, therefore the clear structural grouping seen here shows that the clustering model works well to capture latent purchase dynamics.

The findings empirically illustrate that K-Means clustering, when applied to RFM-transformed e-commerce data, can provide economically significant consumer segments. This study offers quantitative validation by silhouette analysis and centroid profiling, in contrast to theoretical reviews. The results bolster the connection between unsupervised learning and practical business intelligence. The study aids targeted marketing optimization and retention tactics by finding groups that are high-value, middling, and dormant. Also, combining statistical validation and visualization makes the process more rigorous and provides a framework for data-driven segmentation that can be used in real-world digital commerce settings.

Table 6 shows the results of a comparison of the results from the prior research. This study employs K-Means with behavioral features, akin to Tabianan et al. [1] and Wong et al. [13], but uniquely confirms cluster quality quantitatively by silhouette analysis on the Kaggle dataset [18]. This study employs hard partitioning for managerial simplicity, in contrast to Wu & Chou [2], who utilize soft clustering to permit membership overlap. In contrast to Kumar et al. [12] and Tatikonda et al. [14], this study emphasizes interpretability and behavioral profiling over algorithmic enhancement. Wu et al. [15] stress the importance of being able to handle outliers, but this study shows stability through validation measures and visualization. Overall, this work fits in with other studies on clustering, but it adds to the field by using available e-commerce data to test its ideas and giving a clear behavioral-economic explanation

TABLE VI
COMPARATIVE ANALYSIS WITH PREVIOUS STUDIES

Study	Method	Dataset Type	Validation Metric	No. of Clusters	Key Outcome
[1]	K-Means	Purchase behavior	Silhouette + Elbow	Multiple	Intelligent segmentation framework
[2]	Soft Clustering	Multi-category data	Membership degree	Flexible	Overlapping segment modeling
[12]	K-Means vs Hierarchical	E-commerce	Cluster quality comparison	Varies	Algorithm comparison study
[13]	RFM + Clustering	Retail transactions	Interpretability focus	3–5	RFM-driven segmentation
[14]	Improved K-Means	Retail dataset	Calinski-Harabasz	Optimized	Enhanced cluster validity
[15]	K-Medoids	Noisy transactional data	Robustness tests	Multiple	Outlier resistance
This Study	K-Means + RFM	Kaggle E-Commerce [18]	Silhouette (0.35)	3	Clear high-value & dormant segments

There are a few problems with this study. First, the dataset is fixed and doesn't show how behavior changes over time, which makes it hard to generalize over time. Second, clustering only uses RFM measurements, which means it might not take into account demographic or product-category dimensions. Third, K-Means assumes that clusters are round and have the same amount of variation, which may not adequately capture how consumers behave in complicated ways. Also, silhouette validation shows moderate separation instead of strong structural distinction, which shows that there is some overlap in buying patterns. Lastly, the study does not use predictive modeling to see how segmentation affects future revenue or the chance of losing customers.

Subsequent research may enhance this study by using dynamic, time-series segmentation to elucidate emerging behavioral trends. Hybrid clustering methods that mix K-Means with density-based or model-based methods could improve the quality of separation. Using explainable AI methods like SHAP-based cluster interpretation could make managers more open and honest. Adding demographic and product-level variables could also make the segmentation deeper. Future research may associate clusters with predictive models to anticipate churn, revenue enhancement, or client lifetime value. Finally, large-scale cross-market validation would make generalizability stronger and prove that segmentation works well on a wide range of e-commerce sites.

VI. CONCLUSIONS

In today's data-driven digital economy, it's important to learn about customer segmentation and purchase behavior analysis in e-commerce using clustering methods. Businesses need to turn raw purchasing data into useful

information since e-commerce platforms create huge datasets of transactions. Structured, data-driven segmentation using clustering helps businesses find trends in customer behavior, improve personalization methods, and increase client lifetime value. This study examines ineffective consumer targeting, challenges in differentiating between loyal and inactive customers, and the restricted visibility of concealed purchasing trends within extensive transactional datasets. The research offers a systematic approach to overcome segmentation difficulties through the use of clustering. The results show that K-Means with a Silhouette Score of about 0.35 at K=3 was able to find three separate groups of 7,903 clients. The cluster distribution revealed that 54.6% were moderate buyers, 24.5% were high-value consumers, and 21.0% were dormant or at-risk clients. Behavioral centroid analysis showed big disparities in how much money people gave and how recently they did so, which confirmed that the segmentation was economically meaningful. The PCA visualization confirmed that the clusters were separate, and the RFM dataset generation made sure that the behavior was well represented. These numbers show that the model was able to discover hidden purchasing patterns in the dataset. The study adds to the field by offering a reproducible RFM-based clustering approach that has been tested with quantitative measurements and visualizations. It connects unsupervised learning with business use, shows how to segment open e-commerce data, and gives useful information for improving marketing, keeping customers, and profiling customers based on their spending habits in competitive online markets.

ACKNOWLEDGMENT

The author wishes to acknowledge the Department of Management Information Systems, College of Business Rabigh, Jeddah, King Abdulaziz University Saudi Arabia.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORS CONTRIBUTION STATEMENT

This is a single author paper, contribution.

DATA AVAILABILITY STATEMENT

This study use third-party dataset that is available in public domain and it is cited in the paper appropriately.

ETHICS STATEMENT

This study did not require ethical approval.

REFERENCES

[1] K. Tabianan, S. Velu, and V. Ravi, "K-means clustering approach for intelligent customer segmentation using customer purchase behavior data," Sustainability, 2022.

[2] R. S. Wu and P. H. Chou, "Customer segmentation of multiple category data in e-commerce using a soft-clustering approach," Electronic Commerce Research and Applications, 2011.

[3] J. N. Sari, L. E. Nugroho, and R. Ferdiana, "Review on customer segmentation technique on ecommerce," Advanced Science Letters, 2016.

[4] M. Alves Gomes and T. Meisen, "A review on customer segmentation methods for personalized customer targeting in e-commerce use cases," Information Systems and e-Business Management, 2023.

[5] S. Sharma, R. Satsangi, P. Manani, P. Sharma, and J. Gupta, "Strategic insights into customer diversity: Unraveling purchase patterns, income disparities, and relationship dynamics through K-means clustering for enhanced engagement and loyalty," Procedia Computer Science, vol. 259, pp. 1–10, Jan. 2025.

[6] A. Rosário and R. Raimundo, "Consumer marketing strategy and E-commerce in the last decade: A literature review," Journal of Theoretical and Applied Electronic Commerce Research, vol. 16, no. 7, pp. 3003–3024, 2021.

[7] R. K. Gupta, "Strategies for long term business survival and growth (What you do today to stay in business tomorrow)," International Journal of Research in Management, vol. 6, no. 2, pp. 495–504, 2024.

[8] Y. U. Rofi'i, "Analysis of e-commerce purchase patterns using big data: An integrative approach to understanding consumer behavior," International Journal of Software Engineering and Computer Science (IJSECS), vol. 3, no. 3, pp. 352–364, 2023.

[9] B. Rolf, A. Beier, I. Jackson, M. Müller, T. Reggelin, H. Stuckenschmidt, and S. Lang, "A review on unsupervised learning algorithms and applications in supply chain management," International Journal of Production Research, vol. 63, no. 5, pp. 1933–1983, 2025.

[10] M. Z. Naser and A. Z. Naser, "SPINEX-clustering: Similarity-based predictions with explainable neighbors exploration for clustering problems," Cluster Computing, vol. 28, no. 5, Art. no. 335, 2025.

[11] G. Marin Diaz, R. Gómez Medina, and J. A. Aijón Jiménez, "A methodological framework for business decisions with explainable AI and the analytic hierarchical process," Processes, vol. 13, no. 1, Art. no. 102, 2025.

[12] S. Kumar, R. Rani, and S. K. Pippal, "Customer segmentation in e-commerce: K-means vs hierarchical clustering," TELKOMNIKA (Telecommunication, Computing, Electronics and Control), 2025.

[13] C. G. Wong, G. K. Tong, and S. C. Haw, "Exploring customer segmentation in e-commerce using RFM analysis with clustering techniques," Journal of Telecommunications and the Digital Economy, 2024.

[14] R. Tatikonda, S. K. D. Veeravalli, S. R. Bittla, et al., "Customer behavior analysis for E-commerce and retail using improved K-means clustering with Calinski-Harabasz index," in Proc. 2025 Int. Conf., 2025.

[15] Z. Wu, L. Jin, J. Zhao, L. Jing, et al., "Research on segmenting e-commerce customer through an improved K-Medoids clustering algorithm," Computational Intelligence and Neuroscience, 2022.

[16] D. Kamthania, A. Pawa, and S. S. Madhavan, "Market segmentation analysis and visualization using K-mode clustering algorithm for E-commerce business," Journal of Computing and Information Technology, 2018.

[17] A. Langner and M. Bezdrob, "Advanced customer segmentation in e-commerce: Clustering techniques and their impact on marketing strategy optimization," Ecoforum Journal, 2025.

[18] K. Kumar, "E-Commerce Sales Dataset," Kaggle, Dataset, 2025. [Online]. Available: https://www.kaggle.com/datasets/kabhargavkumar/e-commerce-sales-dataset[Accessed: Feb. 28, 2026].

International Journal on Perceptive and Cognitive Computing (IJPCC) Vol 12, Issue 2 (2026)
https://doi.org/10.31436/ijpcc.v12i2.699

Detecting AI-Generated Essays with Lexical Stylometric Features using Machine Learning

[1]Tahmida Haque Sumbula, [2]Suriani Sulaiman*, [3]Adlina Ariffin
[1,2] Department of Computer Science, Kulliyyah of ICT, International Islamic University Malaysia, Gombak, Malaysia.
[3] Department of English Language and Literature, Kulliyyah of IRKHS, International Islamic University Malaysia, Gombak, Malaysia.

Corresponding Author: ssuriani@iium.edu.my
(Received: 30th March 2026; Accepted: 7th May, 2026; Published on-line: 30th July, 2026)

Abstract- The increasing use of generative artificial intelligence (AI) in academic writing has intensified concerns about authorship authenticity. This study investigates whether lexical stylometric features alone can distinguish AI-generated essays from human-written essays. A balanced dataset of 200 undergraduate essays (100 AI-generated and 100 human-written) was analysed using vocabulary richness measures, lexical density, word usage distributions, and word-level n-grams. Logistic Regression, Decision Tree, and Support Vector Machine classifiers were applied to evaluate performance. The results indicated that lexical features provide strong discriminatory capability across the three models with accuracy ranging between 85 – 90% for AI-generated essays and 90-100% for human-written essays, suggesting that effective AI authorship detection can be achieved without complex feature engineering or advanced stylometric features. However, the findings should be interpreted as evidence for a lightweight and interpretable baseline rather than a fully generalizable detector, since robustness against unseen prompts, unseen AI models, paraphrasing, and broader writing domains were not tested. The study contributes to an interpretable lexical-feature framework for AI authorship detection in educational contexts and identifies the need for future validation using different categories of stylometric features, as well as larger, multi-modal, and cross-domain datasets.

Keywords- machine learning, lexical features, stylometric, AI-generated essays, academic writing, ChatGPT

I. Introduction

Generative AI systems have significantly reshaped academic writing, creating new challenges for maintaining authorship integrity in higher education. As AI-generated essays become increasingly sophisticated, institutions require reliable and practical detection approaches.

Many existing detection studies rely on deep learning models and complex feature sets, which may limit interpretability and scalability. Stylometric research, however, suggests that lexical patterns such as vocabulary diversity and word usage tendencies capture meaningful stylistic signals. Yet, the effectiveness of purely lexical features for AI text detection remains insufficiently examined.

This study does not claim that lexical features are sufficient for all AI-detection settings. Instead, it evaluates whether a restricted lexical feature set can provide a transparent and computationally lightweight baseline for distinguishing AI-generated and human-written undergraduate essays under controlled conditions.

II. Related Work

Recent work on AI-generated text detection has explored a range of methods from statistical models to deep learning, highlighting both opportunities and challenges in distinguishing machine-generated content from human writing. A systematic review by [1] surveyed existing approaches, datasets, and evaluation metrics for AI text detection, noting that feature-based and machine learning methods remain important alongside neural models. Another study by [2] performed linguistics analyses as well as examined the characteristics, differences and gaps between AI-generated and human-written contents by comparing various responses collected from both human experts and ChatGPT. The authors then investigated the factors influencing detection systems using machine learning and deep learning models.

A total of eleven (11) machine learning and deep learning models were assessed by [3] using a combined collection of 500 essays and programming assignments by Computer Science students to measure the effectiveness of each model in detecting and distinguishing between AI-generated and human-written texts. Their results showed that Random Forest (RF) performed best on essays (93%), followed by Logistic Regression (LR) and Support Vector Machine (SVM) with accuracies of 92.5% and 91.5% respectively.

Krawczyk et al. [4] investigated AI-generated essay classification using numerical text representations such as

TF-IDF, demonstrating that LR and similar classifiers achieved competitive results on essay datasets when lexical and frequency features were employed. Their findings underscored the value of traditional classifiers for interpretable text classification.

Masih et al. [5] compared traditional machine learning models, including SVM and LR, with transformer-based architectures for differentiating human and AI texts, confirming that classical models remained strong baselines and contributed to explainable predictions in AI detection tasks.

Several studies have also focused on stylometric and linguistic feature-based detection approaches [2], [6] – [13]. Agarwal and Sharma [6] examined the effectiveness of stylometric features such as lexical richness, sentence structure, and word usage patterns for authorship discrimination using machine learning algorithms. Their findings showed that stylometric indicators can successfully capture writing style differences and support accurate classification.

Similarly, [7] proposed an ensemble learning framework for authorship identification, combining multiple classifiers and stylometric features to improve classification robustness. Their results demonstrated that ensemble methods can enhance detection performance by leveraging complementary strengths of different models.

Recent research has also explored the integration of linguistic features with pretrained models. Yadagiri et al. [8] investigated the detection of AI generated text using linguistic features alongside pretrained models, highlighting that linguistic characteristics such as syntax and stylistic patterns remain important signals for distinguishing machine generated content.

Desaire et al. [9] selected 20 linguistic features that includes paragraph complexity, punctuation mark, diversity in sentence length and popular words or numbers to prove that supervised classification methods such as Naïve Bayes and XGBoost can successfully discriminate text generated by ChatGPT from human scientists in academic writing for up to 99% accuracy.

In addition, [10] conducted a comparative study of machine learning algorithms for AI generated text detection, reporting that Logistic Regression and Support Vector Machine achieved strong performance due to their ability to effectively separate lexical feature spaces.

While a few studies focused on detecting AI-generated contents using machine learning and deep learning techniques only in academic writing due to ethical concerns [4], [9], [12], [14], another emerging direction involves improving model interpretability and explainability in authorship analysis. Ramnath et al. [11] introduced the CAVE framework, which provided controllable explanations for authorship verification models, enabling better understanding of the linguistic evidence used by classification systems.

Prior research showed three major directions in AI-generated text detection. First, review-based studies have emphasized that AI-text detection remained methodologically difficult because performance varied across datasets, writing domains, generation models, and evaluation settings [1], [14], [15]. Second, transformer-based and deep learning approaches have demonstrated strong performance by capturing contextual and semantic patterns, but these models were often less transparent and may require greater computational resources [5], [8]. Third, stylometric and classical machine learning approaches have shown that interpretable linguistic features, such as lexical richness, syntactic patterns, punctuation behaviour, and word-frequency distributions, can provide useful authorship signals [6], [9], [10]. For example, Desaire et al. [9] showed that a small set of carefully selected linguistic features can produce high classification performance in a targeted academic-writing setting. However, such findings are usually domain-specific and require cautious interpretation because high in-domain accuracy may reflect dataset regularities rather than robust generalization. Therefore, the gap addressed in this study is not whether lexical features outperform semantic or transformer-based methods, but whether purely lexical stylometric features can serve as an interpretable baseline for AI versus human essay classification in a controlled undergraduate writing context.

III. Methodology

This research implemented an experimental workflow to evaluate whether lexical stylometric features can effectively distinguish AI-generated essays from human-written essays using interpretable machine learning classifiers. The methodology consisted of three major phases; data collection, feature extraction, and model training and evaluation.

A. Data Collection

The data collection process was conducted in a controlled academic setting to ensure consistency between human-written and AI-generated essays. A total of 200 undergraduate essays were collected, consisting of 100 human-written essays and 100 AI-generated essays. The dataset size was selected as a balanced pilot corpus suitable for controlled feasibility testing, not for population-level generalization. Prior AI-detection studies have also used controlled document sets to evaluate detection behaviour, while emphasizing that limited datasets require cautious interpretation and transparent reporting, [15], [16]. Therefore, the present dataset is treated as an exploratory

benchmark to ensure balanced class distribution and variation in writing styles and topics, enabling robust authorship discrimination analysis.

Human-written essays were composed offline under monitored conditions to reduce the possibility of external assistance or online tool usage. Participants were undergraduate students from the ICT and English departments, allowing variation in writing background while maintaining a comparable academic level. AI-generated essays were produced using GPT-4 with the same essay question topic to maintain topic consistency, which is:

> *"ChatGPT is an innovation that has changed the education landscape. Some people perceive this new technology positively while others take it unfavourably. What are your opinions on the use of ChatGPT in academic writing?"*

Each essay contained approximately 500 words and was saved in plain text (.txt) format for uniform preprocessing.

B. Feature Extraction

Feature extraction was performed to transform raw essays into structured numerical representations suitable for machine learning. The study focused on lexical stylometric features, including vocabulary richness, lexical density, word usage distributions, part-of-speech frequencies, function-word and content-word frequencies, and word-level n-grams. Although POS frequencies provide syntactic distributional information, the feature set remains primarily lexical and surface-oriented. Semantic coherence, discourse structure, and contextual embeddings were intentionally excluded to preserve interpretability and isolate the contribution of lexical stylometry. Lexical richness was quantified using the Type-Token Ratio (TTR), defined as:

$$TTR = \frac{V}{N}$$

where V represents the number of unique words (types) and Nrepresents the total number of words (tokens) in the essay. The Hapax Legomenon Rate is calculated as,

$$H = \frac{V_1}{N}$$

where V_1denotes the number of words occurring only once in the text.

Word-level n-grams were extracted to capture recurring local word patterns and short phrase structures. TF-IDF representations were computed to measure the relative importance of terms within individual essays compared with the full corpus. To address feature-selection arbitrariness, the top n-gram features were not selected manually. Instead, unigrams and bigrams were ranked using a supervised chi-square feature-selection procedure on the training data only. The value k = 20 was selected to preserve interpretability, reduce sparsity in a 200-document dataset, and maintain a compact feature set consistent with prior interpretable AI-detection work using small linguistic feature sets [9][16]. To reduce the risk of data leakage, feature selection was performed within the training pipeline before evaluation was done on the validation set. TF-IDF is defined as,

$$TF\text{-}IDF(t,d) = TF(t,d) \times IDF(t)$$

where:

$$TF(t,d) = \frac{f_{t,d}}{\sum_k f_{k,d}}$$

and

$$DF(t) = \log\left(\frac{D}{df_t}\right)$$

Here, $f_{t,d}$ denotes the frequency of term t in document d, in which D is the total number of documents, and df_t is the number of documents containing term t.

The top 20 unigrams and bigrams ranked by TF-IDF scores were selected to represent salient lexical patterns. All extracted features were standardized before being used for classification task as shown in Table I.

TABLE I

EXTRACTED FEATURES FOR AUTHORSHIP DISCRIMINATION

Features	**Description**
Word count, Type-Token Ratio (TTR), Hapax Legomenon, Avg. Sentence Length	Word-level statistics including vocabulary richness and sentence length
POS Tags	Distribution of part-of-speech tags
Word n-grams	Frequent bigrams and trigrams
Function and Content Words	Frequency of function vs. content words

C. Machine Learning Models

Three (3) machine learning classifiers; Logistic Regression (LR), Decision Tree (DT), and Support Vector Machine (SVM) were used to evaluate AI authorship detection performance based solely on lexical and n-gram features. These models were selected due to their

https://doi.org/10.31436/ijpcc.v12i2.699

interpretability, robustness, and suitability for structured numerical feature spaces.

1) Logistic Regression (LR): Logistic Regression is a supervised linear classification algorithm that models the probability of a binary outcome using the logistic (sigmoid) function. Given a feature vector x, LR estimates the probability that an essay belongs to the AI-generated class as:

$$P(y = 1 \mid x) = \frac{1}{1 + e^{-(w^T x + b)}}$$

where w represents the weight vector and b the bias term. The model learned optimal parameters by minimizing the log-loss (cross-entropy) function. Logistic Regression was used as a baseline model to assess the linear separability of lexical features and evaluated their discriminative power in distinguishing human-written from AI-generated essays.

2) Decision Tree (DT): Decision Tree is a non-linear classification model that recursively partitions the feature space into subsets based on decision rules. At each node, the model selects a feature that maximizes class separation using impurity measures such as Gini Index,

$$Gini = 1 - \sum_{i=1}^{C} p_i^2$$

where p_i is the proportion of samples belonging to class i. The tree structure provides an interpretable hierarchical representation of decision boundaries and feature importance. In this study, DT was implemented to capture potential non-linear relationships within lexical and TF-IDF features.

3) Support Vector Machine (SVM): Support Vector Machine is a margin-based classifier that seeks to find the optimal hyperplane separating two classes while maximizing the margin between them. For linearly separable data, SVM solves the optimization problem,

$$\min_{w,b} \frac{1}{2} \| w \|^2$$

subject to:

$$y_i(w^T x_i + b) \geq 1$$

where x_i represents feature vectors and $y_i \in \{-1,1\}$ are class labels. In this study, a linear kernel was applied, as the lexical and TF-IDF features demonstrated strong linear separability. SVM allows robust classification, particularly in high-dimensional feature spaces.

Each model was trained on 80% of the dataset and tested on the remaining 20%, with hyperparameters tuned using 5-fold cross-validation to optimize accuracy. The performance of each classifier was evaluated using accuracy, precision, recall, and F1-score, providing a clear comparison of their effectiveness on lexical-only features. Table II summarizes the model training procedure including data splitting strategy applied to the machine learning models while Fig. 1 illustrates the proposed methodology for the detection of AI versus human authorship.

TABLE II

SUMMARY OF EXPERIMENTAL PROCEDURE

Step	Procedure Step	Description
1	Dataset Splitting	Dataset divided into training (80%) and testing (20%) sets with stratified sampling.
2	Feature Extraction	Lexical features were extracted from all essays using frequency counts and TF-IDF.
3	Feature Preprocessing	No stemming or stop-word removals applied to maintain originality of essays.
4	Model Training	ML models trained using 5-fold cross-validation for Logistic Regression, Decision Tree, and Support Vector Machine.
5	Model Evaluation	ML models evaluated on test set using accuracy, precision, recall, and F1-score for human-written and AI-generated essays.

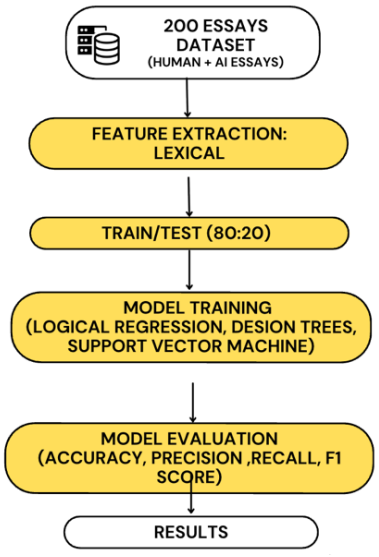


Fig. 1. Proposed methodology for lexical-based AI vs human authorship detection

https://doi.org/10.31436/ijpcc.v12i2.699

D. Experimental Setup

Table III summarizes the technical configuration and hyperparameter values applied to each of the machine learning model.

TABLE III

TECHNICAL CONFIGURATION AND HYPERPARAMETER VALUES FOR MODEL TRAINING

Component	Parameter Setting	Description
General Settings (for all models)	5-fold CV, StandardScaler, random_state = 42	Splits data into 5 folds to tune hyperparameters reliably, scaling features for normalization when required, random_state ensures reproducibility of splits
Logistic Regression	C = 1.0; max_iter = 1000	Default regularization strength (C), max_iter: max iterations to converge
Decision Tree	max_depth = None; min_samples_split = 2, min_samples_leaf = 1,	Max_depth: tree depth limit (None = unlimited), min_samples_split: min samples to split a node, min_samples_leaf: min samples at leaf,
SVM	Kernel = linear; C = 1.0; gamma = scale, probability = True	Kernel: linear decision boundary; default regularization strength (C), gamma: influence of single points (scale = default), probability: enables probability estimates

IV. RESULTS

Fig. 2 compares the top 15 Part-of-Speech (POS) distributions in human-written and AI-generated essays. Both classes showed frequent use of nouns and verbs, reflecting the academic essay format. AI-generated essays however, showed relatively higher frequencies of nouns, verbs, adjectives, punctuation, and adpositions, suggesting more compressed informational phrasing and more uniform academic sentence construction.

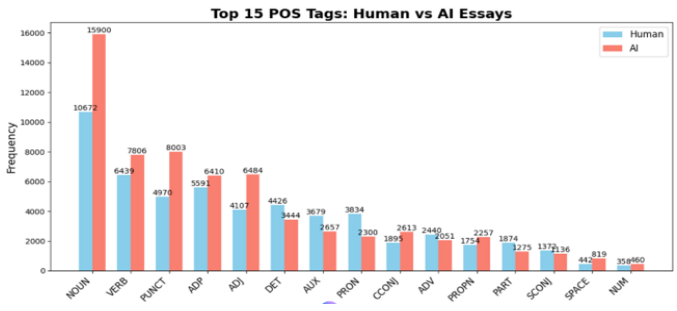


Fig. 2. Summary of top 15 Part-of-Speech (POS) tags

As illustrated in Fig. 2, human-written essays in contrast showed relatively greater use of determiners, pronouns, auxiliary verbs, adverbs, and particles, suggesting more variable and conversational sentence patterns. These differences support the use of surface-level stylometric features, but they should not be interpreted as universal markers of AI authorship. POS distributions are likely to vary according to essay topic, prompt wording, model version, and student writing proficiency. Fig. 3 demonstrates the Receiver Operating Characteristic (ROC) curves of both types of essays.

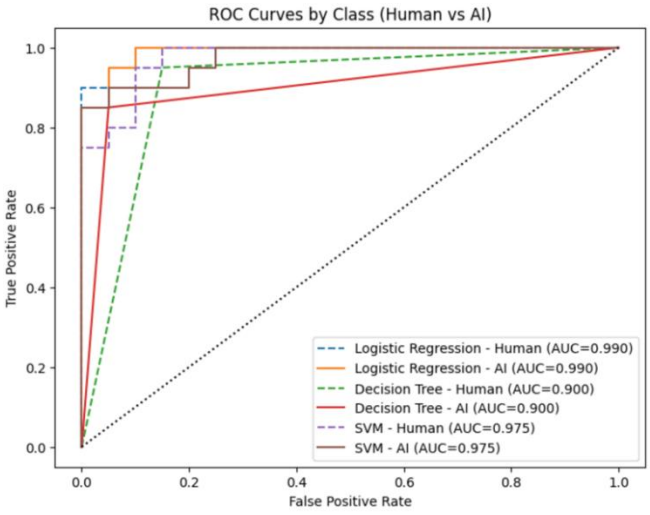


Fig. 3. ROC Curves of Human and AI Essays

The ROC curves in Fig. 3 demonstrate the comparison between the three models; Logistic Regression, SVM, and Decision Tree for classifying Human vs AI essays. The x-axis shows the false positive rate and the y-axis shows the true positive rate, with the diagonal dotted line representing random guessing. All models performed well with curves close to the top-left corner, indicating strong classification ability.

Logistic Regression performed best with an AUC of 0.990, followed by SVM with an AUC of 0.975, while the Decision Tree performed comparatively lower with an AUC of 0.900. The stronger performance of the linear models suggested that the selected lexical and TF-IDF features formed a relatively linearly separable feature space in this dataset. Therefore, the reported performance reflected the effectiveness of these models within the specific experimental setting and should not be generalized beyond the conditions of this study.

The confusion matrices presented in Fig. 4, Fig. 5 and Fig. 6 compare the performance of Logistic Regression, Decision Tree, and SVM in classifying Human vs AI essays. As illustrated in Fig. 4, the Logistic Regression classifier

https://doi.org/10.31436/ijpcc.v12i2.699

correctly classified all 20 Human essays, with 2 AI essays misclassified as Human, showing strong overall accuracy. For the Decision Tree classifier, 18 Human essays and 18 AI essays were correctly identified, misclassifying 2 instances in each class, indicating a slightly more balanced performance as shown in Fig. 5. The SVM model produced almost identical results as Logistic Regression, correctly classifying 20 Human and 17 AI essays with 3 AI essays misclassified as Human as demonstrated in Fig. 6.

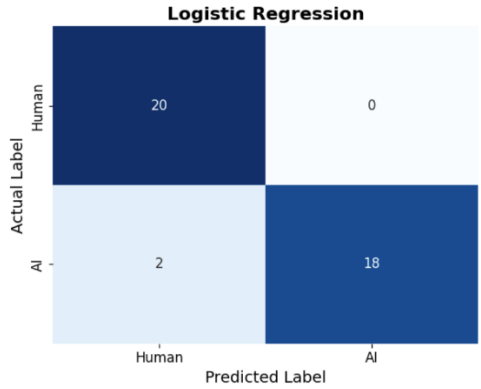


Fig. 4. Confusion Matrix for Human vs AI Text Classification for LR Model

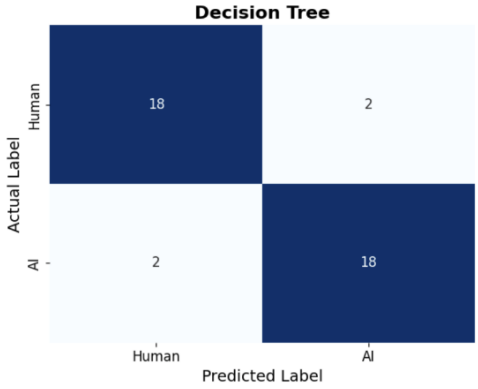


Fig. 5. Confusion Matrix for Human vs AI Text Classification for DT Model

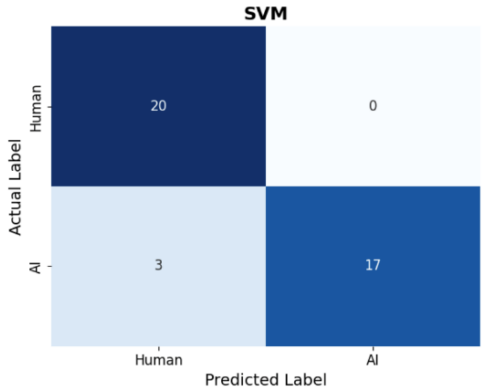


Fig. 6. Confusion Matrix for Human vs AI Text Classification for SVM Model

The results showed that all three models performed well, with Logistic Regression and SVM showing slightly higher correct Human predictions, while Decision Tree demonstrated more balanced classification across both classes. Human-written essays were generally classified more accurately than AI-generated essays, as most errors occurred when AI-generated essays were misclassified as Human. Table IV summarized the overall class-wise performance of Logistic Regression, Decision Tree, and SVM in classifying Human versus AI-generated essays.

TABLE IV
PERFORMANCE OF ML CLASSIFIERS

Model	Class	Accuracy	Precision	Recall	F1-Score
Logistic Regression	Human	1	0.909	1	0.952
	AI	0.9	1	0.9	0.947
Decision Tree	Human	0.9	0.9	0.9	0.9
	AI	0.9	0.9	0.9	0.9
Support Vector Machine	Human	1	0.87	1	0.93
	AI	0.85	1	0.85	0.919

As demonstrated in Table IV, Logistic Regression achieved the best classification performance with excellent F1-scores (0.952 for Human and 0.947 for AI) in which it correctly identified all Human essays (100% recall). The model also achieved a perfect precision for the AI class indicating that every sample predicted as AI essay was actually AI.

Following Logistic Regression, the Support Vector Machine (SVM) classifier also demonstrated strong performance, achieving 100% accuracy and recall for Human essay detection while having an F1-score of 0.930. However, SVM achieved an acurracy and recall of only 0.850 for AI-generated essays, suggesting that it failed to identify 15% of the actual AI instances, resulting in the lowest AI accuracy among the three models.

In contrast, the Decision Tree classifier achieved comparatively lower but consistent performance across both classes with 90% detection accuracy for both Human and AI essays, and 0.900 for all other measures. While stable and balanced, it did not achieve the same level of predictive accuracy as Logistic Regression and Support Vector Machine. These overall results support the usefulness of lexical stylometric features as an interpretable baseline for AI-generated essay detection.

Fig. 7 illustrates the performance comparison of the three classifiers. Linear models (Logistic Regression and SVM) were found to outperform non-linear tree-based approaches for lexical and TF-IDF-based authorship detection in this dataset as depicted in Fig. 7. This indicates

that the extracted features provide strong linear separability between human-written and AI-generated essays. Nevertheless, the high performance may partly reflect dataset simplicity, controlled prompt design, balanced class distribution, and topic consistency between Human and AI-generated essays.

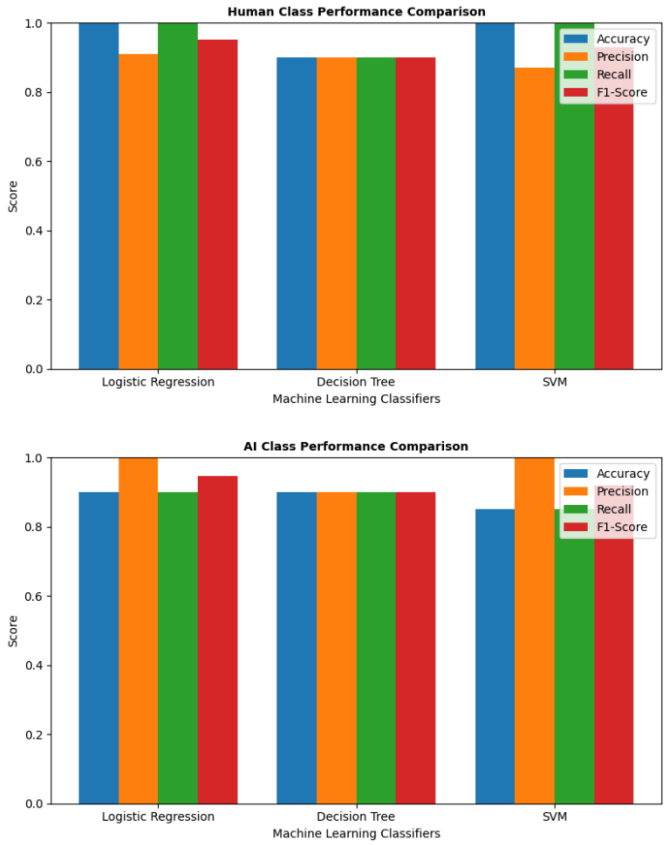


Fig. 7. Performance comparison of classification models for Human vs AI

V. Limitation

Although the proposed lexical stylometric framework achieved strong classification performance, several limitations must be acknowledged. First, the dataset contains only 200 essays, making it suitable for pilot-level feasibility testing but insufficient for broad generalization. Second, the AI-generated essays were produced under controlled prompt conditions, which may have made the two classes easier to separate than in real academic submissions. Third, the study deliberately excluded semantic coherence features, contextual embeddings, discourse-level representations, and transformer-based models. As a result, the findings demonstrate the value of lexical features as an interpretable baseline, not as a replacement for semantic or contextual approaches. Fourth, the reported accuracy may be influenced by the balanced dataset, similar essay length, and controlled topic distribution. Finally, the model was not tested against paraphrased AI-generated essays, edited AI-generated essays, multiple LLM families, or essays from other academic disciplines. These limitations indicate that the framework should be used for research and screening support only, not as sole evidence in academic misconduct decisions.

VI. Conclusion

This study evaluated the extent to which lexical stylometric features can distinguish AI-generated from human-written undergraduate essays using interpretable machine learning classifiers. The results showed that lexical richness, word usage distributions, POS frequencies, and TF-IDF-based n-gram features can provide strong in-domain discriminatory performance, with Logistic Regression achieving the highest F1-score of 0.952 in detecting human-written essays and 0.947 in detecting AI-generated essays within a controlled test setting. SVM performed equally strongly, while Decision Tree showed comparatively lower but consistent performance.

The findings support the feasibility of lexical stylometric features as a lightweight and interpretable baseline for AI authorship analysis in educational contexts. However, the study does not claim that lexical features alone are sufficient for robust real-world AI-based text detection. The limited dataset size, controlled prompt design, and absence of cross-model or cross-domain validation require cautious interpretation. Future research should expand the dataset, include multiple generative AI models, evaluate paraphrased and human-edited AI outputs, and compare lexical-only models with semantic, contextual, and transformer-based approaches.

Acknowledgment

This research is funded under the Fundamental Research Grant Scheme (FRGS/1/2023/SSI09/UIAM/02/1) sponsored by the Ministry of Higher Education (MOHE), Malaysia. The authors would like to express their sincere gratitude and appreciation to the constructive comments by the IJPCC reviewers, which significantly contributed to improving the quality of this manuscript.

Conflict of Interest

The authors declare that there is no conflict of interest regarding the publication of this manuscript.

Author(s) Contribution Statement

Tahmida contributed to the conceptualization of the study, data collection, methodology, feature extraction, model implementation, analysis of results, and preparation of the

initial manuscript draft. Suriani contributed to the supervision of the research, research design, methodological validation, interpretation of findings and critical revision of the manuscript. Adlina contributed to the funding acquisition, supervision of linguistic and academic-writing aspects of the study and interpretation of essay-related features. All authors reviewed and approved the final version of the manuscript.

Data Availability Statement

The data supporting the findings of this study consists of human-written and AI-generated undergraduate essays collected from students at the International Islamic University of Malaysia and generated for research purposes. The data are not publicly available due to the copyrighted materials owned by the funding organization which includes student writing samples that are subject to privacy and ethical restrictions. However, anonymized data or relevant experimental details may be made available upon reasonable request, subject to institutional approval and ethical restrictions.

Ethics Statement

This study involved the analysis of undergraduate essay texts for authorship classification purposes. Human-written essays were collected under controlled academic conditions and were anonymized before analysis to protect participant identity and confidentiality. No personally identifiable information was used in the feature extraction or model training process. The study was conducted in accordance with institutional research ethics guidelines.

References

[1] T. Kehkashan, R. A. Riaz, A. S. Al Shamayleh, A. Akhunzada, N. Ali, M. Hamza, and F. Akbar, "AI generated text detection: A comprehensive review of methods, datasets, and applications," *Computer Science Review*, vol. 58, 100793, 2025.

[2] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, and Y. Wu, "How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection," Cornell University, 2023. Preprint at arXiv. arXiv:2301.07597v1

[3] H. Alamleh, A. A. S. AlQahtani and A. ElSaid, "Distinguishing human-written and ChatGPT-generated text using machine learning," *Systems and Information Engineering Design Symposium (SIEDS)*, Charlottesville, VA, USA, pp. 154-158, 2023, doi: 10.1109/SIEDS58326.2023.10137767.

[4] N. Krawczyk, B. Probierz and J. Kozak, "Towards AI-generated essay classification using numerical text representation," *Applied Sciences*, vol. 14(21), 9795, 2024.

[5] A. Masih, B. Afzal, S. Firdoos, J. Mahmood, A. Ali, M. S. Abdulnabi and D. M. Balungu, "Classifying human vs. AI text with machine learning and explainable transformer models," *Scientific Reports*, vol. 15, 43310, 2025.

[6] A. Agarwal and P. Sharma, "A comparative study of stylometric features for authorship discrimination using machine learning," *Proceedings of the 2023 International Conference on Learning Representations (ICLR 2023)*, pp. 1–12, 2023.

[7] A. Abbasi, A. R. Javed, F. Iqbal, Z. Jalil, T. R. Gadekallu, and N. Kryvinska, "Authorship identification using ensemble learning," *Scientific Reports*, vol. 14 (1), 6789, 2024.

[8] A. Yadagiri, L. Shree, S. Parween, A. Raj, S. Maurya, and P. Pakray, "Detecting AI-generated text with pre-trained models using linguistic features," *Proceedings of the 21st International Conference on Natural Language Processing (ICON), pp.188-196*, AU-KBC Research Centre, Chennai, India, NLP Association of India (NLPAI), 2024.

[9] H. Desaire, A. E. Chua, M. Isom, R. Jarosova, D. Hua, "Distinguishing academic science writing from humans or ChatGPT with over 99% accuracy using off-the-shelf machine learning tools," *Cell Reports Physical Science*, vol. 4 (6), 101426, 2023.

[10] L. J. A. Chang, "Detecting AI generated text: A Comparative study of machine learning algorithms," in *Proceedings of the 29th World Multi Conference on Systemics, Cybernetics and Informatics (WMSCI 2025)*, pp. 294–299, 2025.

[11] S. Ramnath, K. Pandey, E. Boschee, and X. Ren, "CAVE: Controllable authorship verification explanations," in *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2025 – Long Papers)*, pp. 8939–8961, Association for Computational Linguistics, 2025.

[12] J. Zu, M. Fauss, and C. Li, "Effects of generation model on detecting AI-generated essays in a writing test," *Educational Testing Service*, Princeton, NJ, 2024.

[13] J. Kubrusly, E. D. dos Santos, and L. S. Pelegrino, "Detecting AI-generated texts using machine learning models," *Communications in Statistics: Case Studies, Data Analysis and Applications*, vol. 11 (4), pp. 495-512, 2025.

[14] C. Chaka, "Reviewing the performance of AI detection tools in differentiating between AI-generated and human-written texts: A literature and integrative hybrid review," *Journal of Applied Learning & Teaching*, vol. 7(1), pp. 115–126, 2024.

[15] D. Weber-Wulff et al., "Testing of detection tools for AI-generated text," *International Journal for Educational Integrity*, vol. 19, article 26, 2023.

[16] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PLOS ONE*, vol. 14, no. 11, e0224365, 2019.

International Journal on Perceptive and Cognitive Computing (IJPCC) Vol 12, Issue 2 (2026)
https://doi.org/10.31436/ijpcc.v12i2.716

Machine Learning–Based Classification of Mobile Money Transactions

Mohamed Abdinur Hussein[1], Hassan Warsame Mohamed[2], Adamu Abubakar[3]
[1,2]Center for Postgraduate Studies, University of Bossaso – Garowe, Puntland, Somalia
[3]Department of Computer Science, International Islamic University Malaysia

Corresponding Author: adamu@iium.edu.my
(Received: 3th June 2026; Accepted: 4th July, 2026; Published on-line: 30th July, 2026)

Abstract— Mobile money has become the backbone of everyday payments across much of the Horn of Africa, and the transaction records it leaves behind are a valuable but under-used source of insight. This paper investigates whether the type of a mobile money transaction can be recovered automatically from its attributes, a capability that would support transaction monitoring and customer analytics. Working within the Knowledge Discovery in Databases framework and using the WEKA platform, we compared three well-established classifiers—Naïve Bayes, J48 (WEKA's implementation of the C4.5 decision tree), and Random Forest—on a real dataset of 3,731 transactions described by ten attributes and labelled with one of ten transaction types. All models were assessed under 10-fold stratified cross-validation. The two tree-based methods clearly outperformed the probabilistic baseline: J48 reached 76.01% accuracy with a Kappa of 0.6147 and the lowest mean absolute error (0.0615), while Random Forest was almost identical at 75.58% and produced the smallest root mean squared error (0.1821); the difference between the two tree-based models was not statistically distinguishable. Naïve Bayes, hampered by its independence assumption, managed only 51.09%. A per-class breakdown showed near-perfect recognition of common categories such as P2P receive (F-measure above 0.93) but a complete failure on the rarest classes, several of which scored zero. This imbalance effect is quantified by the macro-averaged F1 of just 0.40 and balanced accuracy of 38.02% for the best model, far below its 76.01% overall accuracy. This gap is evidence that headline accuracy is inflated by the dominant classes. On balance we recommend J48 for deployment under the present experimental setting, since it pairs competitive accuracy with fast training and human-readable rules that translate directly into monitoring logic.

Keywords— mobile money; machine learning; transaction classification; class imbalance data mining; classification; decision tree; random forest; naïve Bayes

I. INTRODUCTION

Mobile money services have formed the backbone of financial transactions in Puntland and throughout Somalia, where limited access to traditional banks has hastened the growth of digital financial platforms [1]. Individuals and businesses are rapidly using mobile money for daily financial transactions, including person-to-person transfers, merchant payments, salary disbursements, utility bill payments, overseas remittances, among others [2]. Mobile money is a way to get financial services quickly, easily and cheaply via a mobile device, even in the most distant places, and it allows financial inclusion, unlike traditional banking systems. As transaction volumes increase, service providers generate huge amounts of transactional data that mirror financial behaviour of clients, spending patterns, transaction frequency, payment channels and temporal aspects [3]. These digital records are significant information resources that may be studied to reveal hidden patterns of behaviour, boost operational efficiency, reinforce transaction monitoring, aid fraud detection and enhance decision-making through data-driven intelligence.

The fast rise of mobile money services has greatly increased the number and complexity of transaction data generated on a daily basis [4]. The rapidity of the transactions [5] and the heterogeneity of transaction types have made manual processing of these large-scale information increasingly unfeasible. Transaction classification is often inconsistent across different mobile money platforms due to inadequate records, human error, or operational discrepancies. Wrong or delayed classification decreases the effectiveness of transaction monitoring, financial reporting capabilities, and fraud detection methods [6]. As the transaction volumes are increasing, the financial institutions demand automated techniques to precisely identify the transaction types, being efficient and scalable.

https://doi.org/10.31436/ijpcc.v12i2.716

Although the use of machine learning in analysis of financial data is ubiquitous, the selection of a suitable classification algorithm is still a daunting task because the prediction power of different algorithms varies depending on the nature of the dataset [7]. Transactions in mobile money usually possess highly linked features, non-linear relationship and significant class imbalance where common transaction categories are dominant and crucial minority categories appear rarely. These factors diminish the performance of many standard classification methods and complicate model selection. Hence, it is necessary to identify which machine learning method offers the optimal trade-off between predictive accuracy, computational efficiency and model interpretability for mobile money transaction classification in the Puntland setting.

Correct classification of mobile money transactions is critical to provide secure, dependable and effective delivery of digital financial services. Proper classification of transactions helps service providers to improve customer service, prepare correct financial reports, monitor transaction behaviour, detect operational abnormalities and ensure regulatory compliance [8]. Furthermore, good classification leads to improved resource allocation, increased financial transparency, and greater customer confidence in digital monetary systems. As mobile money is replacing conventional banking in many developing countries, efficient transaction classification is becoming a crucial task to provide sustainable financial service provision and to safeguard the integrity of the digital payment ecosystem [9].

Machine learning can be very helpful in assessing massive volumes of transaction data, in that it can automatically find complicated linkages and behaviour patterns that are hard to spot using standard statistical or rule-based approaches. Unlike manual categorization, machine learning algorithms learn continually from past transaction records and identify unseen transactions with minimal human intervention. Decision tree and ensemble learning methods also provide interpretable decision rules that help financial analysts to analyze transaction behaviour. Therefore, machine learning has become an important technology to improve the accuracy of categorization, reduce the operating costs, speed up the transaction process, and enable the intelligent fraud detection of modern monetary systems.

Many studies have been conducted to apply machine learning methods to financial transaction analysis, fraud identification, customer behavior prediction and electronic payment systems. Naïve Bayes, decision tree models, Support Vector Machines, Artificial Neural Networks, Random Forest and Gradient Boosting classification methods have shown promising results on various financial datasets. However, the performance they describe differs significantly, as transaction datasets differ in size, feature qualities, class distributions and domain-specific behaviours. There are no simple generalizations of banking or credit card transaction studies to mobile money contexts, especially in developing nations where the nature of transactions is very different.

Furthermore, many of the prior works are mostly based on binary fraud detection instead of multi-class transaction classification. Others test algorithms on balanced data sets or generated data that is not representative enough of real-world operational contexts. There have been relatively few studies using real mobile money transaction data from Somalia or Puntland and it is not clear what classification technique would perform best given the highly imbalanced transaction distributions, whilst still generating interpretable decision rules which could be used in practice.

The study will compare the performance of Naïve Bayes, J48 Decision Tree and Random Forest classifiers in classifying mobile money transactions based on Knowledge Discovery in Databases (KDD) framework and identify the most accurate, efficient and interpretable model for knowledge discovery and transaction monitoring in the mobile money environment of Puntland.

This study makes several important contributions to the field of mobile money transaction analytics:

- It is one of the few empirical studies that assess machine learning classification approaches using real mobile money transaction data obtained from Puntland, Somalia.
- It applies whole Knowledge Discovery in Databases (KDD) process including data pre-processing, transformation, model creation, evaluation and knowledge deployment to mobile money transaction analysis.
- It offers a comparative analysis of three popular classification algorithms, namely Naïve Bayes, J48 Decision Tree, and Random Forest, under same experimental settings using 10-fold stratified cross-validation.
- It analyzes the performance of the classifiers by means of a number of complementing performance metrics such as classification accuracy, Kappa statistic, precision, recall, F-measure, Mean Absolute Error (MAE), Root Mean Squared Error (RMSE) and computational time.
- Results indicate that J48 Decision Tree offers the optimum blend of predictive performance, computational efficiency, and model interpretability. It produces transparent decision rules which can enable transaction monitoring, fraud detection, and operational decision making.

https://doi.org/10.31436/ijpcc.v12i2.716

- Finally, the results provide a practical baseline for future research on intelligent mobile money analytics, such as fraud detection, class imbalance management, and sophisticated ensemble learning methods within the expanding digital financial ecosystem.

The remaining sections of the paper are structured as follows: Section 2 presents the relevant work, section 3 outlines the research methods, section 4 details the study results, section 5 comprises the discussion, and section 6 concludes the paper.

II. RELATED WORK

The articles that were studied laid the groundwork for data mining and supervised categorization, both theoretically and in terms of methodology and empirical evidence. They lay out the steps involved in finding new information, provide the most popular classification algorithms, compare how well they forecast, and give statistical metrics for gauging how well two algorithms agree on a classification. Taken together, these research papers back the use of Naïve Bayes, J48, and Random Forest for classifying mobile money transactions, with the caveat that the success of the algorithms relies on the specifics of the dataset.

For quite some time, financial and behavioral data has been subject to classification. When it comes to algorithms and their evaluation, the traditional go-to sources are the standard texts by Han et al. [11] and Witten and Frank [12]. In particular, Quinlan's ID3 and its descendant C4.5 [13], with the latter being implemented in WEKA as J48, are the ancestors of decision trees. In order to create a collection of layered, understandable rules that naturally reflect the interactions between attributes, a tree continually splits the data on the attribute that decreases uncertainty the greatest. Expanding on this concept, Random Forest [14] typically reduces variation and increases stability by averaging the results of constructing several trees on bootstrap samples.

Han et al. [11] proposed a complete architecture of Knowledge Discovery in Databases which includes data cleansing, integration, transformation, data mining, evaluation, and display of knowledge. Using method explanations and application examples, the authors showed how classification, clustering, association analysis, and outlier identification extract useful patterns from enormous data sets. They found that both proper algorithms and thorough data preparation are needed for good knowledge discovery.

T. Witten et al. [12] took a pragmatic approach to experimental machine learning with the WEKA platform. The book provides real-world instances of data preparation, classification, feature selection, cross-validation, and interpretation of performance. Its major finding is that reliable data mining is not only a matter of algorithm choice, but also of good preprocessing, experimental design, validation, and careful interpretation of model outputs.

Quinlan [13] created C4.5 decision-tree approach as an extension of ID3 method. The method chooses splitting attributes, manages continuous and missing data, prunes overfitted branches and turns trees into classification rules. Experiments on different data sets demonstrated that C4.5 could deal successfully with huge instances with many attributes and to provide accurate, interpretable models. It was shown that decision tree models were efficient in performing practical classification.

Breiman [14] proposed Random Forest, an ensemble method that constructs numerous decision tree models from bootstrap samples and a random selection of predictors. Theoretical research shows that the generalization error of its ensemble converges with the number of trees and depends on the strength of the individual trees and their association. Experiments showed that Random Forest was competitive in accuracy, robust to overfitting, able to handle noise and offered important information about variable importance.

Domingos and Pazzani [15] studied both theoretically and empirically the reasons why the simple Bayesian classifier can perform well even when it violates conditional-independence assumptions. They studied zero-one loss and experimented with dependent qualities, showing that accurate probability estimate is not required as long as the proper class still has the highest probability. They showed that Naïve Bayes is best under more general conditions than attribute independence.

Caruana and Niculescu-Mizil [16] performed a comparison of ten supervised learning algorithms on numerous data sets using nine metrics of performance including accuracy, ROC area, precision-recall, squared error and cross-entropy. They found boosted trees, random forests, neural networks, and support vector machines to usually perform well, however the rankings differed by metric. The authors determined that the better model depended on both the data set and the evaluation criterion.

Fernández-Delgado et al. [17] assessed 179 classifiers from 17 method families on 121 real-world datasets. Overall, Random Forest variations performed best, support vector machines and neural networks also performed well. However, no classifier dominated all datasets. The study concludes that substantial empirical comparison is still required as classification success strongly depends on the properties of the application data.

Wolpert [18] compared learning methods with off-training-set error and mathematical analysis without any

https://doi.org/10.31436/ijpcc.v12i2.716

assumptions about the distribution of the problems. The results showed that, averaged uniformly over all conceivable learning tasks, gains of one approach are lost elsewhere. The study proved the No Free Lunch principle, which states that no classifier is better than any other for all possible problems without making assumptions about the target problem.

A statistical method for assessing the agreement between observers for multivariate categorical data was proposed by Landis and Koch [19]. Their technique extended Kappa statistics and agreement functions that account for agreement expected by chance to the observed agreement. An application to categorical reliability data illustrated the effectiveness of the method in assessing consistency in classification. They also offered interpretation ranges for Kappa values from little or poor agreement to practically perfect agreement.

Naive Bayes follows a different path, estimating class probabilities assuming characteristics are independent given the class. The approach can still perform well even when that assumption is broken, as Domingos and Pazzani showed notably [15], but its accuracy does deteriorate when attributes are highly connected. Transaction data is a common example when amount, balance, channel and direction move together. Tree-based and ensemble methods have been shown to be among the strongest general-purpose classifiers in broad empirical surveys [16], [17]. Wolpert's No Free Lunch theorem [18] tells us that no algorithm is the best for all datasets, and this is precisely why an empirical comparison is needed.

Work on transaction and fraud analytics often reports accuracy, precision, recall and the F-measure together with the Kappa statistic interpreted using the benchmarks of Landis and Koch [19], and error terms such as MAE and RMSE validated by cross-validation. One surface gap that keeps coming up is the class imbalance in multi-class transaction data, where the rare categories most relevant for fraud are also the hardest to learn. Our study directly addresses this gap by comparing three classifiers on a real data set from Puntland and studying in detail the effects of imbalance on the outcomes.

III. METHODOLOGY

In this study, we apply the Knowledge Discovery in Databases (KDD) process to build and test a machine learning model to detect fraudulent mobile money transactions. The methodology has successive phases including data collection, pre-processing, transformation, model building, evaluation and knowledge deployment (see Figure 1). Each phase contributes to increasing data quality, establishing robust prediction models, and discovering actionable knowledge to facilitate fraud detection and decision-making in mobile monetary systems.

Figure 1 shows the Knowledge Discovery in Databases (KDD) methodology employed in this study for mobile money transaction classification. The procedure starts with the gathering of a dataset containing 3,731 mobile money transactions. Data pre-processing then follows, such as removing extraneous features, handling missing values, and encoding nominal variables. The cleaned dataset is translated into WEKA ARFF format for analysis. Three classification algorithms including Naïve Bayes, J48 Decision Tree and Random Forest are constructed and tested using the 10-fold stratified cross-validation. Finally, the top performing model is deployed to extract decision rules and to identify critical predictive variables to facilitate transaction classification and knowledge discovery.

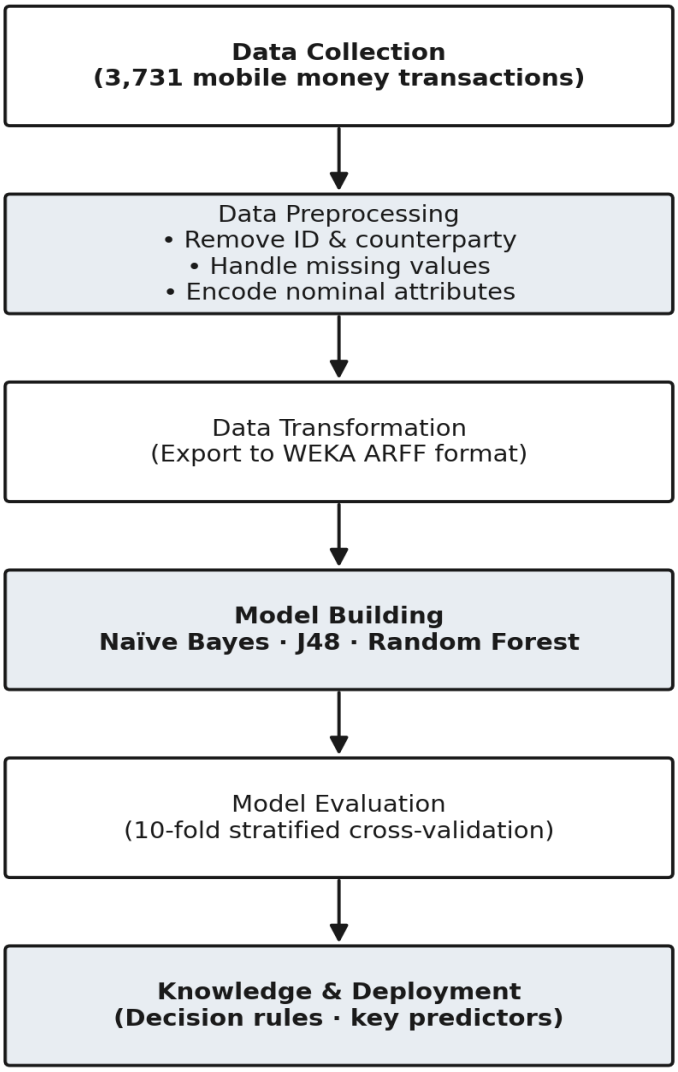


Fig. 1 KDD workflow followed in this study.

https://doi.org/10.31436/ijpcc.v12i2.716

A. Dataset

The data are from mobile money transaction logs gathered over approximately twelve months in a Puntland context. The raw file contains 3,731 transactions. After eliminating a distinctive transaction identification and a high-cardinality counterparty attribute—both of which lack generalizability—ten properties persist, one of which serves as the class label. The label, txn_type, is nominal and encompasses ten values, including merchant payment, two types of person-to-person transfers, overseas remittance, bill payment, and other less common categories. The predictor attributes delineate the transaction's direction, access channel, amount, amount-to-balance ratio, remaining balance, hour of the day, day of the week, whether the counterparty is new, and the time elapsed since the user's last transaction. Table I illustrates a significant skew in class distribution.

TABLE I

CLASS DISTRIBUTION OF TRANSACTION TYPES

Transaction Type	Count	Transaction Type	Count
Merchant payment	1,776	Bill payment	18
P2P send	1,323	Agent deposit	14
P2P receive	503	Reversal	6
Bank withdraw	64	Bank transfer	5
Intl. remittance	19	Airtime/data	3

B. Pre-processing and Attribute Selection

Preparation was carried out entirely in WEKA. The identifier and counterparty fields were removed with the Remove filter; the categorical attributes were encoded as nominal while the numeric ones were kept on their original scale; and the small number of missing entries—chiefly the time-since-previous field for a user's first transaction, together with a few unrecorded channel values—were handled with the ReplaceMissingValues filter where appropriate. The cleaned data were exported to ARFF for modelling. We retained all nine remaining predictors, since each captures a distinct facet of behaviour; an information-gain ranking confirmed that amount, the amount-to-balance ratio, direction, and channel carry the most discriminative power.

C. Experimental Design

Three classifiers were chosen to contrast a probabilistic method with a single tree and with an ensemble of trees. Naïve Bayes served as a fast-probabilistic baseline. J48, WEKA's implementation of C4.5, was run with its default settings (a confidence factor of 0.25 and a minimum of two instances per leaf). Random Forest was likewise left at its defaults, growing one hundred trees. Every model was evaluated using 10-fold stratified cross-validation, which repeatedly trains on nine folds and tests on the tenth while preserving the class proportions in each fold; this uses every record for testing exactly once and gives a more trustworthy estimate than a single train–test split. We recorded the count of correctly and incorrectly classified instances, overall accuracy, the Kappa statistic, precision, recall, the F-measure, MAE, RMSE, and the time taken to build each model.

Fig. 2 depicts the experimental pipeline: the prepared dataset is passed through 10-fold cross-validation to each of the three classifiers in parallel, their performance metrics are collected, and the results are then compared to identify the strongest model.

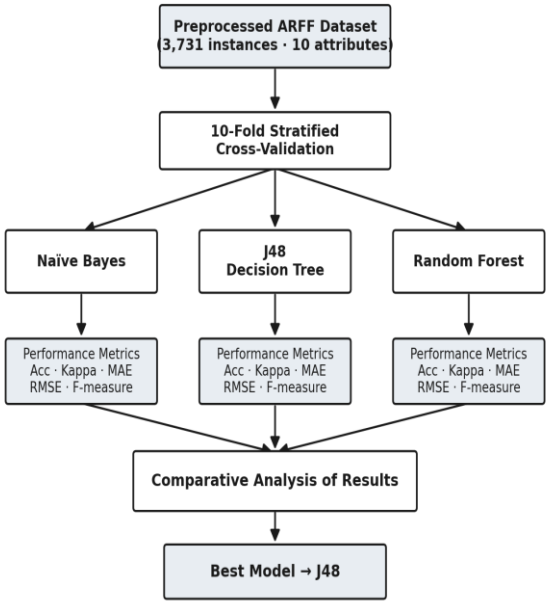


Fig. 2 Model training and comparison pipeline.

IV. RESULTS AND DISCUSSION

Table II summarises how the three classifiers performed across the headline metrics. The gap between the probabilistic baseline and the two tree-based methods is stark. Naïve Bayes classified just over half of the transactions correctly, whereas J48 and Random Forest were both close to three-quarters. J48 edged ahead on most measures the highest accuracy at 76.01%, the highest Kappa at 0.6147, and the lowest mean absolute error while Random Forest returned the smallest root mean squared error, hinting at marginally better calibrated probabilities. In build time the ordering reversed, with Naïve Bayes fastest and Random Forest the most expensive, though none of the models was slow in absolute terms.

https://doi.org/10.31436/ijpcc.v12i2.716

TABLE II

OVERALL PERFORMANCE OF THE THREE CLASSIFIERS

Metric	Naïve Bayes	J48	Random Forest
Correctly classified	1,906	2,836	2,820
Incorrectly classified	1,825	895	911
Accuracy (%)	51.09	76.01	75.58
Kappa	0.2117	0.6147	0.6091
MAE	0.1018	0.0615	0.0651
RMSE	0.2639	0.1935	0.1821
Build time (s)	0.01	0.10	0.57

Read against the Landis and Koch benchmarks [9], the Kappa values tell a clear story: both tree models land in the substantial-agreement band, while Naïve Bayes reaches only fair agreement, confirming that its apparent accuracy owes a good deal to the dominant classes rather than to genuine discrimination.

Averaged figures conceal an important weakness, so Table III reports the F-measure for each transaction type. The common categories were handled well—P2P receive reached an F-measure above 0.93 under both tree models, and merchant payment and P2P send were also recognised reliably. The rare categories were another matter entirely. Classes such as airtime/data, bank transfer, and reversal, each represented by only a handful of records, were effectively invisible to the tree models, which scored zero on them. This is the familiar consequence of class imbalance: with so few examples, there is simply not enough signal for the learner to form a rule.

TABLE III

CLASS-LEVEL F-MEASURE BY CLASSIFIER

Transaction Type	NB	J48	RF
Merchant payment	0.709	0.779	0.769
P2P send	0.244	0.691	0.688
P2P receive	0.243	0.941	0.939
Banks withdraw	0.167	0.455	0.478
Intl. remittance	0.160	0.750	0.788
Bill payment	0.163	0.211	0.095
Agent deposit	0.176	0.000	0.000
Reversal	0.092	0.000	0.000
Airtime/data, Bank transfer	0.000	0.000	0.000

The pattern is easy to explain once the assumptions behind each method are laid out. Naïve Bayes treats the predictors as independent given the class, but in mobile money data the amount, the balance, the channel, and the direction plainly move together, so the assumption is violated and accuracy pays the price [5]. J48 and Random Forest make no such assumption; they carve the feature space into nested regions and thereby capture the very interactions that trip up the Bayesian model. It is worth noting that ensemble theory would lead one to expect Random Forest to beat a single tree [4], yet here the two were separated by less than half a percentage point. That near-tie is not a contradiction so much as a reminder, in the spirit of the No Free Lunch theorem [8], that the best model is a property of the dataset, and on this dataset a single well-pruned tree was enough.

V. DISCUSSION

Mobile money transfers are essential in providing financial services to consumers and businesses with restricted access to conventional banking institutions. They facilitate users in transferring funds, receiving remittances, making merchant payments, purchasing airtime, settling bills, and managing daily financial transactions through their mobile devices. Furthermore, mobile money eliminates geographical, physical, and administrative barriers to financial participation in Somalia and other emerging countries [1], [2]. Its simplicity, cost-effectiveness, speed, and accessibility render it especially vital in rural and underprivileged communities. Robust mobile money platforms are crucial for promoting financial inclusion, enhancing business operations, improving household welfare, and fostering broader economic development.

The swift expansion of mobile money generates extensive transaction records that encompass valuable insights into client behavior, transaction frequency, payment methods, balances, transaction directions, and spending patterns. However, this data holds no use until it is continuously transformed into valuable knowledge. Knowledge Discovery in Databases provides a systematic methodology for cleansing, transforming, analyzing, evaluating, and interpreting transactional data [11]. Knowledge discovery techniques can be utilized to reveal concealed behavioral patterns, aberrant behaviors, trends in service consumption, operational vulnerabilities, and potentially fraudulent transactions. This intelligence assists providers in making superior, evidence-based decisions, enhancing financial reporting, segmenting clients, monitoring transactions, allocating funds, adhering to rules, managing risk, and more.

The increasing volume of transactions is unmanageable through human review, necessitating mobile money providers to seek efficient transaction analysis. Manual classification is labor-intensive, unreliable, prone to human error, and less appropriate for high-volume financial settings. Inaccurate transaction classification can affect financial reporting, client accounts, fraud investigations, and

https://doi.org/10.31436/ijpcc.v12i2.716

regulatory oversight. Automated analysis enables providers to differentiate among merchant payments, person-to-person transfers, international remittances, bank withdrawals, bill payments, deposits, reversals, and other transactions. It can also facilitate the swift identification of anomalous transaction behavior. Consequently, transaction analysis is essential for ensuring service reliability, protecting users, improving financial transparency, and maintaining trust in the digital payment system.

The significance lies in the fact that mobile money variables frequently exhibit complex, non-linear, and interdependent interactions that conventional manual or rule-based methods cannot adequately capture. Machine learning algorithms can identify transaction trends from past data, autonomously categorize new transactions, reduce processing expenses, and produce decision rules that improve monitoring, fraud detection, and operational decision-making [7], [8].

The experiment involved evaluating Naïve Bayes, J48, and Random Forest on 3,731 transactions utilizing 10-fold stratified cross-validation. Naïve Bayes accurately classified 1,906 transactions and erroneously categorized 1,825, resulting in an accuracy rate of 51.09%. The Kappa statistic was 0.2117, the Mean Absolute Error (MAE) was 0.1018, and the Root Mean Square Error (RMSE) was 0.2639. Nonetheless, it was constructed in under 0.01 seconds and exhibited poor accuracy and moderate concordance, signifying insufficient discriminatory power. Thus, computation speed alone is insufficient when the foundational model assumptions do not accurately reflect the linkages present in mobile money data.

J48 accurately identified 2,836 transactions and misclassified 895 transactions, with a maximum accuracy of 76.01%. The highest Kappa value recorded was 0.6147, the lowest MAE was 0.0615, the RMSE was 0.1935, and the model construction time was 0.10 seconds. The Kappa value signifies substantial agreement above random occurrence. The low MAE indicates relatively minor average prediction errors. The findings demonstrate that J48 successfully identified substantial correlations among the transaction amount, balance ratio, direction, and channel. The model features a discernible decision-tree framework, appropriate for operational analysis and practical implementation.

The Random Forest correctly detected 2820 transactions and misclassified 911. The precision was 75.58%. The Kappa statistic was 0.6091 and the Mean Absolute Error (MAE) was 0.0651. It exhibited the lowest RMSE of 0.1821. The model's construction time was 0.57 seconds, slower than Naïve Bayes and J48, however adequate for practical application. The low RMSE signifies that probability estimations are adequately calibrated. Nevertheless, it exhibited an accuracy that was 0.43 percentage points inferior to J48, indicating that the additional complexity of the ensemble did not provide a substantial classification advantage for this dataset.

The J48 and Random Forest models performed effectively for the prevalent transaction types at the class level. The F-measures for merchant payment were 0.709, 0.779, and 0.769 for Naïve Bayes, J48, and Random Forest, respectively. The respective figures for P2P are 0.244, 0.691, and 0.688. P2P reception recorded values of 0.243, 0.941, and 0.939. The bank withdrawal received scores of 0.167, 0.455, and 0.478. International remittance achieved scores of 0.160, 0.750, and 0.788. The findings indicate that tree-based models excel at recognizing prevalent and moderately represented transaction categories.

The performance of minority transaction types was markedly inferior. The F-measures for the bill payment dataset were 0.163, 0.211, and 0.095 for Naïve Bayes, J48, and Random Forest, respectively. The agent deposit was documented as 0.176 for Naïve Bayes and 0.000 for both tree-based models. The reversal achieved a score of 0.092 with Naïve Bayes, and 0.000 with both J48 and Random Forest. All three classifiers received a score of 0.000 for airtime/data and bank transfer. # These outcomes are due to the very low frequency of the classes, comprising merely 18 bill payments, 14 agent deposits, 6 reversals, 5 bank transfers, and 3 airtime/data transactions, which hampers the establishment of dependable patterns.The study demonstrates that mobile money transaction data may be converted into pertinent operational insights via machine learning. J48 provides the optimal balance of classification accuracy, computing efficiency, error minimization, and interpretability. The decision criteria are beneficial for automated transaction classification, customer behavior analysis, financial reporting, transaction monitoring, and preliminary fraud detection. In contexts where probability calibration is crucial, Random Forest proves advantageous, whereas Naïve Bayes serves merely as a first benchmark. The results indicate that overall accuracy should not be evaluated without considering class-level performance, as a model may demonstrate accuracy in most classes while failing to identify infrequent but operationally significant transactions.

The primary concern is the significant imbalance among the classes in the dataset. Merchant payments and P2P transfers recorded 1,776 and 1,323 transactions, respectively. Multiple minority classes contained fewer than twenty observations. The classifiers predominantly identified patterns associated with prevailing transaction types, but they continuously failed to recognize airtime/data, bank transfer, reversal, and agent-deposit transactions. The dataset was obtained in a specific Puntland mobile money context during around twelve months. Transaction patterns,

channel utilization, and class distributions may not accurately reflect different geographical regions, service providers, customer demographics, or temporal contexts.

Subsequent research should compile larger and more diverse datasets from various providers and locations. Enhancing minority classes can be accomplished via additional data acquisition, SMOTE, adaptive synthetic sampling, random undersampling, cost-sensitive learning, or class-weighted models. Moreover, researchers ought to evaluate Gradient Boosting, XGBoost, LightGBM, Support Vector Machines, and deep learning techniques. The models ought to be trained on prior transactions and subsequently evaluated on later periods. This is referred to as temporal validation. Additional behavioral markers, including transaction velocity, device alterations, geographical patterns, recurring counterparties, failed transactions, and account age, may improve classification. Future research should encompass the examination of explainability, concept drift, real-time implementation, and fraud-specific assessment utilizing precision-recall curves and class-sensitive performance metrics.

VI. Conclusions

We applied the KDD workflow to the problem of classifying mobile money transactions, comparing Naïve Bayes, J48, and Random Forest on 3,731 real records under 10-fold cross-validation. The tree-based methods were decisively stronger than the probabilistic baseline. J48 gave the best overall balance 76.01% accuracy, a Kappa of 0.6147, and the lowest MAE—while Random Forest matched it on accuracy and posted the lowest RMSE. Naïve Bayes, at 51.09%, was held back by an independence assumption that the data do not support. We therefore recommend J48 for deployment, not only for its accuracy but for the transparent rules it yields, which are of immediate use in transaction monitoring and fraud detection. The clearest limitation is class imbalance, which left every model unable to learn the rarest transaction types. The obvious next step is to address that imbalance directly through resampling methods such as SMOTE, cost-sensitive learning, or stronger gradient-boosted ensembles—so that the rare but consequential categories can be detected as well as the common ones

Acknowledgment

The author wishes to acknowledge the Center for Postgraduate Studies, University of Bossaso – Garowe, Puntland, Somalia for their support.

Conflict of Interest

The authors declare that there is no conflict of interest.

Authors Contribution Statement

Mohamed Abdinur Hussein: conceptualisation, data collection and pre-processing, WEKA experiments, formal analysis, writing of the original draft. Hassan Warsame Mohamed: methodology design, validation of results, interpretation of findings, review and editing of the manuscript. Adamu Abubakar supervise the work. All authors read and approved the final version.

Data Availability Statement

The dataset used for this study has been cited in the paper.

Ethics Statement

This study did not require ethical approval.

References

[1] P. Chonka, "Displacement stigma, insecurity and inequality in urban Somalia: The limits of digital financial empowerment," Contemporary Politics, pp. 1–21, 2025, doi: 10.1080/13569775.2025.2522448.

[2] T. Suri, J. Aker, C. Batista, M. Callen, T. Ghani, W. Jack, et al., "Mobile money," VoxDevLit, vol. 2, no. 2, p. 3, 2023. [Online]. Available: https://www.voxdev.org/sites/default/files/2023-09/Mobile_Money_2.pdf

[3] S. R. Baker and L. Kueng, "Household financial transaction data," Annual Review of Economics, vol. 14, no. 1, pp. 47–67, 2022, doi: 10.1146/annurev-economics-051520-023646.

[4] D. Namakhwa, B. H. M. Tchereni, W. Masanjala, C. D. Namakhwa, S. L. Kuchande, and W. R. Mgomezulu, "Modelling and forecasting mobile money customer transaction volumes in rural and semi-urban Malawi: An autoregressive integrated moving average spatial decomposition," Scientific African, vol. 26, Art. no. e02430, 2024, doi: 10.1016/j.sciaf.2024.e02430.

[5] M. R. Miah and M. M. Islam, "Data preprocessing and feature engineering strategies for large-scale predictive modeling applications," Review of Applied Science and Technology, vol. 3, no. 1, pp. 263–302, 2024, doi: 10.63125/tqqqed47.

[6] A. Agarwal and V. A. Assenova, "Mobile money as a stepping stone to financial inclusion: How digital multisided platforms fill institutional voids," Organization Science, vol. 35, no. 3, pp. 769–787, 2024, doi: 10.1287/orsc.2022.16562.

[7] H. Gao, G. Kou, H. Liang, H. Zhang, X. Chao, C. C. Li, and Y. Dong, "Machine learning in business and finance: A literature review and research opportunities," Financial Innovation, vol. 10, no. 1, Art. no. 86, 2024, doi: 10.1186/s40854-024-00629-z.

[8] N. Hanbali and A. El-Yahyaoui, "Advanced machine learning and deep learning approaches for fraud detection in mobile money transactions," Innovations in Systems and Software Engineering, vol. 21, no. 2, pp. 333–353, 2025, doi: 10.1007/s11334-025-00605-5.

[9] J. Fan, L. K. Shar, R. Zhang, Z. Liu, W. Yang, D. Niyato, and K. Y. Lam, "Deep learning approaches for anti-money laundering on mobile transactions: Review, framework, and directions," IEEE Internet of Things Journal, early access, 2026, doi: 10.1109/JIOT.2026.3653434.

[10] A. Ali, S. Abd Razak, S. H. Othman, T. A. E. Eisa, A. Al-Dhaqm, M. Nasser, et al., "Financial fraud detection based on machine learning: A systematic literature review," Applied Sciences, vol. 12, no. 19, Art. no. 9637, 2022, doi: 10.3390/app12199637.

[11] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques, 3rd ed. Waltham, MA: Morgan Kaufmann, 2012.

[12] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, Data Mining: Practical Machine Learning Tools and Techniques, 4th ed. Cambridge, MA: Morgan Kaufmann, 2016.

[13] J. R. Quinlan, C4.5: Programs for Machine Learning. San Mateo, CA: Morgan Kaufmann, 1993.
[14] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[15] P. Domingos and M. Pazzani, "On the optimality of the simple Bayesian classifier under zero-one loss," Machine Learning, vol. 29, no. 2–3, pp. 103–130, 1997.
[16] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in Proc. 23rd Int. Conf. Machine Learning, 2006, pp. 161–168.
[17] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" J. Machine Learning Research, vol. 15, no. 1, pp. 3133–3181, 2014.
[18] D. H. Wolpert, "The lack of a priori distinctions between learning algorithms," Neural Computation, vol. 8, no. 7, pp. 1341–1390, 1996.
[19] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," Biometrics, vol. 33, no. 1, pp. 159–174, 1977.

International Journal on Perceptive and Cognitive Computing (IJPCC) Vol 12, Issue 2 (2026)
https://doi.org/10.31436/ijpcc.v12i2.715

A Human-Centred Framework for Ethical Data Visualization and Storytelling

Madihah Sheikh Abdul Aziz
Dept of Information Systems, Kulliyyah of Information and Communication Technology, International Islamic University Malaysia, 53100, Gombak, Selangor, Malaysia

Corresponding Author: madihahs@iium.edu.my
(Received: 13th May 2026; Accepted: 7th June, 2026; Published on-line: 30th July, 2026)

Abstract— As data visualization becomes central to decision-making across domains, educators must prepare students not only to master technical tools, but also to communicate data ethically and effectively. This paper presents a scaffolded instructional approach for teaching data storytelling in an undergraduate information visualization course. Students progressed from designing static infographic posters to building interactive dashboards using the same real-world dataset, guided by structured feedback, user experience (UX) principles, and reflective practice. The course design was informed by a faith-driven educational philosophy, embedding values from Tawhidic Epistemology (TE), such as intention (niyyah), responsibility (amanah), and reflective consultation (shura), into each stage of the learning process. Students worked in groups, selected socially relevant datasets (e.g., mental health, environment), and applied design principles grounded in cognitive theory and user empathy. Outcomes were measured through comparative grade analysis across two cohorts (before and after implementing the structured model), as well as qualitative reflections. Results indicate a marked improvement in academic performance and ethical awareness, with the later cohort achieving a higher proportion of A grades and reporting increased clarity, purpose, and confidence in their designs. This paper contributes a transferable pedagogical model that bridges technical training, design ethics, and values-based learning in data visualization education.

Keywords— Data Visualization Education, User Experience (UX) Design, Tawhidic Epistemology (TE), Scaffolded Learning, Data Storytelling.

I. Introduction

The increasing ubiquity of data in everyday life demands that graduates not only analyze information but also communicate it meaningfully and ethically. In the context of ICT education, teaching data visualization involves more than mastering tools such as Tableau; it also requires cultivating an understanding of audience, clarity, cognitive load, and design ethics. However, many visualization courses still emphasize tool usage and chart diversity, with less focus on narrative design, ethical awareness, or user-centered thinking [2], [3].

This paper addresses that gap by presenting a scaffolded instructional model for teaching data storytelling in an undergraduate visualization course. Students began with a data-driven infographic poster assignment, then expanded their work into an interactive dashboard, using the same dataset across both tasks. This approach emphasized progressive complexity, structured feedback, and iterative learning, aligning with UX and cognitive design principles [4]–[5], [10].

In addition, the course was framed within a faith-driven educational paradigm, inspired by the principles of Tawhidic Epistemology (TE). Rather than being treated as an abstract philosophy, TE was integrated into the design process through practices such as weekly student-led tazkirah (spiritual reflections), ethical dataset selection, peer consultation (shura), and reflective journaling (muhasabah). These values acted as ethical anchors, encouraging students to approach visualization as a trust (amanah) and to consider the broader impact of their designs on audiences.

This study aims to 1)Design and implement a scaffolded instructional model informed by Tawhidic Epistemology (TE) for teaching ethical data storytelling in an undergraduate Information Visualization course, 2)Evaluate the impact of the proposed model on students' academic performance, ethical awareness, and confidence in visualization design, 3)Examine students' reflections and project artefacts for evidence of value-based thinking, visual communication skills, and user-centred design practices.

Accordingly, this study addresses the following research questions:

- **RQ1:** How does a scaffolded, TE-integrated instructional model influence students' performance in data storytelling assignments?

https://doi.org/10.31436/ijpcc.v12i2.715

- **RQ2:** What ethical and user experience (UX) design values do students develop through the proposed instructional approach?
- **RQ3:** How do students perceive their growth in ethical awareness, visual communication, and narrative intentionality throughout the learning process?

This paper reports on the implementation of this model across two student cohorts (2024 vs. 2025), using both quantitative and qualitative measures to evaluate its impact. The findings suggest that structured, feedback-rich, and value-anchored instruction can significantly enhance student performance, ethical reasoning, and design confidence. The paper concludes by offering a transferable teaching model for visualization educators seeking to balance technical instruction with ethical and epistemological depth.

The remainder of this paper is organized as follows. Section 2 outlines the background and problem statement, establishing the context and rationale for the study. Section 3 describes the course design, materials, and instructional strategies implemented. Section 4 details the methodology used to evaluate the approach. Section 5 presents the results and analysis, focusing on student performance, reflections, and project outcomes. Finally, Section 6 offers a synthesis of key findings and implications for the implementation of a scaffolded, faith-informed model in undergraduate visualization education. Throughout, the paper has been structured to clearly articulate the problem addressed, contextualize the pedagogical intervention, and present its objectives in a focused and coherent manner.

II. Related Work

A. Ethical Dimensions of Data Visualization and Storytelling

Ethical considerations in data visualization have emerged as a central research theme over the last decade. Correll [1] was among the first to systematically highlight how choices in visualization design, scales, aggregation, and annotation are not neutral but carry normative implications. Building on this foundation, Bach, Correll, and Kaur [2] identify emerging practices in ethical data storytelling, including transparency in data provenance, audience-centered framing, and the need to surface uncertainty rather than obscure it.

Recent work has advanced this agenda along three dimensions: trust, representation, and pedagogy. Elhamdadi et al. [6] introduced Vistrust, a multidimensional framework that dissects trust in visualizations into cognitive and affective components, showing that even subtle design variations can shift users' confidence in the underlying data. Similarly, Hepworth and Church [7] warn that visualizations embed "arguments" about social phenomena, and failures to account for racial and cultural context can reproduce harm. Al-Hazwani et al. [8] extend this discussion through data humanism, advocating for design practices that deliberately foreground imperfection, emotion, and context to avoid the illusion of objectivity.

Educational perspectives reinforce the importance of early ethical training. Baynes et al. [9] present a case study of embedding ethics modules in introductory visualization courses, while recent efforts to revise teaching codes (2024) emphasize fairness, transparency, and context, moving beyond the narrow concern of "lying with charts." Together, these contributions mark a shift from treating ethics as an add-on to positioning it as a core design principle in visualization pedagogy.

B. UX and Design Learning in Visualization Education

Understanding the perceptual and cognitive underpinnings of visualization design remains foundational. Ware's [10] Information Visualization: Perception for Design continues to serve as a reference for mapping visual encoding to human perception. Yet, the conversation has moved toward embedding these perceptual insights into user experience (UX) frameworks tailored for learning. Nouri, Zhang, and Rapp [11] propose a UX design framework for digital learning environments that aligns usability, motivation, and inclusivity, demonstrating the value of holistic design in education.

Recent research highlights scaffolded and technology-mediated UX design. Menekse et al [12] show how real-time scaffolds can improve reflection quality and academic performance in large lecture courses, while Chaseley and Abercrombie [13] explore how scaffold types (descriptive vs critical prompts) influence reflective depth. Similarly, Liu and Wang [14] find that scaffolded reflective learning enhances both design performance and metacognitive growth in student teachers. From a technology lens, Børte et al. [15] emphasize aligning pedagogy with digital tool design, echoing Gkintoni et al. [16], who integrate AI and neuroscience insights to challenge traditional cognitive load assumptions in learning environments.

These findings suggest that UX in educational visualization must extend beyond interface usability to encompass reflective scaffolds, AI-driven personalization, and the cognitive demands of learners.

C. Faith-Based and Tawhidic Epistemology (TE) Pedagogy

Faith-based perspectives contribute a unique epistemological grounding to visualization education. Ibrahim and Osman [17] propose a conceptual model for integrating Islamic values into computing education, aligning learning processes with ethical virtues such as amanah (trustworthiness) and shura (consultation). Aljunied [18] frames this within the broader discourse of

Islamisation of knowledge, urging Islamic universities to harmonize empirical reasoning with Qur'anic principles.

Recent scholarship reflects increasing concern with digital transformation in Islamic pedagogy. Yunita [19] explores how educational technologies can support Islamic pedagogy by fostering justice and compassion, while Zahraini et al. [20] highlight the challenges of Islamic curriculum reform in the digital era. Other works [21, 22] illustrate the application of digital resources in Islamic religious education, showing both opportunities for innovation and risks of value dilution.

Together, these studies underscore that TE-aligned pedagogy provides more than cultural context: it establishes an ethical compass for visualization design, ensuring that narratives remain trustworthy, just, and spiritually grounded. This integration resonates with the call in visualization ethics for transparency and responsibility, offering a uniquely Islamic framing of the "why" behind ethical design.

D. Reflective and Scaffolded Learning for Visualization Pedagogy

Reflection is widely recognized as a driver of deep learning [23]. In visualization education, reflection helps students critique not only their design outputs but also their underlying assumptions and narrative framing. Recent empirical studies have sharpened our understanding of how scaffolding enhances reflective practice. Liu and Wang [14] demonstrate that scaffolded reflection improves both design performance and reflective thinking, while Rook et al. [24] report that students desire varied scaffolds and clear framing when reflection is integrated into higher education.

At the intersection of pedagogy and technology, new approaches leverage AI-mediated scaffolds. Yuan and Hu [25] speculate on how generative AI could serve as a reflective companion, delivering personalized feedback at scale. These developments align with the shift toward digital learning ecosystems where reflection is not an isolated activity but a guided, interactive process.

Scaffolded reflection also overlaps with ethical visualization practices. By prompting students to consider questions of representation, bias, and trust, scaffolds can bridge technical design skills with ethical awareness. This connection suggests that reflective pedagogy is not only about metacognition but also about cultivating professional responsibility.

E. Data Analysis as a Precursor to Visualization

A crucial but sometimes underemphasized component of visualization education is data analysis. Visualization is most effective when grounded in robust data wrangling, descriptive statistics, and exploratory analysis [26], [27]. Regression, clustering, and forecasting techniques not only generate insights but also shape visualization choices and narrative framing.

An equally critical step is the choice of chart type, which serves as the bridge between analysis and visual communication. Chart selection is not arbitrary; it must align with the analytical task and the audience's interpretive needs. For instance, trends over time are best represented by line charts, distributions by histograms, comparisons by bar charts, and relationships by scatterplots [27]. Ware [10] emphasizes that perceptual principles such as clarity, salience, and cognitive efficiency determine whether a chart supports or hinders understanding.

From an educational standpoint, teaching chart choice cultivates both analytical literacy and visual reasoning. Students are challenged not only to analyze data but also to design visuals that truthfully represent distributions, comparisons, or correlations. Neri et al. [29] confirm in their review of data visualization in AI-assisted decision-making that the credibility of dashboards often hinges on whether the chosen charts faithfully represent the data's structure and uncertainty. Poor chart choice not only reduces interpretability but also raises ethical concerns by distorting meaning or amplifying bias.

Finally, the ethical dimension of visualization often originates in the analysis and charting stages. Okorie et al. [28] note that biases in sampling, cleaning, and modeling directly influence fairness and credibility. If these biases are compounded by inappropriate chart choice, the result can be highly misleading. Teaching students to interrogate both what the data says and how it is shown ensures accountability across the full pipeline, from raw analysis to final storytelling.

F. Implications for Visualization Education

The reviewed literature highlights that visualization education must evolve beyond tool proficiency toward a multidimensional model that equips students with both technical expertise and ethical responsibility. Five interconnected implications stand out as discussed below:

1) Data Analysis and Chart Choice as Core Competence: Visualization education must emphasize not only statistical literacy but also the ability to map analytical tasks to appropriate chart types. Choosing the right chart, line for trends, bar for comparisons, scatter for relationships, pie for part-to-whole, ensures that insights are communicated clearly and truthfully [27], [10]. Students who understand this analysis-to-visualization pipeline are better prepared to create visuals that are both informative and trustworthy.

2) Visual Design Principles as Foundational Literacy: At the heart of effective visualization lies perceptual alignment. Ware [10] underscores that clarity, salience, and hierarchy reduce cognitive load, while Gkintoni et al. [16] show how

poor design can overwhelm learners. Embedding design principles ensures students move beyond attractive visuals toward perceptually efficient and cognitively accessible representations.

3) Ethics as a Central Learning Outcome: Correll [1] and Bach et al. [2] emphasize that visualizations are never neutral, they carry arguments. Ethical education in visualization must therefore cover transparency, fairness, uncertainty, and context. Students should learn to anticipate ethical risks, from misleading encodings to biased data sources, and develop habits of responsible design.

4) UX and Reflective Scaffolding as Pedagogical Drivers: Research on scaffolded reflection [14], [12] demonstrates that structured support enhances both design quality and critical judgment. Incorporating UX frameworks [11] ensures that learning environments are inclusive and user-centered. Together, scaffolding and UX make visualization education more reflective, interactive, and adaptive to diverse learners.

5) Faith-Based and TE-Aligned Values as Ethical Anchors: In Islamic and values-driven contexts, visualization is also a moral and spiritual act. TE-aligned pedagogy [17] situates visualization within the virtues of amanah (trust), shura (consultation), and ihsan (excellence). This approach complements global ethical discourses by embedding visualization practice within cultural and spiritual integrity.

Together, these five pillars, data analysis and chart choice, visual design principles, ethics, UX and reflection, and TE values, form the foundation of a holistic model of visualization education. Such a model prepares graduates not only to use tools but also to design with clarity, empathy, and integrity in an era where visualization is central to decision-making, storytelling, and public trust.

III. COURSE DESIGN AND IMPLEMENTATION

This section outlines the instructional framework for the course, with a focus on the pedagogical strategies and philosophical principles that guided the design and delivery of the course. The approach centers on scaffolded learning, integration of TE, UX principles, and reflective teaching methods.

A. Course Overview and Learning Outcomes

The study was conducted in the undergraduate course for Information Visualization, offered at the Kulliyyah of ICT, International Islamic University Malaysia (IIUM). It aims to equip students with both technical and cognitive skills in designing effective, purposeful data visualizations. The course is a core requirement for final-year students from multiple ICT specializations, including Business Intelligence, Information Security and Assurance, and Digital Media Design, Enterprise Management and Technology, and general specialization. While most students had prior exposure to Human-Computer Interaction (HCI) concepts through a prerequisite course, they had no prior experience in data visualization tools such as Tableau, nor in information or communication design theory.

The Course Learning Outcomes (CLOs) included, 1)Describe the underlying principles, techniques, and practices of the information visualization for effective decision making, 2)Demonstrate the ability to use a structured design process and an industrial tool to create effective visualizations, 3)Develop an interactive visualization system based on the storytelling principles to design coherent and insightful visualizations.

A total of 45 students were involved in the 2025 cohort, split across two sections (classes). The diversity of specializations, combined with their inexperience in visualization tools, made this course an ideal setting for implementing a scaffolded, ethically grounded instructional model.

B. Teaching Methodology: Scaffolded Project-Based Learning

The instructional strategy adopted a scaffolded learning model, where students progressed from a data-driven infographic poster assignment to an interactive dashboard project, using the same real-world dataset throughout. This approach enabled students to develop a strong conceptual understanding of data storytelling before translating it into interactive, functional formats.

Each stage of the project was supported by structured consultations, peer critiques, and iterative feedback loops that emphasized design clarity, ethical communication, and user orientation. The dataset selection was left to the students, who were guided to choose socially relevant, ethically safe topics such as mental health awareness, blood donation trends, food insecurity, or environmental sustainability.

C. Integration of Tawhidic Epistemology (TE)

The course was designed in alignment with the TE, which emphasizes the integration of faith, ethics, and knowledge. This philosophical foundation was embedded both explicitly and implicitly throughout the course delivery and assessments.

Key values included:

1) Niyyah (intentionality): Encouraging students to begin with ethical and purposeful design intentions.

2) Amanah (trust): Emphasizing the responsibility of data representation.

3) Ihsan (excellence): Striving for clarity, integrity, and balance in visual communication.

https://doi.org/10.31436/ijpcc.v12i2.715

4) Shura (consultation): Practiced through class critique sessions and peer feedback.

5) Muhasabah (self-reflection): Embedded in the post-project reflections submitted by students.

Weekly 5-minute student-led tazkirah sessions were also integrated to strengthen spiritual engagement and ethical awareness.

D. UX, Visual Cognition, and Design Principles

The instructional content emphasized a human-centered, design-first approach rooted in principles from user experience (UX) and cognitive design theory. Students were introduced to:

1) Visual hierarchy and pre-attentive attributes (e.g., color, shape, size, alignment)
2) Gestalt principles (e.g., proximity, similarity, closure)
3) Story-first vs. chart-first design strategies
4) Accessibility guidelines for colour and layout (e.g., users with colour-blindness)
5) Ethical considerations in representing marginalized data or sensitive issues

These concepts were integrated through lectures, critique activities, and visual storytelling tasks that mirrored real-world scenarios.

E. Assignment and Project Structure

Both tasks were group-based, and together contributed 30% of the students' total coursework marks (Infographic: 10%, Dashboard: 20%). The rubrics assessed data analysis, and interpretation, storytelling clarity, visual hierarchy, ethical alignment, UX considerations, and presentation.

TABLE 1:
THE FLOW OF ASSIGNMENT AND PROJECT STRUCTURE

Flow	Task	Description	Learning Outcome
1	infographic poster	static visual narrative using real-world dataset	understand basic visual encoding, structure, and narrative clarity
2	feedback & consultation	instructor-led critiques and peer feedback	develop critical thinking and reflective revision
3	interactive dashboard (tableau)	functional, multi-screen dashboard based on the same dataset	apply UX and storytelling principles in interactive design
4	presentation & reflection	final walkthrough of dashboard and student reflection	communicate design choices, receive live critique, and reflect on process

F. Tools and Instructional Support

To support the scaffolded learning process, students were equipped with a range of accessible tools. For data preparation and basic cleaning, most groups utilized Google Sheets or Microsoft Excel. For visual storytelling, students developed their infographic posters using Canva or, in some cases, Adobe Illustrator, enabling them to experiment with layout, visual hierarchy, and colour schemes. The interactive dashboards were built using Tableau, chosen for its intuitive drag-and-drop interface, industry relevance and strong support from its global user community.

Instructional support was delivered through multiple channels. The course included guided tutorials and hands-on activities to build foundational skills. Throughout the semester, students participated in in-class UX critiques and visual construction exercises, allowing them to refine their ideas collaboratively. The instructor provided formative, iterative feedback during consultation slots, reinforcing design intent, ethical framing, and alignment with storytelling objectives. Additionally, both the infographic assignment and the final dashboard project were evaluated using rubric-based assessments, providing structured expectations and reinforcing key learning outcomes.

G. Cohort Comparison and Iteration

This model was iteratively improved based on feedback from the 2024 cohort, who followed a more open-ended, less scaffolded project structure. The 2025 cohort received clearer milestones, integrated feedback points, and stronger ethical framing. This study compares outcomes between both cohorts to evaluate the pedagogical effectiveness of the structured model.

IV. METHODOLOGY

This study adopts a design-based research (DBR) approach, integrating pedagogical innovation with iterative evaluation in a real-world classroom context. The aim is to explore how a scaffolded, faith-informed instructional model anchored in the Tawhidic paradigm and UX principles affects student engagement, ethical awareness, and visual communication performance in an undergraduate Information Visualization course.

A. Research Design and Approach

A mixed-methods design was employed, combining quantitative data (e.g., assessment performance, grade distributions) with qualitative insights (e.g., open-ended reflections, student feedback, artifact analysis). The study design draws from constructivist learning theory, emphasizing the role of structure, reflection, and ethical intention in the development of data storytelling skills.

https://doi.org/10.31436/ijpcc.v12i2.715

A pre-test and post-test structure was implemented to compare students' conceptual understanding before and after completing the course. Additionally, project artifacts and student reflections were collected to triangulate learning outcomes and assess deeper impacts beyond grades.

B. Participants and Course Context

The study involved Level 3 and final-year undergraduate students enrolled in the Information Visualization course across two different cohorts, Cohort 1 (2024, Semester 1) as the pre-intervention and Cohort 2 (2025, Semester 2) as the post-intervention.

Both cohorts were taught by the same instructor and followed the same syllabus structure. Students came from diverse ICT specializations, as mentioned in Section 3, but shared the same prerequisites, including a prior course in Human-Computer Interaction (HCI). Importantly, none had prior experience in Tableau or formal visualization training, as confirmed through a pre-test survey conducted in the beginning of the semester.

C. Learning Model and Interventions

The five-phase scaffolded model was designed to support students' growth from ideation to ethical communication, integrating modern UX principles with faith-based values. Each phase was aligned with practical tools and corresponding TE principles. Students in the 2025 cohort received a redesigned, scaffolded model with the features as depicted in the Figure 1.

The scaffolded structure emphasized intentionality in design, iterative feedback, and reflection, both cognitive and spiritual. Weekly 5-minute tazkirah sessions, led by students, provided a rhythm for ethical anchoring and peer connection. The 2024 cohort, in contrast, used the same assessment tasks but received minimal scaffolding, no formal design critique sessions, and no embedded reflective components.

The scaffolded learning model comprises five interconnected phases that progressively develop students' technical competence, ethical awareness, and reflective practice:

- Plan: Students identify a real-world problem, select an appropriate dataset, and define the intended audience while embedding niyyah (purposeful intention) and amanah (responsibility) in the design process.
- Design: Students transform the dataset into an infographic poster by applying principles of visual hierarchy, colour, typography, and storytelling, reflecting ihsan (excellence) and 'aql (intellectual reasoning).
- Critique: Students present their work for instructor consultation and peer review, practising shura (consultation), constructive feedback, and iterative refinement.
- Build: Students extend the infographic into an interactive Tableau dashboard, integrating usability, interaction, and analytical exploration while applying hikmah (wisdom) and tadabbur (thoughtful reflection).
- Reflect: Students complete reflective activities and participate in weekly student-led tazkirah, encouraging muhasabah (self-reflection) and reinforcing ethical responsibility in communicating data.

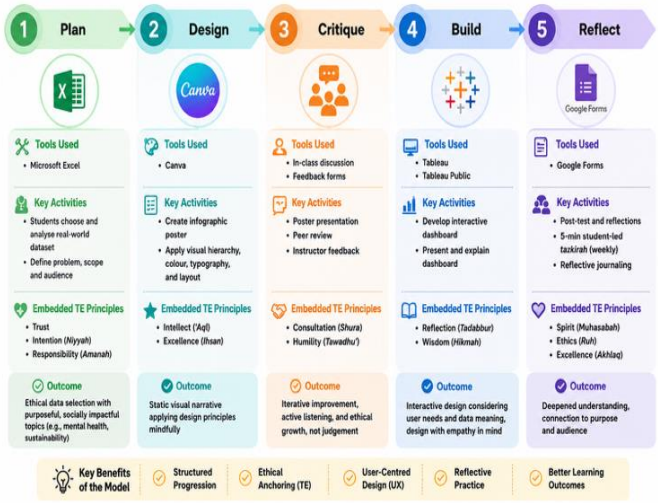


Fig. 1: Scaffolded Learning Model with Ethical Anchors from Tawhidic Epistemology (TE)

D. Instruments and Data Collection

To assess the impact of the revised model, data was collected using the following instruments in Table 2.

TABLE 2:
INSTRUMENTS USED IN THE STUDY

Instrument	Purpose	Data Type
Pre-test survey	Establish baseline for visualization and tool knowledge	Quantitative
Post-test survey	Assess understanding, confidence, ethical awareness, and students reflections (to capture student perception and personal growth)	Quantitative Qualitative
Final grade records	Measure academic performance across cohorts	Quantitative
Assignment artifacts	Evaluate design quality, storytelling clarity, UX and presentation	Qualitative (rubric-based)

https://doi.org/10.31436/ijpcc.v12i2.715

V. RESULTS AND DISCUSSION

A. Data Analysis

1) Quantitative Analysis

The grade distribution comparison was conducted across both cohorts using the percentage of students achieving each final grade category as a descriptive indicator of academic performance. As shown in Figure 2, the 2025 cohort recorded a higher proportion of A grades (85%) compared with the 2024 cohort (55%), while the proportion of students obtaining A– and B+ or below decreased from 24% to 12% and from 21% to 3%, respectively. These descriptive findings suggest that the scaffolded, feedback-rich instructional model was associated with improved learning outcomes and more consistent student performance. The observed improvements align with previous studies highlighting the value of structured feedback [3], iterative design practice [30], and student agency [31] in promoting deeper learning and design quality. The embedded TE values may have further amplified student motivation and ethical awareness, encouraging purposeful work and deeper engagement.

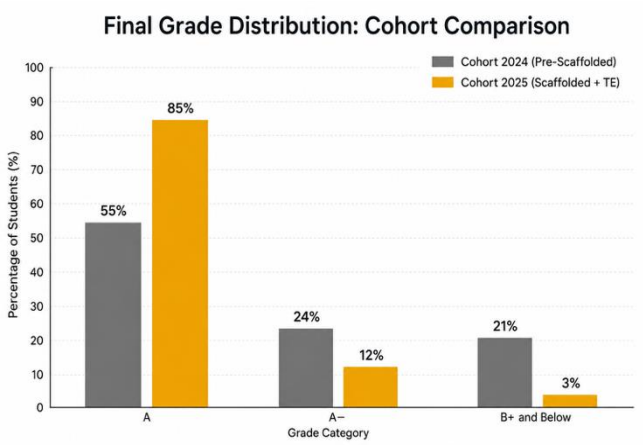


Fig. 2: Comparison of final grade distributions between the 2024 and 2025 cohorts. The scaffolded TE-integrated approach was associated with a higher proportion of A grades and fewer students in the B+ and below category.

2) Qualitative Analysis

The open-ended responses (students' reflections) were coded thematically using inductive analysis. Categories included: learning clarity, ethical decision-making, user empathy, confidence in storytelling, and value reflection (e.g., niyyah, amanah). Sample quotes were extracted to demonstrate students' evolving mindset regarding purpose-driven design and responsibility in communicating data.

B. Validity Considerations

Both cohorts used the same instructor, syllabus, and grading rubric, reducing instructional bias. While pre/post tests were anonymous, baseline analysis confirmed all students began with zero prior exposure to Tableau or information visualization. Although small in size, the study focused on depth of reflection and impact, typical of design-based research. The inclusion of student-led tazkirah and values reflection was not treated as a religious evaluation but as a pedagogical strategy to embed intentionality and ethics, consistent with TE-based education [31]-[32].

C. Thematic Insights from Student Reflections

Open-ended responses gathered from both sections revealed strong qualitative support for the scaffolded, reflective learning model. Five dominant themes emerged: intentional design, confidence growth, peer empathy, UX awareness, and faith-driven reflection.

1) Many students expressed increased confidence through hands-on practice:

> "Tableau made it easy for me by drag and drop I could design without needing to code."
> "I work better by doing. The Tableau project helped me explore data and tell the story I want."

2) Others highlighted the value of critique and teamwork:

> "Critique sessions helped me see my blind spots and improve with confidence."
> "We learn from the feedback, how to be better, not just what's wrong."

3) Several reflections revealed awareness of design as ethical communication, not just visual creativity:

> "The most challenging part was making sure the dashboard is clean and not misleading."
> "It's hard to balance between beauty and usefulness."
> "Now I know how colours and layout can change what users feel."

4) Students also recognized their growth in applying UX and design principles intentionally:

> "I learned how to reduce clutter and highlight what matters to the user."
> "Visual hierarchy is so important, I didn't realize it before this class."

D. Impact of TE-Embedded Reflection

Beyond technical competence, the integration of Tawhidic Epistemology (TE) encouraged students to reflect on the ethical and social responsibilities associated with

https://doi.org/10.31436/ijpcc.v12i2.715

data visualization. Analysis of the students' reflective responses revealed four key outcomes:

- Purposeful design through niyyah: Students became more intentional in selecting socially relevant topics and designing visualizations that conveyed meaningful messages.
- Ethical responsibility through amanah: Students recognised that visualizations should communicate information truthfully, clearly, and without misleading the intended audience.
- Continuous self-improvement through muhasabah: Reflection activities encouraged students to evaluate their own design decisions and identify opportunities for refinement.
- Values reinforcement through weekly tazkirah: Short student-led reflections strengthened awareness of ethical communication and connected technical work with broader personal and professional responsibilities.

These findings are illustrated through representative student reflections:

> "The tazkirah helped me think of design as part of responsibility and ethics."
>
> "I chose a topic that aligns with my values, because I believe in doing what matters."
>
> "Designing with niyyah made me rethink how I communicate visually."
>
> "This course reminded me that data is not just numbers, it's trust."

These reflections support TE emphasis on integrating 'aql (intellect) and ruh (spirituality) in the pursuit of knowledge. By embedding Islamic values into the learning process, not through overt preaching, but through reflective design, the course facilitated not only technical growth but also moral and spiritual development.

E. Representative Students' Projects

The following student projects were selected to illustrate how the scaffolded, TE-embedded instructional model fostered ethical, purposeful, and visually effective data storytelling. Each group was required to conceptualize a real-world problem, design a data-driven infographic poster, and later extend it into an interactive Tableau dashboard. Weekly critique sessions, peer feedback, and reflective prompts were embedded throughout the project timeline.

1) Project 1 – Plastic Waste and Environmental Impact in Malaysia

This group focused on plastic pollution in Malaysia by integrating diverse datasets from local environmental agencies (CAP, Ministry of Housing) to academic sources (NCBI). Their infographic poster provided a well-structured visual narrative showing the lifecycle of plastic waste, recycling challenges, and microplastic effects on marine life. The message-first layout incorporated iconography, simplified graphs, and color harmony, conveying urgency without overwhelming the viewer.

Their Tableau dashboard extended the narrative with interactive components:

- A forecasting chart on recycling rates from 2020–2030 using polynomial regression
- A geospatial map visualizing countries exporting plastic waste to Malaysia
- Treemaps and donut charts showing composition of waste types and environmental impact

The team used annotations, guided titles, and clear legends to maintain readability. Through this, they demonstrated mastery of ethical framing (amanah), visual hierarchy (ihsan), and intentionality (niyyah). Their topic choice and design decisions reflected both data literacy and spiritual responsibility.

2) Project 2 – Public Transport Ridership Trends in Klang Valley (2020–2024)

This group investigated shifts in public transport usage across MRT, LRT, BRT, and KTM systems in Klang Valley. Their infographic highlighted yearly ridership trends, weekday vs. weekend differences, and the effect of government initiatives. Strategic colour use and visual storytelling created a narrative flow titled "From Monday Rush to Sunday Calm."

The Tableau dashboard evolved into a robust, multi-layered system featuring:

- Time-series ridership trends annotated with public holidays and MRT expansion phases
- A calendar-based view mapping spikes in travel to national events
- Fare structure comparisons using stacked bars and custom tooltips

> "We didn't just visualize numbers, we started seeing patterns in society."
>
> "The dashboard answered a real-world question for us."

This project showcased the ability to move beyond descriptive charts into interpretive storytelling. They exemplified hikmah (wisdom) in visual synthesis, shura (consultation) through weekly feedback loops, and muhasabah (reflection) in their ethical consideration of design bias.

Both projects illustrate the pedagogical value of combining structured scaffolding with reflective and ethical framing. Students not only developed technical proficiency

https://doi.org/10.31436/ijpcc.v12i2.715

but demonstrated sensitivity toward audience, ethics, and narrative integrity. The outcomes support the central claim of this study that scaffolded, TE-informed instruction cultivates holistic data storytellers ready for real-world impact. Figures 3–5 present representative examples of students' progression from static infographic posters to interactive Tableau dashboards.

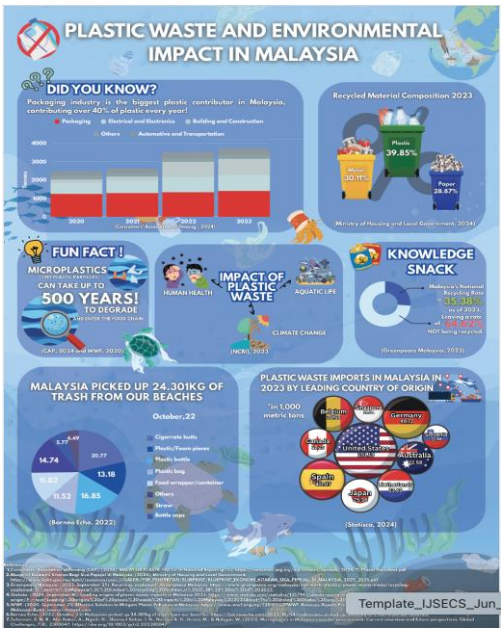
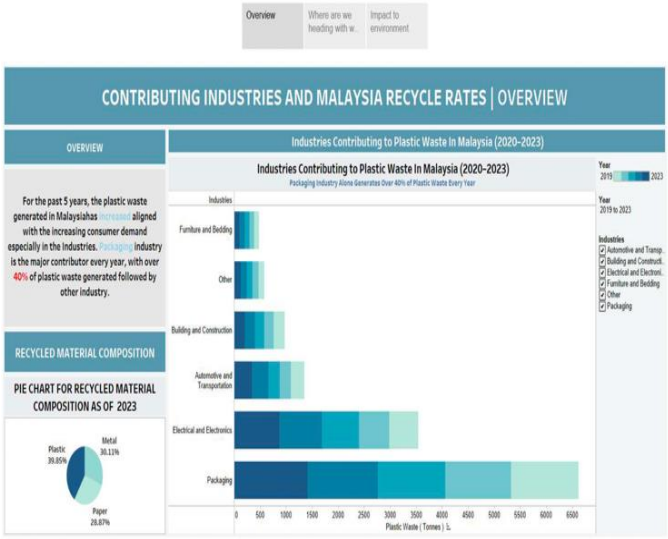
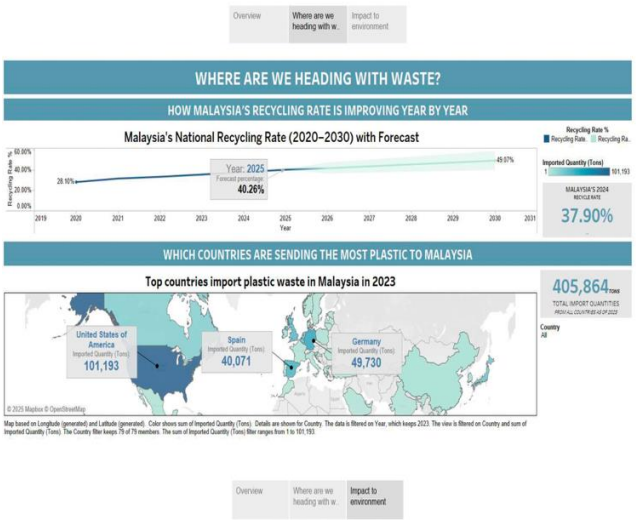


Fig. 3: Scaffolded progression of the plastic-waste project 1, from a static infographic poster to selected interactive Tableau dashboard views using the same dataset.





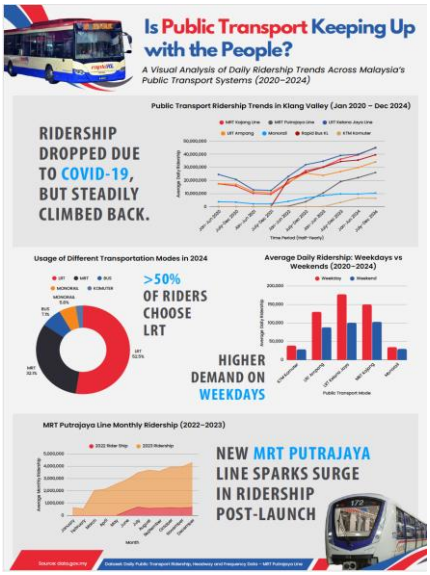
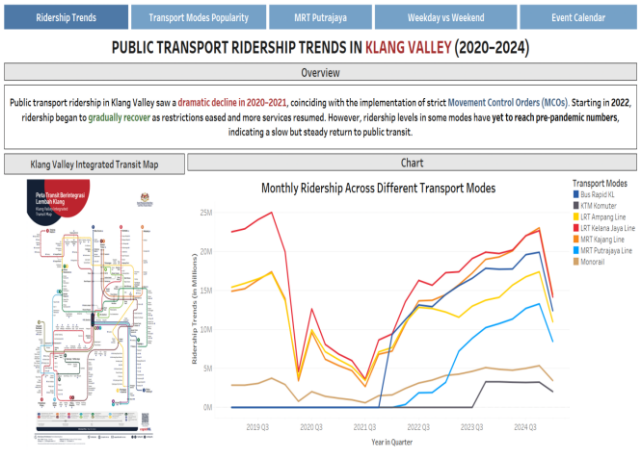
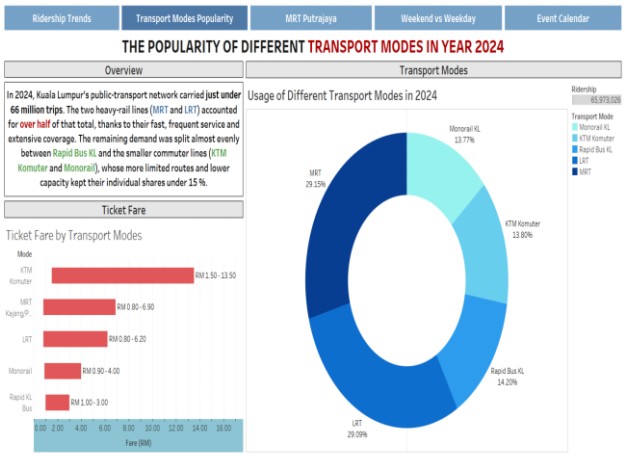


Fig. 4: Scaffolded progression of the public-transport project 2, from a static infographic poster to selected interactive Tableau dashboard views using the same dataset.

https://doi.org/10.31436/ijpcc.v12i2.715





Together, the paired artefacts illustrate how students retained continuity in topic, dataset, and narrative purpose while progressing from static visual communication to interactive analytical exploration. The transition demonstrates the development of visual hierarchy, user-centred interaction, data storytelling, and ethical design judgement through the scaffolded learning process.

VI. Limitations and Lessons Learned

While the outcomes of the scaffolded, TE-integrated teaching model are encouraging, several limitations must be acknowledged. Firstly, the study did not employ a matched pre-test/post-test design, which limits the ability to conduct rigorous statistical analysis of individual learning gains. Although all students reported no prior knowledge of Tableau or data storytelling concepts, this claim was self-reported and not validated through standardized measurement instruments.

Secondly, the final grade distribution used for quantitative comparison reflects students' overall course performance, which includes assessments beyond the infographic and dashboard project. As such, isolating the direct impact of the scaffolded design process on those specific assessments requires a more focused grading structure in future iterations.

Additionally, the study was conducted in a single-instructor, single-course context within one institution, involving a relatively small number of students across two cohorts. These factors may limit the generalizability of the findings to broader educational settings. Despite this, the structured approach supported by shared rubrics, datasets, and reflective activities provides a transparent and replicable model that can inform similar interventions in other UX, HCI, or visualization-related courses.

VII. Summary of Insights

The findings from both quantitative and qualitative analyses offer strong support for the effectiveness of the scaffolded, TE-informed instructional model. Across both sections, students demonstrated clear gains in performance, design maturity, and ethical awareness. This section synthesizes key observations in relation to the study's three research questions.

- RQ1: How does a scaffolded, TE-integrated instructional model affect students' performance in data storytelling assignments?

The comparative grade distribution across cohorts revealed that the 2025 cohort who engaged in the scaffolded assignment-to-dashboard model achieved higher proportions of A grades, with substantially fewer lower-grade outcomes than the previous cohort. This contrasts with the previous semester where similar project outcomes showed more grade variability. These results suggest that the structured and reflective approach not only enhanced design quality but also supported consistency in learning outcomes.

- RQ2: What ethical and UX design values do students develop through this approach?

Students consistently applied UX and visual design principles such as clarity, hierarchy, alignment, and narrative

sequencing in both infographics and dashboards. Moreover, qualitative responses and final project artifacts reflected an awareness of amanah (trust in representing data truthfully), niyyah (purposeful storytelling), and ihsan (excellence in design). Weekly tazkirah sessions and critique activities supported ethical reflection and the translation of Islamic values into visual communication practice.

- RQ3: How do students perceive their growth in ethical awareness, visual communication, and narrative intentionality?

Post-course reflections and quotes revealed increased confidence in decision-making, heightened awareness of design ethics, and appreciation for the connection between intention and audience impact. Students expressed that they no longer saw data visualization as "just charts," but as a form of communication with meaning and consequence. The integration of faith-based values with technical content deepened their engagement and sense of accountability in design work.

In summary, the model fostered technical proficiency, ethical literacy, and narrative empathy attributes critical for future-ready, faith-conscious digital storytellers.

In addition to project outcomes and post-course reflections, the Student Feedback on Teaching and Learning (SFS) data provided further validation of the course's effectiveness. Across both sections, students rated the instructor's teaching effectiveness between 95.96% and 97.69%, and overall course satisfaction between 94.34% and 97.36%. Notably high scores were given in areas such as critical thinking development, Islamic integration, constructive feedback, and respect for students. Written feedback praised the structured, step-by-step delivery, clarity of instruction, and engaging classroom environment. As one student remarked:

"Arguably the best course I've ever enrolled in. The productive classes made me not want to skip any sessions."

These results strengthen the qualitative findings and suggest that the scaffolded, TE-informed instructional model not only enhanced performance and ethical design awareness but also contributed to a highly positive and motivating learning experience.

VIII. CONCLUSIONS AND FUTURE WORK

This study presents a structured and ethically anchored teaching model for data storytelling, implemented in an undergraduate Information Visualization course. Rooted in Tawhidic Epistemology and guided by scaffolded learning principles, the model moved students through a five-phase progression from dataset selection to visual reflection framing visualization not merely as a technical output but as a trust-bearing act of communication.

The integration of spiritual reflection (via weekly tazkirah), guided design tools (Canva and Tableau), and iterative feedback (UX) led to improvements in student engagement, academic performance, and ethical awareness. Students not only demonstrated design competence but developed a deeper sense of niyyah (intention) and amanah (responsibility) in how they represented and narrated data. These findings echo emerging perspectives on holistic education [23], [31]-[32], where technical mastery is balanced with spiritual, ethical, and emotional intelligence.

For future iterations, improvements may include matched pre- and post-assessments, rubric-linked tracking of narrative and visual clarity, and the expansion of this model to postgraduate or interdisciplinary contexts. Additionally, future research may explore the longitudinal impact of this approach on students' ethical reasoning, professional decision-making, and design mindset beyond the classroom.

Ultimately, this study contributes a practical and transferable framework for integrating scaffolded learning, user experience principles, and Tawhidic Epistemology into data visualization education. By demonstrating how ethical reflection can be embedded alongside technical skill development, the proposed model extends conventional visualization pedagogy beyond tool proficiency towards purposeful, human-centred, and responsible data storytelling. The framework may serve as a useful reference for educators seeking to cultivate technically competent, ethically grounded, and socially responsible future data storytellers.

CONFLICT OF INTEREST

The author declares that there is no conflict of interest regarding the research, authorship, and publication of this article.

AUTHOR CONTRIBUTION STATEMENT

Madihah Sheikh Abdul Aziz solely conceived and designed the study, developed the instructional framework, conducted the course implementation, collected and analysed the data, interpreted the findings, prepared the figures, and wrote, reviewed, and revised the manuscript.

DATA AVAILABILITY STATEMENT

The data supporting the findings of this study consist of anonymized student assessment records, reflective responses, and project artefacts collected during the Information Visualization course. Due to institutional and ethical considerations, these data are not publicly available but may be made available by the corresponding author upon reasonable request, subject to institutional approval.

ETHICS STATEMENT

This study was conducted within the context of normal teaching and learning activities in an undergraduate Information Visualization course. Student assessment records, reflections, and project artefacts were anonymized prior to analysis, and no personally identifiable information is reported. The study complies with the ethical principles governing educational research at the authors' institution.

ACKNOWLEDGMENT

The author sincerely thanks the reviewers and editors of the International Journal on Perceptive and Cognitive Computing for their constructive comments and valuable suggestions, which significantly improved the quality of this manuscript. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

[1] M. Correll, "Ethical dimensions of visualization research," in Proc. 2019 CHI Conf. Human Factors Comput. Syst., pp. 1–13, 2019. doi: 10.1145/3290605.3300411.

[2] B. Bach, M. Correll, and K. Kaur, "Emerging practices in ethical data storytelling," IEEE Trans. Vis. Comput. Graph., vol. 29, no. 1, pp. 1423–1437, Jan. 2023. doi: 10.1109/TVCG.2022.3150889.

[3] J. Roberts, C. Headleand, and P. Ritsos, "Teaching data visualization: Promoting critical thinking and creativity," IEEE Comput. Graph. Appl., vol. 42, no. 3, pp. 89–98, 2022. doi: 10.1109/MCG.2022.3153350.

[4] S. Krug, Don't Make Me Think, Revisited: A Common Sense Approach to Web Usability, 3rd ed. Berkeley, CA, USA: New Riders, 2014.

[5] R. E. Mayer, Cognitive Theory of Multimedia Learning, 2nd ed. Cambridge, U.K.: Cambridge Univ. Press, 2005.

[6] H. Elhamdadi, A. Stefkovics, J. Beyer, E. Moerth, C. X. Bearfield, and C. Nobre, "Vistrust: A multidimensional framework and empirical study of trust in data visualizations," arXiv preprint arXiv:2309.16915, 2023.

[7] A. Hepworth and A. Church, "Racism in the machine: Visualization ethics in digital humanities projects," arXiv preprint arXiv:2411.17704, 2024.

[8] M. Al-Hazwani, Y. Zhang, L. Garrison, and J. Bernard, "Data humanism decoded: A characterization of its principles to bridge data visualization researchers and practitioners," arXiv preprint arXiv:2509.00440, 2025.

[9] A. Baynes, L. Busby, M. Correll, and A. Johnson, "Case study on including ethics into introductory data visualization," in Proc. IEEE VIS, 2023. [Online]. Available: https://www.researchgate.net/publication/367208307

[10] C. Ware, Information Visualization: Perception for Design, 4th ed. San Francisco, CA, USA: Morgan Kaufmann, 2020.

[11] J. Nouri, J. Zhang, and C. Rapp, "A UX design framework for digital learning environments," Educ. Technol. Res. Dev., vol. 70, pp. 921–945, 2022. doi: 10.1007/s11423-021-10039-z.

[12] M. Menekse, J. Li, and E. Brown, "A quasi-experimental study in a large lecture course: Real-time, technology-enhanced scaffolding effects on reflection quality and academic performance," Learn. Instr., vol. 83, 101799, 2025. doi: 10.1016/j.learninstruc.2025.101799.

[13] T. L. Chaseley and S. Abercrombie, "Using scaffolds to support preservice teachers' reflective practice," Front. Educ., vol. 10, 1621269, 2025. doi: 10.3389/feduc.2025.1621269.

[14] Q. Liu and H. Wang, "Exploring the effects of scaffolded reflective learning on student teachers' design performance and reflective thinking," Learn. Instr., vol. 86, 101844, 2025. doi: 10.1016/j.learninstruc.2024.101844.

[15] K. Børte, I. J. Ness, and D. G. Lorentzen, "Learning to teach: Aligning pedagogy and technology in a digital learning design tool," Comput. Educ., vol. 215, 104895, 2024. doi: 10.1016/j.compedu.2024.104895.

[16] E. Gkintoni, C. Tsichouridis, and M. Darra, "Challenging cognitive load theory: The role of AI, educational neuroscience, and machine learning," Front. Psychol., vol. 16, 1185272, 2025. doi: 10.3389/fpsyg.2025.11852728.

[17] R. Ibrahim and N. H. Osman, "Integrating Islamic values into computing education: A conceptual model," J. Islamic Ethical Comput., vol. 1, no. 1, pp. 45–58, 2023.

[18] S. M. K. Aljunied, "Islamisation of knowledge and the future of Islamic universities," Islamic Quarterly, vol. 65, no. 2, pp. 139–158, 2021.

[19] N. Yunita, "Towards Islamic pedagogy by exploring the applications of educational technology," J. Islamic Educ. Technol., vol. 3, no. 1, pp. 45–59, 2024.

[20] N. Zahraini, I. Akib, D. Rosidin, and R. Sulaeman, "Islamic education reform in the digital age: Challenges and opportunities for a modern curriculum," Int. J. Islamic Educ. Stud., vol. 4, no. 1, pp. 55–72, 2025.

[21] S. Munawati, "Innovation in Islamic religious education for elementary students—Using digital educational resources," Int. J. Educ. Learn., vol. 6, no. 2, pp. 112–120, 2024. [Online]. Available: https://files.eric.ed.gov/fulltext/EJ1460268.pdf

[22] M. Sanusi, "Transforming Islamic education in the digital age," Attractive: Innovative Education Journal, vol. 6, no. 2, pp. 1479, 2024. [Online]. Available: https://attractivejournal.com/index.php/aj/article/view/1479

[23] S. D. Brookfield, Becoming a Critically Reflective Teacher, 2nd ed. San Francisco, CA, USA: Jossey-Bass, 2021.

[24] L. Rook, S. White, and S. Callaghan, "Student feedback for integrating reflection into higher education," High. Educ. Res. Dev., vol. 44, no. 2, 2548274, 2025. doi: 10.1080/07294360.2025.2548274.

[25] L. Yuan and X. Hu, "Generative AI as a tool for enhancing reflective learning in students," arXiv preprint arXiv:2412.02603, 2024.

[26] S. Few, Now You See It: Simple Visualization Techniques for Quantitative Analysis, 2nd ed. Burlingame, CA, USA: Analytics Press, 2021.

[27] T. Munzner, Visualization Analysis and Design. Boca Raton, FL, USA: CRC Press, 2014.

[28] V. O. Okorie, C. A. Ezeh, and C. Ogbu, "Ethical considerations in data collection and analysis: A review investigating ethical practices and challenges in modern data collection and analysis," Int. J. Ethics Inf. Technol., vol. 10, no. 1, pp. 1–14, 2024.

[29] G. Neri, S. Marshall, H. K. Chan, A. Yaghi, D. Tabor, R. Sinha, and S. Mazumdar, "Data visualization in AI-assisted decision-making: A systematic review," Front. Commun., 2025. [Online]. Available: https://eprints.whiterose.ac.uk/id/eprint/231013/1/fcomm-2-1605655.pdf

[30] S. Sedig, P. Parsons, and A. Babanski, "Exploring visual literacy for complex and dynamic data: A case for information visualization education," Inf. Vis., vol. 11, no. 4, pp. 283–293, 2012. doi: 10.1177/1473871612455743.

[31] A. Mostefai, "Student agency in visualization pedagogy: Balancing structure and freedom," Int. J. Educ. Technol. High. Educ., vol. 22, no. 3, pp. 223–238, 2025.

[32] S. M. K. Aljunied, "Islamisation of knowledge: Contemporary directions," J. Islamic Thought & Civiliz., vol. 14, no. 1, pp. 44–61, 2022.

International Journal on Perceptive and Cognitive Computing (IJPCC) Vol 12, Issue 2 (2026)
https://doi.org/10.31436/ijpcc.v12i2.705

Debouncing-Based Automated Network Failover System with Multi-Channel Real-Time Notifications

Febrian Wahyu Christanto[1*], Samuel Vittorio Rivaldo[1], Firdaus Laia[2], Eryan Ahmad Firdaus[3]
[1]Department of Information Technology, Universitas Semarang, Semarang, Indonesia
[2]Department of Computer Science, Universitas Nias Raya, South Nias, Indonesia
[3]Department of Informatics, Indonesia Defense University, Bogor, Indonesia

*Corresponding Author: febrian.wahyu.christanto@usm.ac.id
(Received: 10th May 2026; Accepted: 17th June, 2026; Published on-line: 30th July, 2026)

Abstract— Internet connectivity failures in single-ISP environments may interrupt business operations, increase network recovery time, and reduce service availability. This research proposes a debouncing-based automated network failover system integrated with multi-channel real-time notifications to improve failover stability and operational responsiveness in small-to-medium business environments. The proposed system was implemented on a MikroTik RouterOS platform by integrating Netwatch, Scheduler, automated routing scripts, Telegram Bot API, and email notification services. Unlike conventional failover implementations that rely on single connectivity checks, the proposed approach incorporates a debouncing-based verification mechanism and edge-triggered notification logic to prevent false failover decisions, reduce route flapping caused by transient network disturbances, and minimize redundant notification delivery. The system development followed the PPDIOO (Prepare, Plan, Design, Implement, Operate, and Optimize) methodology to ensure a structured implementation and evaluation process. Experimental evaluation was conducted under real operational conditions by measuring failover recovery time, notification delay, failover stability, and client-side connectivity recovery during ISP transitions. The proposed system achieved an average failover recovery time of 4.2 seconds, while Telegram and email notifications recorded average delivery delays of 2.22 seconds and 8.61 seconds, respectively. Furthermore, connection tracking optimization improved effective client-side recovery by reducing session persistence issues following route switching. The implementation successfully reduced administrator intervention during network failures while improving service continuity, monitoring efficiency, and operational reliability. These results demonstrate that integrating debouncing-based failover control with edge-triggered multi-channel notifications provides an effective and practical solution for enhancing network availability and resilience in operational business networks.

Keywords—Failover System, Network Monitoring, MikroTik, Debouncing Mechanism, Real-time Notification

I. INTRODUCTION

Reliable internet connectivity has become an essential requirement for modern business operations, particularly for organizations that depend on digital transactions, cloud-based services, and real-time communication. In Indonesia, internet penetration continues to increase significantly, reaching more than 80% of the national population in recent years [1], [2]. This growing dependence on internet infrastructure also increases the operational impact of network downtime, especially in business environments with limited redundancy mechanisms [3], [4]. Kampoeng Semarang, a souvenir and handicraft business center located in Semarang, Indonesia, relies heavily on internet connectivity to support daily operational activities. The existing network infrastructure utilizes a single Internet Service Provider (ISP) with a bandwidth capacity of 150 Mbps serving approximately 30 active users per day. However, the absence of an automated backup connection mechanism causes operational disruptions whenever the primary ISP experiences connectivity failure. Field observations indicated that network interruptions occurred between 20 and 40 times per month, with manual recovery processes requiring approximately 5 to 20 minutes depending on administrator response time and ISP handling.

In this research, network monitoring refers to the continuous observation of internet connectivity and network operational status to determine whether failover actions should be executed. The monitoring mechanism utilizes ICMP reachability tests implemented through MikroTik Netwatch, while repeated verification is performed using Scheduler to distinguish temporary packet loss from actual link failures. Operational monitoring parameters include host reachability, packet response status, failover recovery time, notification delay, and client-side connectivity recovery after ISP transitions.

However, practical failover implementation still faces several operational challenges. In real-world networking

environments, transient packet loss, unstable latency, and intermittent ISP connectivity may trigger false failover events if monitoring decisions rely solely on single connectivity checks [5], [6], [7]. Such conditions can produce route flapping behavior, where the network repeatedly switches between primary and backup links within short periods. Frequent route transitions not only degrade user experience but may also introduce session persistence problems, connection tracking inconsistencies, and delayed client-side recovery. Furthermore, repeated notifications generated during unstable conditions may reduce administrator responsiveness and increase operational noise. Therefore, failover systems require not only rapid switching capability but also stable decision-making mechanisms capable of distinguishing temporary disturbances from actual link failures [8], [9], [10].

To address these limitations, this research proposes a debouncing-based automated network failover system integrated with multi-channel real-time notifications. The proposed system introduces a debouncing mechanism using repeated verification through MikroTik Scheduler to prevent unstable failover decisions caused by temporary packet loss or intermittent connectivity fluctuations. In addition, an edge-triggered notification strategy is implemented to ensure that notifications are only transmitted during actual state transitions between UP and DOWN conditions, thereby minimizing redundant alerts [11]. The system was developed using the PPDIOO (Prepare, Plan, Design, Implement, Operate, and Optimize) framework to ensure structured deployment, operational validation, and optimization processes [12], [13], [14]. The proposed architecture integrates MikroTik Netwatch, automated routing scripts, Telegram Bot API, and email notification services to provide stable failover behavior and real-time operational awareness.

This research contributes to the field of network reliability and operational monitoring in three main aspects: (1) the implementation of a debouncing-based anti-flapping failover mechanism, (2) the integration of edge-triggered multi-channel notification delivery, and (3) operational evaluation in a real-world small-to-medium business environment. Experimental results demonstrate improvements in failover consistency, recovery responsiveness, and client-side connection stability during ISP transitions.

II. Related Work

Several researches have implemented failover mechanisms and network monitoring systems to improve network availability and reduce downtime caused by ISP failures. Bustomi and Tahir [15] implemented a MikroTik-based failover system using a GSM backup connection to maintain service continuity during primary ISP disruptions. Similarly, Putra et. al. [16] applied a Netwatch-based failover mechanism to automatically switch traffic to a backup ISP when connectivity problems occurred.

Recent researches have emphasized that increasing internet dependency requires network infrastructures capable of maintaining high availability and service continuity under dynamic operational conditions. Howell [4] discussed internet ecosystem resilience, while Elsayed and Abohany [7] highlighted the importance of resilient enterprise network architectures to reduce operational disruption during connectivity failures. These studies underline the necessity of automated recovery mechanisms for maintaining business continuity. Previous studies have also investigated intelligent failover decision mechanisms. Anbalagan [8] proposed proactive failover automation for mission-critical systems, whereas Ryu et al. [9] introduced distributed failover controllers with embedded monitoring capabilities. Although these studies improve failover reliability, they do not specifically address route flapping caused by transient packet loss in MikroTik-based operational environments.

Other researches focused on monitoring and notification systems to improve administrator responsiveness. Zikra et. al. [17] developed a MikroTik-based monitoring system integrated with Telegram and email notifications to accelerate network issue detection. Vingestin et. al. [18] and Nurjanah et. al. [19] also utilized Telegram Bot integration for real-time monitoring notifications and demonstrated that Telegram-based alerts could improve operational awareness during network disturbances. Artificial Intelligence has also been explored to improve high-availability systems. Fotia et al. [10] reviewed AI-based approaches for improving system availability and fault tolerance. Nevertheless, most existing approaches focus on large-scale infrastructures and require more sophisticated architectures than those commonly deployed in small-to-medium business environments. Although previous researches successfully improved network monitoring and failover automation, most implementations still rely on single connectivity checks for failover decisions. This approach may cause route flapping during transient network instability and generate repetitive notifications under unstable conditions. In addition, previous works mainly focused on successful route switching without evaluating effective client-side recovery after ISP transitions. Notification effectiveness has also become an important aspect of operational monitoring systems. Li et al. [11] demonstrated that notification relevance significantly affects administrator responsiveness, indicating that reducing redundant alerts is essential for improving operational awareness.

https://doi.org/10.31436/ijpcc.v12i2.705

TABLE I
COMPARISON OF PREVIOUS RESEARCHES AND PROPOSED SYSTEM

R	F	M	A	C	O
Hoesiin and Yuliandi [6]	✓	✗	✓	✗	✗
Anbalagan [8]	✓	✗	✓	✗	✗
Ryu et. al. [9]	✓	✗	✓	✗	✗
Bustomi and Tahir [15]	✓	✗	✗	✗	✗
Putra et. al. [16]	✗	✓	✗	✗	✗
Zikra et. al. [17]	✓	✗	✗	✗	✗
Vingestin et. al. [18]	✗	✓	✗	✗	✗
Nurjanah et. al. [19]	✗	✓	✗	✗	✗
Proposed System	✓	✓	✓	✓	✓

Legend:
R: Research
F: Failover
M: Multi-channel Notification
A: Anti-Flapping/ Debouncing
C: Client-side Recovery Analysis
O: Operational Evaluation

Based on the comparison presented in Table 1, this research proposes a debouncing-based automated network failover system integrated with edge-triggered multi-channel notifications. The proposed system applies repeated verification using MikroTik Scheduler to improve failover stability and reduce false failover events caused by temporary disturbances. Furthermore, edge-triggered notifications are implemented to minimize redundant alerts, while client-side recovery behavior is evaluated to improve operational reliability in real-world business environments.

III. METHODOLOGY

This research employed a methodology focused on the design, implementation, and evaluation of an automated network failover system integrated with multi-channel real-time notifications. The proposed system was implemented in the operational network environment of Kampoeng Semarang to improve network availability and reduce downtime caused by ISP connectivity failures.

A. Research Framework

The research adopted the PPDIOO (Prepare, Plan, Design, Implement, Operate, and Optimize) framework to ensure a structured system development and evaluation process as shown in Figure 1.

The Prepare phase focused on identifying operational problems related to network downtime and connectivity instability in the existing infrastructure. The Plan phase involved determining system requirements, selecting supporting technologies, and preparing failover testing scenarios. During the Design phase, the network failover architecture, monitoring mechanism, and notification workflow were developed. The system design integrated MikroTik Netwatch, Scheduler, automated routing scripts, Telegram Bot API, and email notification services. Figure 2 shows the failover logic topology used in this research. The Implement phase focused on configuring the failover mechanism and integrating the notification system into the operational network environment.

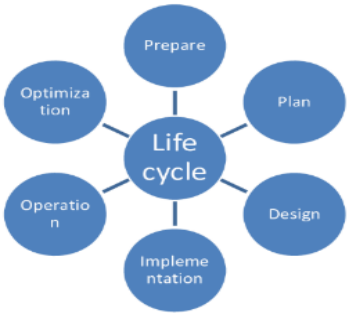


Fig. 1 PPDIOO Framework [12]

The Operate phase involved conducting failover testing, notification delay measurements, and operational monitoring under real network conditions.

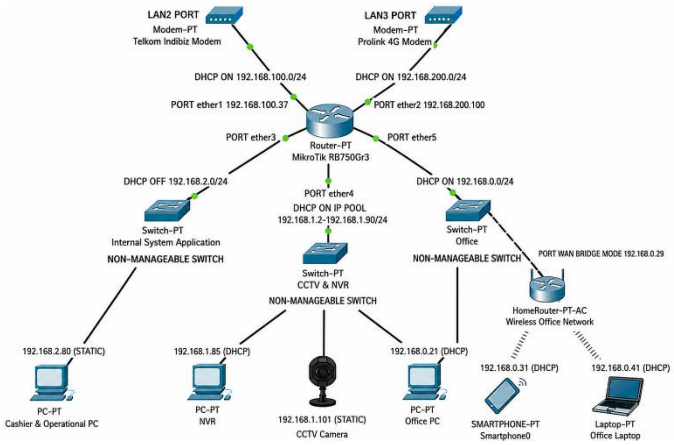


Fig. 2 Failover Network Topology

Finally, the Optimize phase focused on improving failover stability and client-side recovery behavior through connection tracking optimization and debouncing-based verification mechanisms to reduce route flapping during transient network disturbances.

B. Data Collection Method

Data collection was conducted through observation, interviews, and literature review [20]. Observations were performed directly within the operational environment of Kampoeng Semarang to identify network behavior, downtime frequency, failover performance, and notification responsiveness before and after system implementation. Interviews were conducted with operational staff and network administrators to gather information regarding existing network management procedures, operational impacts of downtime, and administrator responses during ISP disruptions. The collected information was used to support operational analysis and system evaluation.

In addition, a literature review was carried out to obtain theoretical references and implementation approaches

https://doi.org/10.31436/ijpcc.v12i2.705

related to network failover systems, MikroTik Netwatch, real-time notifications, and network monitoring mechanisms. Previous researches and technical documentation were also used to identify evaluation metrics and implementation strategies relevant to this research.

C. System Architecture

The proposed system utilized a dual-ISP architecture consisting of a primary ISP connection and a backup ISP connection managed by a MikroTik RouterOS device. Under normal conditions, all network traffic was routed through the primary ISP. When connectivity failure was detected on the primary connection, the failover mechanism automatically redirected traffic to the backup ISP. The architecture integrated MikroTik Netwatch for connectivity monitoring, Scheduler for repeated verification processes, automated routing scripts for failover execution, and Telegram Bot API and email notification services for notification delivery. Notification messages were transmitted automatically whenever network status transitions occurred between UP and DOWN conditions. This can be seen in Figure 3 below.

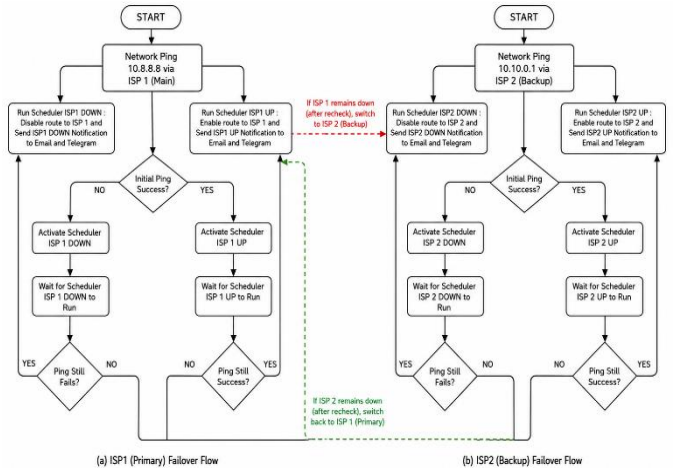


Fig. 3 Proposed Dual-ISP Failover System Architecture

The primary ISP served as the main operational internet connection, while the secondary ISP functioned as a backup link to maintain service continuity during connectivity failures. The router continuously monitored network reachability using connectivity verification mechanisms configured through Netwatch and Scheduler.

D. Proposed Debouncing-Based Failover Mechanism

The proposed failover mechanism applied a debouncing-based verification strategy to improve failover stability and reduce false failover events caused by temporary packet loss or unstable latency conditions. Instead of relying on a single connectivity check, the system performed repeated verification using MikroTik Scheduler before determining whether the primary connection should be considered unavailable. When repeated monitoring results indicated that the primary ISP connection failed consecutively, the system executed automated routing scripts to activate the backup ISP route. Conversely, when the primary ISP connection returned to stable conditions, the system restored traffic routing to the primary link. The proposed debouncing-based failover workflow shown in Figure 4.

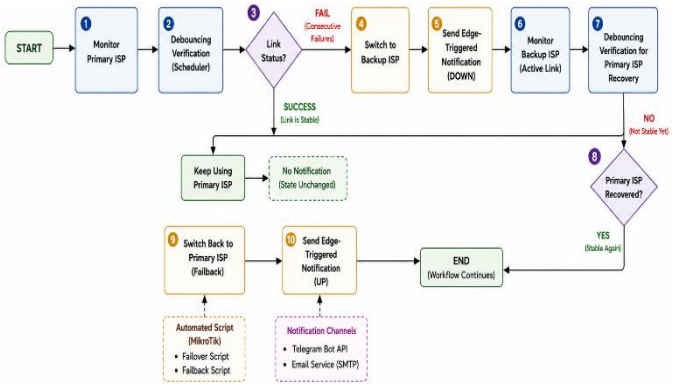


Fig. 4 Debouncing-based Failover Workflow

To reduce notification flooding during unstable network conditions, the system implemented an edge-triggered notification mechanism. Notifications were transmitted only when actual status transitions occurred between UP and DOWN states, preventing repetitive alerts under unchanged conditions. The failover process also incorporated connection tracking optimization to improve effective client-side recovery after ISP transitions. This optimization aimed to minimize session persistence issues and ensure that newly established client connections followed the currently active ISP route.

E. Experimental Setup

The experimental environment consisted of a MikroTik RouterOS device connected to two ISP connections. The primary ISP provided a bandwidth capacity of 150 Mbps, while the secondary ISP functioned as a backup connection using a GSM-based internet network. Approximately 30 active users utilized the network during operational hours. System testing was conducted by intentionally disconnecting the primary ISP connection to observe failover behavior, recovery delay, notification delivery performance, and client-side recovery consistency. Multiple testing scenarios were performed to evaluate failover responsiveness under operational conditions.

Experiments were conducted repeatedly under different operational conditions to evaluate the consistency and stability of the proposed failover mechanism as shown in Table 2 below.

https://doi.org/10.31436/ijpcc.v12i2.705

TABLE II
EXPERIMENTAL ENVIRONMENT CONFIGURATION

Component	Specification
Router Device	MikroTik RouterOS
Primary ISP	150 Mbps Fiber Connection
Backup ISP	GSM-based Internet Connection
Notification Platform	Telegram Bot API, Email
Monitoring Mechanism	Netwatch + Scheduler
Active Users	Approximately 30 users

Testing scenarios included ISP disconnection events, restoration events, notification delivery measurements, and client-side connectivity recovery evaluation after route transitions.

F. Evaluation Metrics

Several evaluation metrics were used to evaluate the performance and operational reliability of the proposed debouncing-based failover system. The primary metric was failover recovery time, which measured the duration required for the system to redirect network traffic from the primary ISP to the backup ISP after connectivity failure detection. This metric was used to assess how quickly the proposed mechanism could maintain network availability during ISP disruptions. In addition, notification delay was measured to determine the responsiveness of the Telegram and email notification services by calculating the time difference between failover event occurrence and notification reception by the administrator.

The evaluation also focused on failover stability during transient network disturbances and repeated connectivity fluctuations. This metric was used to analyze the effectiveness of the debouncing verification mechanism in preventing false failover events and reducing route flapping behavior caused by temporary packet loss or unstable latency conditions. Furthermore, client-side recovery was evaluated to determine whether internet connectivity from the user perspective could be restored consistently after ISP transitions occurred. This evaluation was important because successful route switching at the router level does not always guarantee immediate recovery for connected client devices. The complete evaluation metrics in this research are presented in Table 3.

TABLE III
EVALUATION METRICS

Metric	Description
Failover Recovery Time	ISP switching duration
Notification Delay	Notification transmission latency
Failover Stability	Resistance to route flapping
Client-Side Recovery	Effective recovery experienced by users

The collected evaluation data were analyzed to determine the overall effectiveness of the proposed debouncing-based failover mechanism and edge-triggered multi-channel notification system in improving operational network reliability. The evaluation results were then used to assess system responsiveness, failover consistency, notification performance, and effective connectivity recovery under real operational conditions. In addition, repeated testing scenarios were conducted to observe system behavior during both failover and failback processes under varying network conditions. The analysis also aimed to identify whether the proposed debouncing mechanism could minimize unstable route transitions and reduce redundant notifications caused by temporary connectivity disturbances.

IV. RESULTS AND DISCUSSION

This section presents the implementation results and operational evaluation of the proposed debouncing-based automated network failover system integrated with multi-channel real-time notifications. The discussion focuses on failover performance, notification responsiveness, failover stability, and client-side recovery behavior under real operational conditions at Kampoeng Semarang.

A. System Implementation

The proposed failover system was implemented using MikroTik RouterOS with dual-ISP connectivity consisting of a primary ISP and a backup GSM-based ISP. The implementation integrated MikroTik Netwatch, Scheduler, automated routing scripts, Telegram Bot API, and email notification services to support automatic failover and real-time monitoring.

The debouncing mechanism was implemented through repeated verification using Scheduler to ensure that failover decisions were not triggered by temporary packet loss or intermittent network disturbances. In addition, edge-triggered notifications were configured to ensure that notifications were transmitted only during actual status transitions between UP and DOWN conditions.

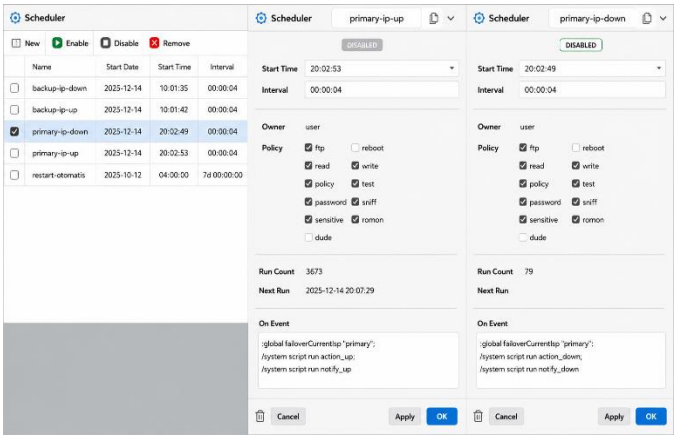


Fig. 5 Failover Scheduler Configuration

https://doi.org/10.31436/ijpcc.v12i2.705

Figure 5 presents the Scheduler configuration used to implement the debouncing-based failover verification mechanism in the proposed system.

B. Operational Downtime Observation

Before the implementation of the proposed system, network recovery procedures were performed manually whenever the primary ISP experienced connectivity failure. Operational observations showed that network interruptions occurred repeatedly during business hours and required several minutes for administrator intervention and recovery. To illustrate the operational downtime conditions that motivated the implementation of the failover system, several downtime events recorded during the observation period are presented in Table 4.

TABLE IV
SAMPLE OPERATIONAL DOWNTIME EVENTS

Date	Time DOWN	Time UP	Duration
03/05/2025	09:08:05	09:08:25	00:00:20
03/05/2025	10:20:20	10:20:40	00:00:20
03/05/2025	10:20:50	10:21:10	00:00:20
03/05/2025	11:15:23	11:15:46	00:00:23
05/05/2025	09:46:20	09:46:40	00:00:20

Table 4 presents several operational downtime events observed before the implementation of the proposed automated failover system. The recorded incidents indicate that internet connectivity disruptions occurred repeatedly during operational hours and required manual intervention from network administrators to restore connectivity. The recovery duration varied depending on the type of disturbance, administrator response time, and ISP recovery process, resulting in service interruptions that affected daily operational activities. These findings demonstrate the necessity of an automated failover mechanism capable of reducing recovery dependency on manual handling while improving network availability and operational continuity.

C. Failover Recovery Performance

The core evaluation focused on measuring the failover recovery time required to redirect network traffic from the primary ISP to the backup ISP after connectivity failure detection. The proposed debouncing-based mechanism successfully performed automatic failover within several seconds while maintaining stable routing behavior.

Multiple failover testing scenarios were conducted by intentionally disconnecting the primary ISP connection during operational conditions. The results showed that the proposed system achieved relatively consistent failover recovery performance with minimal variation between test scenarios.

TABLE V
FAILOVER RECOVERY TIME FROM ISP 1 TO ISP 2

Scenario (Date)	Failover Duration (seconds)
ISP 1 to ISP 2 (05-08-2025)	4,0
ISP 1 to ISP 2 (06-08-2025)	4,0
ISP 1 to ISP 2 (07-08-2025)	4,0
ISP 1 to ISP 2 (08-08-2025)	5,0
ISP 1 to ISP 2 (09-08-2025)	4,0
Average Recovery Time	4,2

Table 5 presents the failover recovery time measured during the transition process from ISP 1 (primary connection) to ISP 2 (backup connection). The testing results show that the proposed debouncing-based failover mechanism successfully performed automatic route switching within a relatively consistent time range, achieving an average recovery time of approximately 4.2 seconds. The relatively small variation between testing scenarios indicates that the repeated verification process implemented through MikroTik Scheduler was able to maintain stable failover behavior while minimizing false failover events caused by temporary connectivity disturbances. The recovery duration was influenced by connectivity verification intervals, repeated Scheduler-based checks, and routing table update processes within MikroTik RouterOS, demonstrating that the proposed mechanism could maintain both failover responsiveness and operational stability during ISP disruptions.

D. ISP Performance Evaluation

The performance of both ISP connections was evaluated to determine whether the backup ISP could maintain operational connectivity during failover conditions. Network speed measurements included download throughput, upload throughput, ping latency, and public IP verification. The evaluation results showed that the primary ISP provided significantly better throughput and lower latency compared to the backup ISP. However, the backup connection remained sufficient to maintain essential operational services during failover events.

TABLE VI
ISP PERFORMANCE AND PUBLIC IP VERIFICATION

Path	Download (Mbps)	Upload (Mbps)	Ping (ms)	Public IP
ISP 1 (Main)	146,0	23,9	101	36.80.254.36
ISP 2 (Backup)	31,4	19,9	120	114.10.8.72

Table 6 presents the performance evaluation results of both ISP connections, including download throughput, upload throughput, ping latency, and public IP verification. The results indicate that the primary ISP provided significantly better network performance compared to the backup ISP, particularly in terms of throughput capacity and lower latency, making it more suitable for daily operational

https://doi.org/10.31436/ijpcc.v12i2.705

traffic. However, the backup ISP was still capable of maintaining essential internet connectivity during failover conditions despite its lower performance characteristics. In addition, the different public IP addresses obtained during testing confirm that the failover mechanism successfully redirected network traffic to a separate ISP connection rather than maintaining local routing behavior, demonstrating that the proposed failover system operated correctly during ISP transition processes.

E. Notification Performance Evaluation

The notification system was evaluated to measure the responsiveness of Telegram and email notifications during failover and failback events. Notification delay was calculated based on the time difference between failover event occurrence and notification reception by the administrator.

The experimental results demonstrated that Telegram notifications consistently achieved lower delivery delay compared to email notifications. Telegram provided near real-time operational alerts, while email notifications required additional processing time due to SMTP transmission and client synchronization processes.

TABLE VII
NOTIFICATION DELAY EVALUATION

Scenario	Average Telegram Delay (seconds)	Average Email Delay (seconds)
ISP 1 to ISP 2	2,22	8,61
ISP 2 to ISP 1	1,89	8,56

Table 7 presents the notification delay evaluation results for both Telegram and email notification services during failover and failback scenarios. The results show that Telegram notifications consistently achieved lower delivery latency compared to email notifications, with average delays of approximately 2 seconds, while email notifications required approximately 8–9 seconds to reach the administrator. This difference is primarily influenced by the lightweight API-based delivery mechanism used by Telegram, whereas email notifications require additional SMTP transmission and client synchronization processes before messages are received. The implementation of edge-triggered notifications also helped reduce redundant alert transmissions during unstable connectivity conditions, allowing administrators to receive more efficient and relevant operational status information during ISP transition events.

F. Notification Verification

The functionality of the multi-channel notification system was further validated by observing notification messages received by administrators during failover events [21]. Notifications were generated automatically whenever actual status transitions occurred between UP and DOWN conditions. The notification mechanism successfully delivered operational status information through both Telegram and email platforms, improving administrator awareness regarding network disturbances and recovery events.

Figure 6 and Figure 7 presents the output of the Telegram and email notification system generated automatically during failover and failback events. The notifications provide real-time operational information regarding ISP status transitions, including DOWN conditions when the primary ISP becomes unavailable and UP conditions when connectivity is restored. The results demonstrate that the proposed multi-channel notification mechanism successfully delivered status updates through both Telegram Bot API and email services, thereby improving administrator awareness and enabling faster operational response during network disturbances. In addition, the implementation of edge-triggered notification behavior ensured that alerts were transmitted only during actual status changes, reducing redundant notifications and minimizing alert flooding during unstable network conditions.

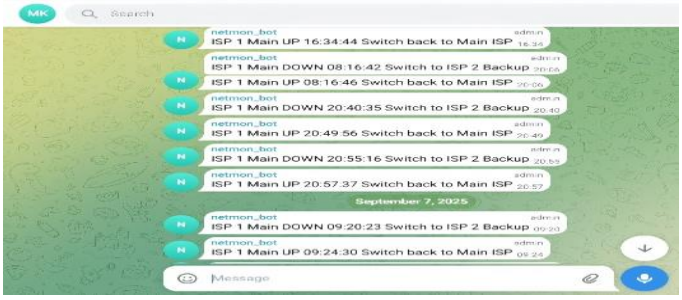


Fig. 6 Telegram Notification Output

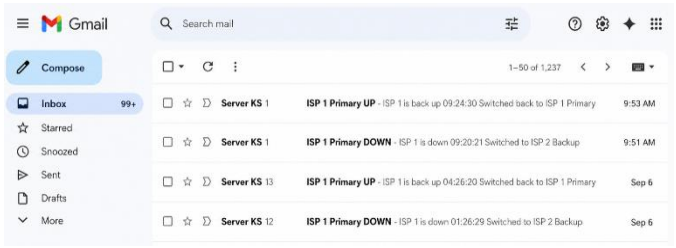


Fig. 7 Email Notification Output

G. Failover Optimization and Client-Side Recovery

During operational testing, an important issue was identified in which several client devices experienced delayed internet recovery after failover transitions despite successful route switching at the router level. This condition was caused by connection tracking persistence and session stickiness behavior that maintained existing sessions on the previous ISP route.

To improve effective client-side recovery, optimization was performed by clearing connection tracking sessions and

https://doi.org/10.31436/ijpcc.v12i2.705

adjusting failover handling mechanisms to ensure that newly established client connections followed the currently active ISP route.

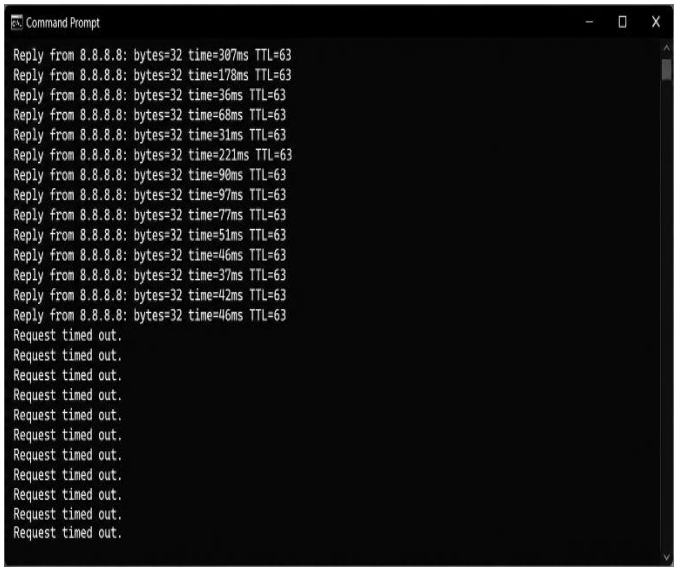


Fig. 8 Client Connectivity Issue Before Optimization

Figure 8 illustrates the client-side connectivity issue observed before the optimization process was applied to the proposed failover system. Although the router had successfully performed failover routing from the primary ISP to the backup ISP, several client devices still experienced temporary connectivity loss indicated by repeated request timeout responses during continuous ping testing. This condition was primarily caused by connection tracking persistence and session stickiness behavior, where existing client sessions remained associated with the previous ISP route even after routing changes had been executed. The results demonstrate that successful failover at the router level does not always guarantee immediate recovery from the client perspective, highlighting the importance of connection tracking optimization to improve effective client-side recovery during ISP transition processes.

After optimization was applied, additional testing showed that client-side connectivity recovery became significantly more stable during ISP transitions. The optimization successfully minimized delayed recovery behavior and improved effective internet accessibility from the user perspective.

Figure 9 presents the client-side connectivity condition after optimization was applied to the proposed failover system. The continuous ping results show that internet connectivity was restored more consistently during ISP transition processes, with significantly reduced request timeout occurrences compared to the condition before optimization. This improvement was achieved through connection tracking optimization and failover handling adjustments that ensured newly established client sessions followed the currently active ISP route. The results indicate that the optimization process successfully improved effective client-side recovery, minimized session persistence issues, and enhanced overall connectivity stability during failover and failback operations.

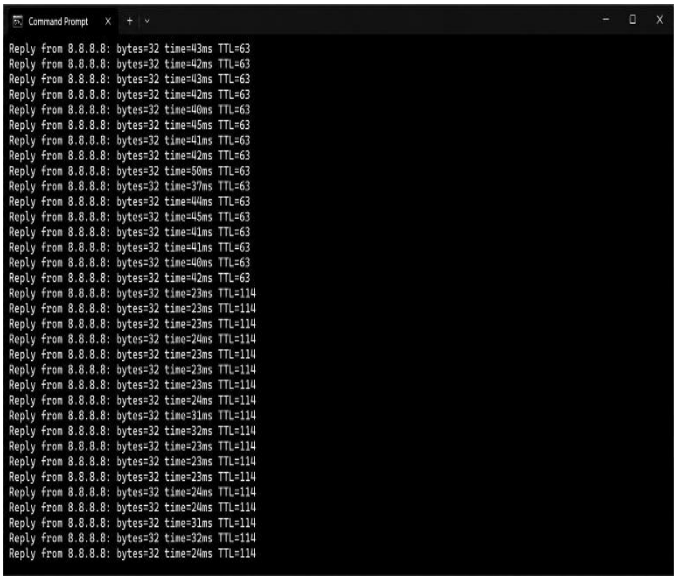


Fig. 9 Client Connectivity Recovery After Optimization

H. Optimization Impact Summary

The optimization process improved not only router-level failover behavior but also effective client-side recovery consistency during ISP transitions. The proposed debouncing mechanism, edge-triggered notifications, and connection tracking optimization collectively improved operational reliability under real-world network conditions.

The Table 8 summarizes the impact of the optimization process applied to the proposed failover system. The results indicate that the implemented optimizations successfully improved failover stability, client-side recovery consistency, and overall operational reliability during ISP transition processes. Before optimization, several issues such as unstable route transitions, session stickiness, delayed client recovery, and repeated request timeout occurrences were still observed during failover events. After optimization, the system demonstrated more stable failover and failback behavior, reduced route flapping, improved connection tracking handling, and more consistent internet recovery from the client perspective. In addition, the implementation of edge-triggered notifications reduced redundant alert transmissions during unstable network conditions, thereby improving monitoring efficiency and administrator awareness.

https://doi.org/10.31436/ijpcc.v12i2.705

TABLE VIII
OPTIMIZATION IMPACT SUMMARY

Optimization Aspect	Before Optimization	After Optimization	Impact
Failover Stability	Occasional unstable route transitions during transient disturbances	More stable failover and failback behavior	Reduced route flapping
Client-Side Recovery	Several clients experienced delayed internet recovery	Faster and more consistent connectivity restoration	Improved effective recovery
Connection Tracking Behavior	Existing sessions remained attached to previous ISP route	Sessions followed the active ISP route correctly	Reduced session stickiness
Ping Test Results	Frequent request timeout occurrences during ISP transition	Minimal timeout occurrences during failover	Improved connectivity continuity
Notification Behavior	Potential redundant notifications during unstable conditions	Notifications sent only during actual state changes	Reduced alert flooding
Operational Reliability	Recovery still partially affected by unstable transitions	More reliable and consistent network operation	Improved operational continuity

I. User Satisfaction Evaluation

To evaluate operational impact from the user perspective, a user satisfaction survey was conducted after system implementation and optimization. The survey results indicated that users perceived improvements in connectivity stability, operational responsiveness, and downtime handling performance.

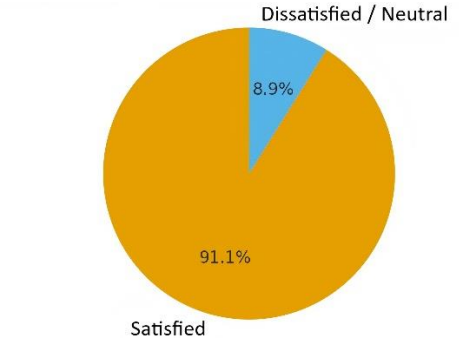


Fig. 10 User Satisfaction Survey Results

Figure 10 presents the results of the user satisfaction survey conducted after the implementation and optimization of the proposed failover monitoring system. The survey results indicate a high level of user satisfaction, with approximately 91.1% of respondents reporting positive perceptions regarding network stability, failover responsiveness, and operational reliability during internet disruptions. The improved satisfaction level suggests that the proposed debouncing-based failover mechanism and multi-channel notification system successfully enhanced service continuity and reduced the operational impact of ISP connectivity failures. These findings also demonstrate that the implemented optimizations contributed not only to technical performance improvements but also to better user experience in real operational environments.

J. Discussion

The experimental results demonstrate that the proposed debouncing-based failover mechanism successfully improved both failover responsiveness and operational stability within a real-world networking environment. The implementation of repeated verification using MikroTik Scheduler effectively reduced false failover events caused by transient packet loss and unstable latency conditions, thereby minimizing route flapping behavior during ISP disruptions. Unlike conventional failover implementations that rely primarily on single connectivity checks, the proposed mechanism introduced a more stable decision-making process before route switching was executed. Despite the additional verification process, the system was still able to maintain relatively fast recovery performance with an average failover recovery time of approximately 4.2 seconds. Furthermore, the integration of edge-triggered multi-channel notifications improved operational monitoring efficiency by reducing redundant alerts during unstable connectivity conditions. Telegram notifications also demonstrated lower latency compared to email notifications, making them more suitable for real-time operational monitoring.

Another important finding of this research is the impact of connection tracking optimization on effective client-side recovery during ISP transition processes. Initial testing showed that successful failover at the router level did not always guarantee immediate internet recovery from the client perspective due to session persistence and connection tracking behavior. After optimization was applied, client-side connectivity recovery became significantly more stable with fewer request timeout occurrences during failover and failback operations. These results indicate that effective failover implementation should not only focus on automatic route switching but also

consider session handling and user-side recovery consistency. Overall, the proposed framework demonstrates that combining debouncing-based failover verification, edge-triggered notifications, and connection tracking optimization can improve network availability, operational reliability, and service continuity in small-to-medium business environments.

V. Conclusion

This research successfully implemented a debouncing-based automated network failover system integrated with multi-channel real-time notifications using MikroTik RouterOS, Telegram Bot API, and email services within the operational environment of Kampoeng Semarang. The proposed system applied repeated verification through MikroTik Scheduler to improve failover stability and reduce false failover events caused by transient network disturbances. Experimental results showed that the system achieved an average failover recovery time of approximately 4.2 seconds, while Telegram and email notifications achieved average delivery delays of approximately 2 seconds and 8–9 seconds, respectively. The implementation of edge-triggered notifications also reduced redundant alert transmissions during unstable connectivity conditions, improving operational monitoring efficiency and administrator awareness.

In addition, the optimization process successfully improved effective client-side recovery by minimizing session persistence issues and enhancing connectivity stability during ISP transition processes. The user satisfaction evaluation further indicated that the proposed system improved perceived network reliability and operational continuity in real-world conditions. Overall, the proposed framework demonstrates that combining debouncing-based failover verification, automated route switching, and multi-channel notification mechanisms can improve network availability and operational resilience in small-to-medium business environments. Future work may include integrating centralized logging systems, adaptive monitoring thresholds, dashboard-based monitoring platforms, and support for multi-ISP scalability to further enhance system reliability and operational management capabilities.

Acknowledgment

The research team extends its heartfelt gratitude to the Universitas Semarang, Universitas Nias Raya, and Indonesia Defense University for their invaluable support and contribution to this research.

Conflict of Interest

Authors state no conflict of interest.

Author (s) Contribution Statement

F.W.C: Conceptualization, Data curation, Formal analysis, Methodology, Project administration, Writing—original draft, and Writing—review and editing. S.V.R: Validation, Software, Visualization, Writing—original draft, and Writing—review and editing. F.L: Resources, Investigation, Supervision, Writing—original draft validation, and Writing—review & editing. E.A.F: Validation and Writing—review and editing.

Data Availability Statement

The data that support the findings of this research are available from the corresponding author upon reasonable request.

Ethics Statement

This research did not require ethical approval.

References

[1] A. Abdurrahman, "Extending the IBCDE Framework to Explore Barriers and Drivers in Indonesia ' s Digital Economy," *J. Digit. Econ.*, vol. 4, no. August, pp. 123–143, 2025, doi: 10.1016/j.jdec.2025.08.003.

[2] R. E. Caraka *et al.*, "Evaluating Internet Accessibility and Equity in Indonesia using Integrated SUSENAS and OOKLA Data," *Discov. Sustain.*, vol. 7, no. 497, pp. 1–31, 2026, doi: 10.1007/s43621-026-02882-x.

[3] A. Djenna, S. Harous, and D. E. Saidouni, "Internet of Things Meet Internet of Threats : New Concern Cyber Security Issues of Critical Cyber Infrastructure," *Appl. Sci.*, vol. 11, no. 4580, pp. 1–30, 2021, doi: 10.3390/app11104580.

[4] B. Howell, "Beyond Infrastructure : Internet Ecosystem Resilience and the Public Good," *Telecomm. Policy*, vol. 49, no. 7, p. 102998, 2025, doi: 10.1016/j.telpol.2025.102998.

[5] F. Toriq and B. Santoso, "Effect of Load Balancing Bonding and Failover on Speed, Latency, Average, and Packet Loss," *J. Appl. Informatics Comput.*, vol. 8, no. 2, pp. 326–331, 2024, doi: 10.30871/jaic.v8i2.8276.

[6] Hoesiin and B. Yuliandi, "Optimizing Network Uptime Through Dual ISP Failover Implementation Using Netwatch," *J. Informatic, Educ. Manag.*, vol. 7, no. 2, pp. 336–349, 2025.

[7] W. M. Elsayed and A. A. Abohany, "Strengthening Network Resilience and Optimizing Resources in Large Enterprises: Innovative Infrastructure and Cybersecurity Strategies," *Cluster Comput.*, vol. 29, no. 252, pp. 1–31, 2026, doi: 10.1007/s10586-026-06075-z.

[8] B. Anbalagan, "Proactive Failover and Automation Frameworks for Mission-Critical Workloads : Lessons from Manufacturing Industry," *Int. J. Res. Appl. Innov.*, vol. 6, no. 1, pp. 8279–8296, 2023, doi: 10.15662/IJRAI.2023.0601004.

[9] C. K. Ryu, M. C. Lee, I. H. Hong, J. H. Park, J. D. Lee, and S. Y. Choi, "Heterogeneous PLC-Based Distributed Controller with Embedded Logic-Monitoring Blackbox for Real-Time Failover," *Electronics*, vol. 14, no. 4359, pp. 1–36, 2025, doi: 10.3390/electronics14224359.

[10] L. Fotia, R. Gaeta, F. Messina, D. Rosaci, and G. M. L. Sarné, "Artificial Intelligence for High-Availability Systems: A Comprehensive Review," *Computers*, vol. 15, no. 4, p. 231, 2026, doi: 10.3390/computers15040231.

[11] T. Li, J. K. Haines, M. F. R. De Eguino, J. I. Hong, and J. Nichols, "Alert Now or Never : Understanding and Predicting Notification

Preferences of Smartphone Users," *ACM Trans. Comput. Interact.*, vol. 29, no. 5, p. 39, 2022, doi: 10.1145/3478868.

[12] P. Petrov, I. Kuyumdzhiev, R. Malkawi, G. Dimitrov, and J. Jordanov, "Digitalization of Educational Services with Regard to Policy for Information Security," *TEM J.*, vol. 11, no. 3, pp. 1093–1102, 2022, doi: 10.18421/TEM113.

[13] R. Rayendra, A. K. Setyanegara, M. Raafi, and F. Tara, "Development of a Web-Based Server Monitoring System Using the PPDIOO Method," *Int. J. Adv. Technol. Eng. Inf. Syst.*, vol. 4, no. 2, pp. 207–224, 2025, doi: 10.55047/ijateis.v4i2.1690.

[14] A. Purwanto and B. Soewito, "Optimization Problem of Computer Network using PPDIOO," *ICIC Express Lett.*, vol. 15, no. 7, pp. 769–777, 2021, doi: 10.24507/icicel.15.07.769.

[15] Bustomi and M. Tahir, "Evaluation of PCC Load Balancing for Dual ISP Networks : Enhancing Throughput and Traffic Stability," *Edumatic J. Pendidik. Inform.*, vol. 10, no. 10, pp. 240–249, 2026, doi: 10.29408/edumatic.v10i1.34243.

[16] M. F. I. P. Putra, J. M. Parenreng, and M. Yahya, "Integration of Telkom ISP and 3 LTE Using PCC Method to Improve Internet Connection Stability," *Iota*, vol. 04, no. 02, pp. 244–256, 2024, doi: 10.31763/iota.v4i2.728.

[17] F. Zikra, Anwar, and I. Safar, "Monitoring Networks Based on Simple Network Management Protocol With Cacti Application and Telegram Notifications on Linux Ubuntu," *J. Teknol. Rekayasa Inf. dan Komput.*, vol. 9, no. 1, pp. 55–65, 2026, doi: 10.30811/jtrik.v9i1.8701.

[18] I. Vingestin, T. U. Kalsum, and Y. Mardiana, "The Design Of Network Monitoring System Using SNMP Protocol With Telegram Notification," *J. Media Comput. Sci.*, vol. 2, no. 1, pp. 93–100, 2022, doi: 10.37676/jmcs.v2i1.3441.

[19] S. Nurjanah, F. Sembiring, and R. R. Ayuningsih, "Integration of Telegram Bot and Uptime Kuma for Wi-Fi Network Monitoring using Mikrotik," *Pilar Nusa Mandiri J. Comput. Inf. Syst.*, vol. 20, no. 2, pp. 118–126, 2024, doi: 10.33480/pilar.v20i2.5535.

[20] B. P. Neupane, N. Dahal, R. K. Dhakal, M. K. Hasan, J. A. Villarama, and B. G. Fabros, "Data Collection Methods in Qualitative Research: Researchers' Reflections," *Front. Res. Metrics Anal.*, vol. 11, no. 1778160, pp. 1–9, 2026, doi: 10.3389/frma.2026.1778160.

[21] H. Zhang, R. Zhang, and J. Sun, "Developing Real-time IoT-based Public Safety Alert and Emergency Response Systems," *Sci. Rep.*, vol. 15, no. 29056, pp. 1–25, 2025, doi: 10.1038/s41598-025-13465-7.

International Journal on Perceptive and Cognitive Computing (IJPCC) Vol 12, Issue 2 (2026)
https://doi.org/10.31436/ijpcc.v12i2.624

Analysing Educational Disparities Between Urban and Rural Areas in Malaysia

Muhammad Faiq Muhammad Fauzi[1], Muhammad Amir Farhan Muhammad Kamal[2], Norzariyah Yahya[3]*

[1,2]Dept. of Information Systems, International Islamic University Malaysia, 53100 Kuala Lumpur, Malaysia
[3]Dept. of Computer Science, International Islamic University Malaysia, 53100 Kuala Lumpur, Malaysia

*Corresponding Author: norzariyah@iium.edu.my
(Received: 6th August 2025; Accepted: 3rd September, 2025; Published on-line: 30th July, 2026)

Abstract— Educational disparities between urban and rural areas remain a critical challenge for Malaysia's national development. This study investigates the factors driving rural–urban academic achievement gaps by combining non-parametric inferential statistics with supervised machine learning data-mining techniques. Primary data were collected from $N = 80$ respondents via a bilingual online survey capturing demographic attributes, learning environments, support systems, and material/ICT resource availability. Non-parametric hypothesis testing confirms a statistically significant difference in overall satisfaction with educational opportunities between independent urban and rural cohorts. To explain this divide, three machine learning models (J48, Random Forest, and Logistic Regression) were evaluated using 10-fold cross-validation in Weka to predict tertiary academic attainment. Random Forest achieved the highest classification performance (75.00% accuracy, $\kappa = 0.451$, ROC Area $= 0.794$). Information Gain feature evaluation reveals that age (0.279 bits) and family academic guidance (0.097 bits) are the dominant drivers of higher education attainment, whereas geographical location alone contributes minimally (0.007 bits). Furthermore, J48 decision tree logic indicates that for young rural learners with adequate home environments, access to ICT and study materials is the decisive factor in reaching tertiary education. These findings demonstrate that Malaysia's educational gap is primarily mediated by household resources and academic support rather than geography itself, suggesting that policy interventions should prioritise household-level digital access and family engagement programs over purely regional funding.

Keywords— Rural education, urban education, academic performance, socioeconomic status, machine learning classification, Weka.

I. Introduction

Educational gaps between urban and rural areas have long been a major concern for policymakers, educators, and scholars. These gaps usually appear as differential access to quality education, learning materials, and academic support, affecting students' academic performance and prospects. Even though equal educational access is one of the primary objectives of Malaysia's national education initiatives, students from rural areas often face systematic disadvantages. These include limited access to academic enrichment activities, fewer qualified educators, and poor school infrastructure.

Although several studies have addressed this topic, little is known about Malaysia's unique circumstances, which are influenced by its socioeconomic diversity, geographical features, and public education funding. This study aims to investigate how academic performance is impacted by region and pinpoint the fundamental causes of these disparities. This study provides valuable insights that help improve educational policies and ensure more equitable opportunities for students across the country.

II. Significance of Study

Understanding the factors contributing to educational disparities is crucial for informing policy development and implementing context-specific solutions. This research provides insights into the specific challenges faced by rural students and offers recommendations to policymakers and educators for bridging the academic gap. While the study provides valuable findings, its limitations, such as small sample size and self-reported data, should be addressed in future research.

III. Literature Review

Educational disparities between rural and urban areas have been extensively studied. According to Xiang and Stillwell [1], despite rising investment in rural education in China, significant gaps remain, particularly in high school education, which is not compulsory and receives less government support. Similarly, Ahsan et al. [2] stated that

urban parents usually invest more in their children's education, leading to better academic achievement than in rural areas. The disparity often stems from differences in school quality, the availability of qualified teachers, and access to learning materials [3]. Wood [4] highlighted that urban students benefit from better facilities, qualified educators, and enriching learning environments, whereas rural students often confront poor infrastructure and limited educational opportunities. Drescher et al. [5] added that although overall differences between rural and nonrural students in the U.S. are modest, disparities become more pronounced among disadvantaged groups, with local conditions significantly influencing outcomes.

Socioeconomic status (SES) may or may not be a major determinant of academic success[3]. Rodríguez-Hernández et al. [6] conducted a systematic assessment of 42 studies and discovered a weak but positive link between socioeconomic status and academic success in higher education. Prior academic achievements, university experiences, and employment status proved to be stronger measures than SES alone. Munir et al. [3] revealed that students from families with higher socioeconomic status tend to perform better due to greater access to educational resources, parental support, and extracurricular activities. Research also shows that, while SES influences educational attainment, its effects can be mitigated by factors such as parental engagement and school quality. Ahsan et al. [2] discovered that intelligence is particularly crucial in rural areas, where school quality is generally lower than in other areas. Similarly, Xiang and Stillwell [1] proposed a capability-based framework for assessing rural-urban education gaps, emphasising the role of regional policies in shaping academic success.

Parental involvement has also been found to affect academic achievement across different socioeconomic groups significantly. A study by Şengönül [7] established that early parent-child interaction, reading activities, and school engagement positively influence performance in subjects like math and science. Furthermore, the paper highlights that families with higher SES are more likely to be actively involved in their children's education, reinforcing the importance of early engagement strategies.

Several studies suggested methods to address educational inequality. Wood [4] suggested increasing funding for rural education, improving infrastructure, and implementing e-learning alternatives. Xiang and Stillwell [1] emphasised the need for targeted policies, advocating for more context-sensitive policies that are adapted to different developmental levels. Furthermore, Ahsan et al. [2] found that government investment in rural school quality could enhance both equity and efficiency in education. Kawuryan et al. [8] supported this by highlighting the importance of teacher quality, pointing out that a lack of qualifications and certifications among educators in Indonesia affects student performance and reinforces inequality.

International perspectives on urban-rural educational divides show similar patterns. In the Netherlands, van Maarseveen [9] found that children in urban areas achieve higher levels of education than their rural peers, even when controlling for cognitive ability and family background. Higher expected returns, lower perceived costs, and access to better information drive this urban advantage. Likewise, Lu [10] analysed urban and rural schools in Yuyao, China, and found that urban schools are generally better resourced, have more qualified teachers, and offer a broader range of extracurricular and technological opportunities. Lu said that to close the educational gap, there needs to be more investment in rural areas, better training for teachers, and new ways of running schools.

IV. Methodology

This study adopts a computational data-mining approach to explore the educational disparities between rural and urban areas in Malaysia. The methodology is structured as a four-stage pipeline, consisting of (A) Data collection, (B) data pre-processing and encoding, (C) descriptive and inferential statistical analysis, and (D) supervised classification and attribute-importance analysis using the Weka [11] data-mining workbench. Below is a description of each stage.

A. Data Collection

A structured questionnaire was developed to collect data on five construct groups related to academic achievement, demographic characteristics (age, gender, area of residence, highest level of education), learning environment (home, school), social influence (peer influence), family support (emotional, financial, academic guidance) and availability of learning resources (physical, digital study materials). The survey was written in both English and Malay to make it accessible to respondents.

The questionnaire was distributed through a cloud-hosted Google Forms instrument and through two popular social messaging platforms (Telegram and WhatsApp). 250 invitations were issued, and 80 valid responses were received over two weeks (response rate 32%). Google Forms automatically records submissions into a spreadsheet, automating the entire ingestion path from respondent to analysable dataset, with no manual transcription or paper-based data entry involved. Then, the raw response spreadsheet was exported and converted into the Attribute-Relation File Format (ARFF) as required by the Weka workbench.

https://doi.org/10.31436/ijpcc.v12i2.624

B. *Data Preprocessing and Encoding*

Data preprocessing was carried out in Weka's Preprocess module. Nominal attributes were declared as categorical items, including age bracket, gender, area, and highest level of study. The opinion items (learning environment, family support, peer influence and access to study materials) on a Likert scale were declared as numeric attributes on their native 1–5 scale. In accordance with the educational-inequality literature review, the last attribute, 'highest level of study', was recoded into a two-class target, i.e., tertiary (Degree, Master or PhD) versus non-tertiary (SPM, Foundation, Diploma or below). The access to study materials, which explicitly includes online resources, practice questions, videos, and physical materials, is treated as an ICT access indicator at the household level. The final analysable dataset consists of 80 instances, 10 predictor attributes and a binary class attribute.

C. *Descriptive and Inferential Analysis*

Each attribute was characterised by calculating measures of central tendency and dispersion (mean, median, mode, variance and standard deviation). Chebyshev's theorem $\left(1-\frac{1}{k^2}\right)$ was used to generalise the distributions of continuous attributes. The significance of the raw urban-rural gap in educational satisfaction was tested using the Wilcoxon signed-rank test, a non-parametric test appropriate for ordinal Likert data. The outputs show an urban-rural difference but do not identify the underlying factors that drive it. This step motivates the subsequent data-mining stage.

D. *Supervised Classification and Attribute-Importance Analysis*

The Weka Classify module was used to train and compare three supervised classifiers to determine the factors predicting academic achievement and their relative importance: J48 (Weka implementation of C4.5 decision-tree algorithm), RandomForest (ensemble of 100 unpruned trees) and Logistic (multinomial logistic regression with ridge regularisation). These three algorithms were selected to cover the interpretability and accuracy range: J48 produces a human-readable tree, Logistic provides signed coefficient weights, and RandomForest typically achieves higher predictive accuracy at the cost of a less transparent decision surface.

All classifiers were evaluated using 10-fold cross-validation, the default protocol in Weka, and were reported based on classification accuracy, Cohen's Kappa, the confusion matrix, and the per-class precision, recall, and F-measure. Weka's Select Attributes module was used with the InfoGainAttributeEval evaluator and a Ranker search over the full training set to rank attribute importance, which is used to measure how much information each attribute contributes to predicting the class label.

The descriptive-inferential stage (C) and the data-mining stage (D) together establish a two-layer analytical strategy, of which the first identifies the existence of the urban–rural gap, and the following stage explains its mechanism by ranking the predictors that mediate it.

V. FINDINGS AND DISCUSSIONS

A. *Descriptive Analysis:*

The pie chart in Fig. 1 shows a significant age distribution, with the largest age group being 18-24 years old, accounting for 41.3% of the 80 respondents. This group is likely young adults, possibly students or early-career individuals. The second largest age group is 45-54 years old, making up 35% of respondents. This group likely represents mature adults, parents, or working professionals, providing valuable perspectives on educational experiences. The 25-34 age range, comprising 11.3% of respondents, may represent young professionals or postgraduate students. The smallest proportions are under 18, 45-54, and above 54. This distribution suggests that the data primarily reflects the views and experiences of the younger to middle-aged demographic, which enhances the credibility of the findings on education access and quality.

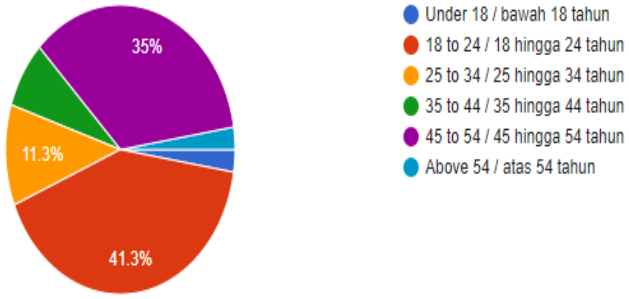


Fig. 1 Pie chart showing the percentage of respondents and their age group

B. *Preliminary Analysis*

The bar chart in Fig.2 shows the distribution of answers about opinions about the learning environment in urban areas. According to the findings, a sizable majority of respondents have a favourable opinion of the urban learning environment. 34.07% of participants agree, 52.41% strongly agree, and 86.48% have a positive attitude. On the other hand, discontent is minimal, since only 2.16% strongly disagree and another 2.16% disagree. Conversely, 9.2% of those surveyed expressed no opinion. The findings imply that most people view the urban learning environment as helpful and productive.

https://doi.org/10.31436/ijpcc.v12i2.624

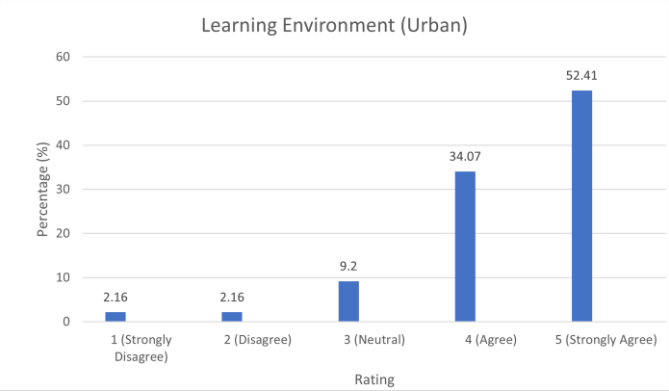


Fig. 2 Bar chart showing the percentage of respondents ' opinions about the learning environment in urban areas

The bar chart in Fig.3 shows data on how people in urban regions feel about the assistance they get from their family. With 44.62% of respondents strongly agreeing and 36.56% agreeing that they receive family support, the data demonstrate a very positive perception. As a result, 81.18% of the answers were positive overall. Just 5.92% of people disagreed, while 12.9% of participants were neutral. Interestingly, no responders expressed strong disagreement. According to these results, most people in metropolitan areas feel supported by their families, which is indicative of a supportive and upbeat family atmosphere.

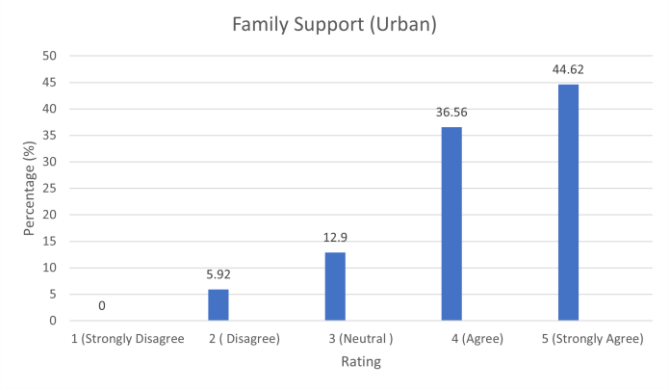


Fig. 3 Bar chart showing the percentage of respondents' opinions about family support in urban areas

The bar chart in Fig. 4 shows how people in cities rate the availability of study materials on a 5-point scale. With 35.48% of respondents rating their access as good and 37.1% as very good, the data shows a generally favourable perspective. This suggests that a sizable majority (more than 72%) think their access to study materials is beneficial. In contrast, only 3.23% of respondents assessed the access as poor, while 24.19% said it was simply adequate. Interestingly, all respondents said they had at least fair access. According to the data, urban students typically have access to study materials that range from adequate to exceptional.

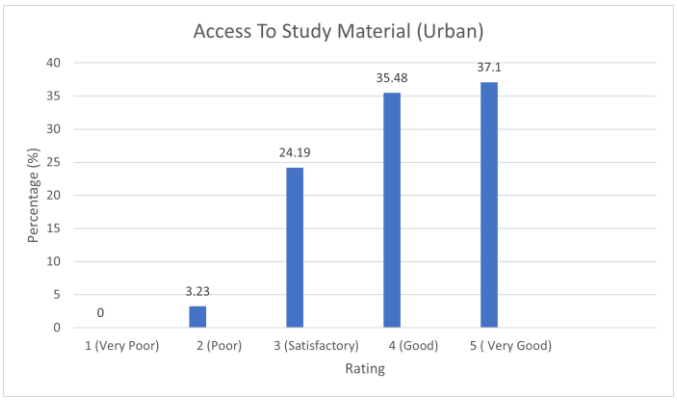


Fig. 4 Bar chart showing the percentage of respondent about opinions about access to study materials in urban areas

The bar chart in Fig. 5 displays the distribution of answers on a five-point Likert scale about the quality of the learning environment in rural areas. There were 3.67% of respondents who disagreed and 3.67% who strongly disagreed with the positive opinion. 18.52%, a higher percentage, stayed neutral. Notably, 33.37% of respondents agreed that the learning environment was favourable, while 52.41% strongly agreed. More than 85% of participants in rural areas expressed agreement or strong agreement, indicating that most of them had a positive perception of their learning environment.

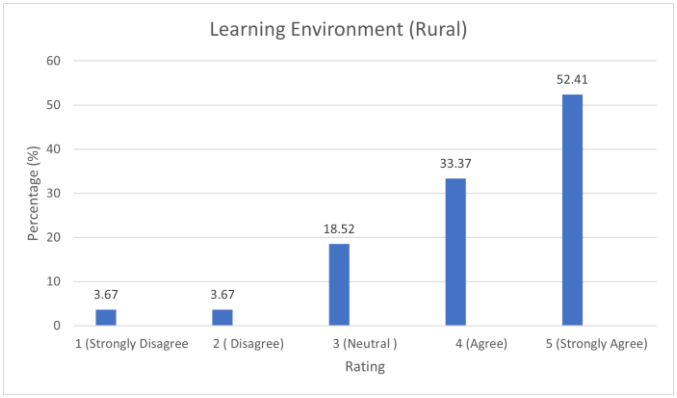


Fig. 5 Bar chart showing the percentage of respondents about their opinions on the learning environment in rural areas

The bar chart in Fig. 6 shows the percentage distribution of rural respondents' opinions about family support. Most individuals agreed that family support is important; only 3.72% disagreed, and none strongly disagreed. Conversely, 22.22% of those surveyed took a neutral position. A large majority, 27.78%, agreed that having strong family support is important, and an even larger majority, 44.62%, strongly agreed. Overall, more than 72% of respondents said they agreed or strongly agreed, indicating that rural participants have a very positive opinion of family support.

https://doi.org/10.31436/ijpcc.v12i2.624

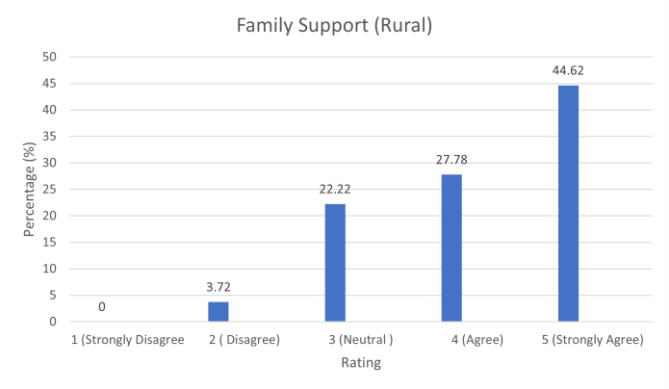


Fig. 6 Bar chart showing the percentage of the respondents' opinions on family support in rural areas

The bar chart in Fig. 7 displays the distribution of responses on a five-point rating system for the availability of study materials among participants in rural areas. A few respondents (5.56%) assessed access as poor, and others ranked it as low or very low. A sizable portion (33.33%) evaluated access as satisfactory, suggesting considerable satisfaction. 22.22% assessed access as very good, while the largest group, 38.29%, ranked it as good. In comparison to other aspects like family support or the learning environment, the majority of respondents, more than 60% perceived their access to study materials favourably overall, albeit a sizable portion still believes there is space for improvement.

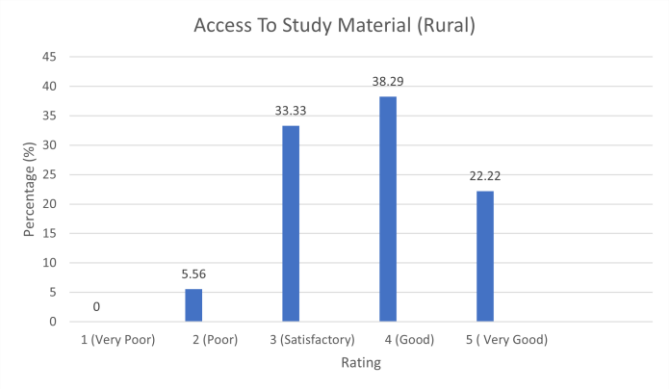


Fig. 7 Bar chart showing the percentage of respondents about their opinions on access to study material in rural areas

B. *Academic Achievement*

The pie chart in Fig. 8 shows a diverse range of educational achievement among respondents. The largest group, 38.8% out of 80 respondents, holds a bachelor's degree or equivalent, providing informed insights on academic systems and challenges. The second-largest segment, 22.5%, has a Master's degree. Other notable groups include 15% with a Diploma/IB/A-level or equivalent, 10% with SPM/IGCSE/O-level, and 8.8% who have completed Foundation/STPM/Matriculation. The data suggests that the majority of respondents are well-educated, potentially influencing the survey's perspective on education quality, access, and disparity.

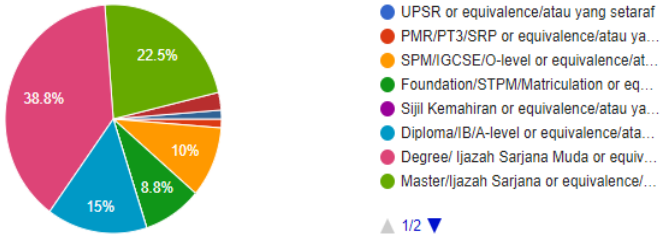


Fig. 8 Pie chart showing the percentage of respondents and their academic achievement

Mode: Based on the data, about 38.8% of the respondents currently have a degree. This figure shows the mode of the data out of 80 respondents; 31 graduated with at least a degree.

Median (MD): Since n = 80, which is even, the median lies between the (n/2)th and (n/2+1)th terms, i.e., the 40th and 41st values. Both fall in the Degree column, so the median is 'Degree'.

C. *Final Analysis*

The data collected show that all data are interconnected with one another. Central tendencies such as mean, median, mode, variance, and standard deviation are measured in this study. In addition, the use of histogram graphs helps provide precise observations of the correlations between each piece of data. Chebyshev's theorem is also implemented to validate and confirm the graphs created. It is noted that all responses to the online survey distributed are based on their life experience as students and workers. The relationship between dependent and independent variables must be identified to draw conclusions from the calculated data description. In this section, we will relate the data descriptions of independent variables with their respective dependent variables. The dependent and independent variables are the level of education and educational opportunities.

Step 1: Stating the hypotheses.

H_0: There is no significant difference in the average satisfaction with educational opportunities between rural and urban areas. H_1: There is a significant difference in average satisfaction with educational opportunities between rural and urban areas.

Step 2: Finding the critical value. Since $\alpha = 0.05$ and this test is a two-tailed test, use the z-values of –1.96 and 1.96.

Step 3: Finding the test value. Data from the two samples (urban and rural) are integrated and ranked. Positive Rank Sum = 1 + 2 + 3 + 5 + 6 + 7 = 24. Negative Rank Sum = –4. Wilcoxon Signed-Rank statistic W = 4.

https://doi.org/10.31436/ijpcc.v12i2.624

Step 4: Making a decision. Since W = 4 > critical value = 2, we reject H_0.

Step 5: Summarise the result. There is enough evidence at α = 0.05 to support the claim that there is a significant difference in average satisfaction with educational opportunities between rural and urban areas.

VI. Predictive Modelling and Feature Importance Analysis

A. Motivation and Data-Mining Pipeline

To understand more than the descriptive analysis and to understand the relative influence of different factors on academic achievement, a supervised data-mining approach was employed using Weka 3.8 [11]. Weka is a widely used open-source machine-learning platform. The analysis followed four main stages. First, responses collected through Google Forms were automatically exported in .xlsx format and converted to Weka's Attribute-Relation File Format (ARFF). Second, the dataset was prepared by defining and cleaning the attributes using Weka's Preprocess module. Third, three classification models were developed and evaluated through cross-validation using the Classify module. Finally, the importance of each attribute was assessed using the Select Attributes module, while Weka's visualisation tools were used to examine confusion matrices, receiver operating characteristic (ROC) curves, and decision-tree structures. To support transparency and reproducibility, the ARFF dataset and the corresponding Weka experiment settings are provided in the Data Availability Statement as the software artefacts associated with this study.

B. Target Variable and Features

The prediction target was defined as a binary measure of tertiary-level academic attainment, where respondents with a Degree, Master's, or PhD qualification were classified as *tertiary*, while those with SPM, Foundation, Diploma, or lower qualifications were classified as *non-tertiary*. This binary classification is widely adopted in educational inequality research. A total of ten predictor variables were included in the analysis, comprising three nominal variables (age group, gender, and urban or rural residence) and seven numerical variables measured using a five-point Likert scale. These numerical variables represented the home learning environment, school learning environment, peer influence, family emotional support, family financial support (used as a proxy for socioeconomic status), academic guidance from family, and access to study materials. The final predictor was considered an indicator of ICT access, as the corresponding survey item captured respondents' access to both digital and physical learning resources, including textbooks, lecture notes, online materials, practice questions, and educational videos.

C. Classifiers and Evaluation Protocol

Three algorithms were compared, corresponding to the reviewers' explicit suggestions: J48 (Weka's implementation of the C4.5 decision-tree algorithm, with confidence-factor pruning at 0.25), Random Forest (an ensemble of 100 unpruned trees with default bagging), and Logistic (multinomial logistic regression with ridge regularisation). All models were evaluated using 10-fold cross-validation. Reported metrics include overall classification accuracy, Cohen's Kappa, per-class precision, recall, F-measure, and the confusion matrix.

D. Results

Table I summarises the cross-validated performance of the three classifiers.

TABLE I

CROSS-VALIDATED PERFORMANCE OF THE THREE CLASSIFIERS

Classifier	Accuracy	Kappa (approx.)
J48 (Decision Tree)	71.25%	0.3826
Random Forest	75%	0.4509
Function Logistic	72.5%	0.396

RandomForest was the most accurate of the three models tested. RandomForest correctly classified 60 of the 80 instances using 10-fold stratified cross-validation, resulting in an overall accuracy of 75.0%, a Cohen's Kappa statistic of 0.451 (indicating moderate agreement beyond chance), and an area under the ROC curve of 0.794. The weighted average F-measure for the two classes was 0.748.

The confusion matrix as shown in Fig. 9 indicates that 42 of the 51 tertiary respondents were correctly identified (recall = 0.824, precision = 0.792, F = 0.808), while 18 of the 29 non-tertiary respondents were correctly identified (recall = 0.621, precision = 0.667, F = 0.643). The model performs more confidently on the majority class, as expected given the imbalance between tertiary (n = 51) and non-tertiary (n = 29) respondents. The ROC curve for the tertiary class as in Fig. 10 confirms the moderate discrimination ability with AUC = 0.7945, well above the 0.5 random-classifier baseline.

```
=== Confusion Matrix ===

  a  b   <-- classified as
 42  9 |  a = tertiary
 11 18 |  b = non_tertiary
```

Fig. 9 RandomForest confusion matrix

https://doi.org/10.31436/ijpcc.v12i2.624

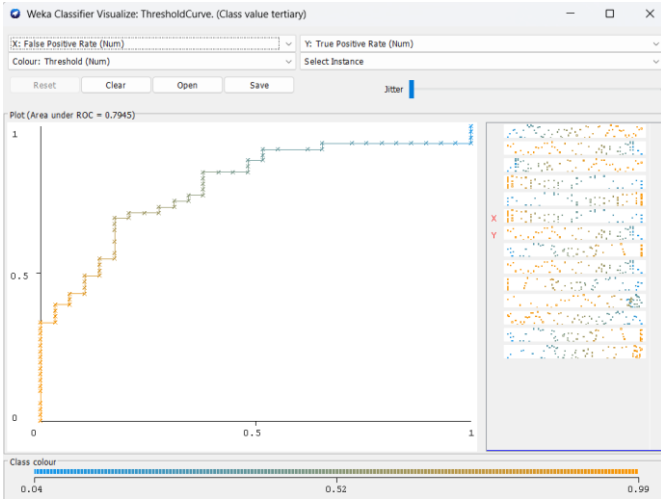


Fig. 10 ROC Curve

```
=== Run information ===

Evaluator:     weka.attributeSelection.InfoGainAttributeEval
Search:        weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:      education_disparity
Instances:     80
Attributes:    11
               age
               gender
               area
               home_env
               school_env
               peer_influence
               emotional_support
               financial_support_SES
               academic_guidance
               ict_access_study_materials
               attainment
Evaluation mode:    evaluate on all training data


=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (nominal): 11 attainment):
        Information Gain Ranking Filter

Ranked attributes:
 0.27858    1 age
 0.09654    9 academic_guidance
 0.05677    7 emotional_support
 0.0067     3 area
 0.00136    2 gender
 0          4 home_env
 0         10 ict_access_study_materials
 0          6 peer_influence
 0          8 financial_support_SES
 0          5 school_env

Selected attributes: 1,9,7,3,2,4,10,6,8,5 : 10
```

Fig. 11 Attribute-Importance Ranking

Weka's InfoGainAttributeEval with a Ranker search as illustrates in Fig. 11 produced the following information-gain values (in bits) for each predictor: age = 0.279, academic_guidance = 0.097, emotional_support = 0.057, area = 0.007, gender = 0.001, with the remaining Likert-scale predictors (home_env, ict_access_study_materials, peer_influence, financial_support_SES, and school_env) registering zero information gain under the default evaluation. Age is by far the dominant predictor, contributing nearly three times the information of academic_guidance, the next-strongest variable. Critically, urban-versus-rural area contributes only 0.007 bits of information, which was roughly 2.5% of what age contributes and less than one-tenth of academic_guidance. This finding refines and clarifies the descriptive gap reported in Section V: while the Wilcoxon Signed-Rank Test confirms that a raw urban–rural difference exists, the InfoGain analysis demonstrates that the difference is not geographic. Rather, the urban–rural gap operates almost entirely through mediating variables, which are age (which reflects accumulated years of educational opportunity) and family-level academic and emotional support.

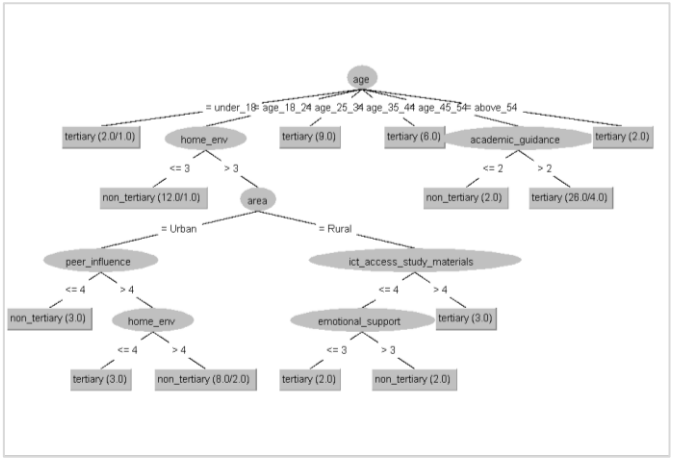


Fig. 12 J48 Decision Tree

Fig. 12 presents the J48 decision tree, which exposes the model's decision logic in human-readable form and enriches the InfoGain ranking. The root split is on age: respondents aged 25–34 and 35–44 are classified directly as tertiary attainers, capturing 15 instances at high purity. For the 45–54 bracket, the tree splits on academic_guidance, with strong family academic guidance (> 2) predicting tertiary attainment for 22 of 26 respondents (84.6% purity). The most analytically interesting subtree is the 18–24 branch, which first splits on home_env: respondents with weak home learning environments (≤ 3) are classified as non-tertiary with 92% purity. For 18–24 respondents with adequate home environments, the tree then splits on area, and within each area applies a different mediating variable. For urban respondents in this subtree, peer_influence and home environment govern the classification. For rural respondents, however, the decisive variable is ict_access_study_materials: rural young adults with strong access to study materials (> 4) are classified as tertiary attainers, while those with weaker access face a further split on emotional support. This is a substantively important

pattern for the paper's central question: for young rural respondents with adequate home learning conditions, access to ICT and study materials becomes the decisive predictor of whether they reach tertiary attainment. This directly operationalises the digital-divide dimension of educational inequality that motivated this study's ICT framing.

The combined findings from the InfoGain ranking and the J48 decision tree provide a clearer understanding of how demographic, family, and ICT-related factors influence academic attainment. Although age recorded the highest information gain (0.279 bits), it is not a factor that can be influenced through policy or intervention. Instead, age likely reflects the cumulative educational opportunities available to respondents over time. When focusing on factors that are more amenable to intervention, academic guidance from family (0.097 bits) emerged as the most influential predictor, followed by family emotional support (0.057 bits). Both variables highlight the important role of family involvement in supporting educational achievement. By comparison, area of residence contributed only 0.007 bits, making it substantially less informative than family-related factors and far less influential than age.

The J48 decision tree further supports these findings. After the initial split on age, subsequent branches were primarily determined by family-related and household characteristics rather than by urban or rural residence. Area of residence appeared only as a secondary decision point within the 18–24 age group and was further divided based on access to study materials, indicating that ICT and learning-resource accessibility had a greater influence than location itself. Overall, these results suggest that while age is the strongest predictor, it mainly represents differences in educational exposure rather than a direct causal factor. More importantly, the findings indicate that disparities in tertiary educational attainment between urban and rural populations are better explained by differences in family support and access to learning resources, including ICT, than by geographical location alone.

E. Implications for the Research Question

This predictive-modelling layer strengthens this study's contribution in three ways that are directly tied to the research question of what drives educational disparity between urban and rural Malaysia.

Firstly, it confirms and refines the Wilcoxon-based finding of urban-rural difference reported in Section V by showing that the difference is explained, rather than simply observed: the InfoGain analysis indicated that geographic location itself contributes only 0.007 bits of information, less than one-tenth of academic guidance and less than one-thirtieth of age. The urban–rural gap exists, but it operates through mediating variables, including family academic guidance, home learning environment, family emotional support, and access to study materials, rather than as an independent effect of geographical location.

Secondly, it operationalises ICT access via the study-materials indicator as a measurable predictor of academic achievement. The J48 decision tree indicates that ICT access becomes the decisive splitting variable for young rural respondents with adequate home learning conditions (Fig. 12). This provides direct model-based evidence for the education disparity dimension of the digital divide.

Lastly, the Weka pipeline itself, which includes reproducible ingestion, attribute definition, classifier training and computational visualisation, is a lightweight information-systems tool that policymakers, NGOs, or state education departments can re-run on future survey waves to track the gap over time.

Together, these findings support a shift in policy focus from geography-based interventions toward household-level supports and targeted ICT-access programmes, particularly for rural respondents in the youngest age bracket, where the tree analysis showed ICT access has the strongest discriminating effect.

CONCLUSION

The study investigates educational disparities between rural and urban Malaysian communities, with a focus on geographic location, socioeconomic background, family support, and access to educational resources. Urban students achieve better academic outcomes, largely due to superior facilities, experienced teachers, and more structured learning environments. As shown in Fig. 2, 86.48% of urban respondents agreed that the learning environment significantly influences academic performance. Family support is another key factor, with 81.18% of urban participants affirming its positive impact, as illustrated in Fig. 3. Additionally, 72.58% of respondents indicated that access to study materials is better in urban areas, as depicted in Fig. 4. In comparison, rural respondents showed slightly lower but still substantial levels of agreement: 85.78% for learning environment, 72.78% for family support, and 60.51% for access to study materials. These differences highlight the relative advantages of urban settings.

Rural students face challenges such as limited access to qualified educators, insufficient school infrastructure, and fewer academic support systems, as shown in Fig.5. However, the study also found that non-material factors, including parental involvement, personal motivation, and supportive social networks, can significantly influence academic success, as illustrated in Fig.6. Some rural students performed equally well or better than their urban

counterparts, highlighting the importance of these non-material influences.

In addition, the supervised data-mining analysis reported in Section VI confirms and clarifies this pattern, where, as discussed in Section VI.D, the strongest modifiable predictors of tertiary-level achievement are family-level variables (academic guidance, home environment, financial support) rather than urban location itself, with ICT access to study materials also making a meaningful contribution. This suggests that the urban–rural gap in Malaysian educational results is mediated by household resources and support rather than being caused by geography itself.

This study suggests that reducing the rural-urban academic gap must go beyond basic resource allocation. Policymakers should prioritise targeted investment in rural education, particularly in infrastructure development, digital learning accessibility, and teacher training. Society should also support active parental engagement and broader community efforts to promote it. Educational institutions can benefit from flexible teaching approaches that respond to the specific needs of rural learners. By recognising and addressing these factors, Malaysia can ensure that every student, regardless of location, has the opportunity to succeed academically and contribute meaningfully to the nation's development.

ACKNOWLEDGMENT

The author expresses gratitude to the respondents for their cooperation in conducting this work and appreciation for the applicable knowledge of the CSCI 1304 Probability and Statistics course offered by the Kulliyyah of Information and Communication Technology.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHOR(S) CONTRIBUTION STATEMENT

The author contributed to the study conception and design, manuscript writing, and approval of the final version.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ETHICS STATEMENT

This research did not require ethical approval.

REFERENCES

[1] L. Xiang and J. Stillwell, "Rural–Urban Educational Inequalities and Their Spatial Variations in China," *Appl. Spat. Anal. Policy*, vol. 16, no. 2, pp. 873–896, Jun. 2023, doi: 10.1007/s12061-023-09506-1.

[2] N. Ahsan, M. Shahe Emran, F. Shilpi, and W. Bank, "The Rural-Urban Divide and Intergenerational Educational Mobility in a Developing Country: Theory and Evidence from Indonesia," Oct. 2020.

[3] J. Munir, M. Faiza, B. Jamal, S. Daud, and K. Iqbal, "The Impact of Socio-economic Status on Academic Achievement," *Journal of Social Sciences Review*, vol. 3, no. 2, pp. 695–705, Jun. 2023, doi: 10.54183/jssr.v3i2.308.

[4] R. M. Wood, "A Review on Education differences in Urban and Rural Areas," *International Research Journal of Educational Research*, vol. 14, no. 2, pp. 1–3, 2023, doi: 10.14303/2141-5161.2023.254.

[5] J. Drescher, A. Podolsky, S. F. Reardon, and G. Torrance, "The Geography of Rural Educational Opportunity," *RSF*, vol. 8, no. 3, pp. 123–149, May 2022, doi: 10.7758/RSF.2022.8.3.05.

[6] C. F. Rodríguez-Hernández, E. Cascallar, and E. Kyndt, "Socio-economic status and academic performance in higher education: A systematic review," *Educ. Res. Rev.*, vol. 29, p. 100305, Feb. 2020, doi: 10.1016/j.edurev.2019.100305.

[7] T. Şengönül, "A Review of the Relationship between Parental Involvement and Children's Academic Achievement and the Role of Family Socioeconomic Status in this Relationship," *Pegem Egitim ve Ogretim Dergisi*, vol. 12, no. 2, pp. 32–57, 2022, doi: 10.47750/pegegog.12.02.04.

[8] S. P. Kawuryan, S. A. Sayuti, Aman, and S. I. A. Dwiningrum, "Teachers quality and educational equality achievements in indonesia," *International Journal of Instruction*, vol. 14, no. 2, pp. 811–830, Apr. 2021, doi: 10.29333/iji.2021.14245a.

[9] R. Van Maarseveen, "The urban-rural education gap: Do cities indeed make us smarter?," *J. Econ. Geogr.*, vol. 21, no. 5, pp. 683–714, Sep. 2021, doi: 10.1093/jeg/lbaa033.

[10] J. Lu, "Study on the Difference between Urban and Rural Education in Yuyao," *Journal of Education and Educational Research*, vol. 10, 2024.

[11] E. Frank, M. A. Hall, I. H. Witten, and M. Kaufmann, "WEKA Workbench Online Appendix for 'Data Mining: Practical Machine Learning Tools and Techniques,'" 2016.

International Journal on Perceptive and Cognitive Computing (IJPCC) Vol 12, Issue 2 (2026)
https://doi.org/10.31436/ijpcc.v12i2.712

Usability and Accessibility Evaluation of Political Party Websites in Nigeria

Solomon Adelowo Adepoju[1], Ayobami Ekundayo[2], Boyi-Musa Yakubu Lasotte[3], Oluwafemi Osho[4], Saidu Ahmed Abubakar[5], Luke-Metibaiye Oluwashina[6]

[1, 2,3,5,6]Department of Computer Science, Federal University of Technology Minna, Nigeria
[4]School of Computing Clemson University, Clemson, South Carolina, USA

*Corresponding Author: solo.adepoju@futminna.edu.ng
(Received: 16th February 2026; Accepted: 13th Mayr, 2026; Published on-line: 30th July, 2026)

Abstract— In this digital age, political party websites serve as pivotal tools for communication, information dissemination, and civic engagement. However, the effectiveness of these websites depends on their usability and accessibility, particularly in a diverse and dynamic context such as Nigeria. This study assesses the usability and accessibility of eight popular political party websites in Nigeria by using four automated tools which are AChecker, Web Accessibility Evaluation (WAVE), Test Accessible Web (TAW), and GTMetrix. The websites include All Progressives Congress (APC), Peoples Democratic Party (PDP), New Nigeria People's Party (NNPP), All Progressives Grand Alliance (APGA), Labour Party (LP), Social Democratic Party (SDP), Accord Party (Accord), and African Democratic Party (ADP). The findings, after multiple testing on the tools, revealed disparities among the websites. APC's website emerged as the most balanced, with strong HTML5/ARIA adoption and moderate structural soundness, though its performance was weakened by slow load speeds. The APGA website was the fastest and most stable, generating few errors or alerts. However, it provided almost no accessibility features. The results further show that all the websites of all the evaluated political parties have usability and accessibility issues which is a violation of Web Content Accessibility Guidelines (WCAG 2.0). Our evaluation indicates that none of the Nigerian political party websites fully comply with modern accessibility and performance standards. These findings highlight the need for web developers to carefully consider established design guidelines when developing and maintaining political party websites.

Keywords— Usability, Accessibility, Political Party Websites, WCAG2.0, Website Evaluation tools.

I. Introduction

The growing adoption of Information and Communication Technology (ICT), especially the Internet, has significantly reshaped the way organizations, including political parties' function [1]. In this digital age, political engagement increasingly dependent on online platforms, making the usability and accessibility of political party websites crucial for fostering inclusive participation [2], [3]. The websites of political parties serve as key information hubs for party policies, candidate profiles, and electoral processes.

In recent years, the internet has become an essential tool for political communication and engagement. It offers political parties an effective platform to disseminate information, mobilize supporters, and garner public backing. In Nigeria, where democracy is evolving and political participation is increasingly digitized, the usability and accessibility of political party websites are critical in shaping how citizens perceive and engage with the political process. Also, internet penetration rate in Nigeria has continued to grow and hence websites serve now serve as a vital tool to disseminate political information by political party members. electoral processes. At present there are over 100 million internet users in the country and the number is expected to increase to about 103 million by 2027 [4], [5]. With the return of the country to democracy since 1999, various political parties have emerged and merged. According to the Independent National Electoral Commission (INEC), Nigeria has 19 registered political parties and each of them is to intimate the public with their manifesto [6]. Apart from public or open campaign, one-way citizen can know more about them is via their websites. Hence, it is expected that with a click of the button vital information needed by the citizen and party member should be readily available and accessible online. In order to ensure this, the various website must comply with accessibility guidelines. However, the effectiveness of these political party websites in reaching diverse user groups, including individuals with disabilities, remains largely unexplored.

A website must be user-friendly and accessible to all categories of users including people with special needs and

older adults [7]. Everyone, including the elderly and disabled people has the right to access the Internet for using a variety of different websites [8]. The web has diverse users with each exhibiting needs based on their ability to interact, understand, perceive and navigate with web contents. For example, accessing textual information by visually impaired users differs from those with good sight [9]. The primary purpose of political party websites is to facilitate effective communication with voters, clearly articulate the party's policies, and provide compelling reasons why the party deserves electoral support. It is aimed to support each party in the direction of effective communication and services to the public or citizens via the websites [3], [10]. Unfortunately, many political party websites in Nigeria are designed primarily for representation rather than interaction the citizens. This is due to the fact that their websites are not effective, efficient and satisfactory enough to the users. They often fail to meet accessibility and usability standards, making it difficult for individuals with disabilities and those who are less tech-savvy to access information and services [11]. This lack of accessibility and usability has contributed to the marginalization of some political parties, limiting their recognition and engagement with the public [12]

Authors from various countries have carried out research on evaluating the accessibility and usability of different genre of websites in education, banking, airline, government, e-commerce and many others [7], [13], [14], [15], [16], [17], [18], [19], [20], [21]. Till date, little or no attention has been paid to political parties' websites evaluation especially in developing countries where democracy is being practiced or where political parties exist. Hence, this study evaluates the usability and accessibility of political parties' websites in Nigeria, with the aim to confirm their adherence or otherwise to international web standards such as the WCAG.

The study closely examines the design elements, navigation efficiency, readability and assistive technology compatibility of Nigerian political parties' websites. It highlights the potential barriers that may exclude certain demographics from assessing the websites. The comprehensive findings provide insights into the current digital inclusivity of Nigerian political parties and offer recommendations for improving user experience, ultimately contributing to more accessible and democratic online engagement. To the best of the authors knowledge, this paper represents the first attempt to evaluate the usability and accessibility of political websites in Nigeria.

The remaining part of the paper is structured as follows: Section II provides the review of literature focusing on website accessibility, usability and review of related works. Section III discusses the methodology while section IV and V present the discuss of results and conclusion respectively.

II. LITERATURE REVIEW

A. *Web Accessibility and Usability*

Web accessibility refers to the development of websites and digital content (e.g., multimedia, images, audio, text, video,) that diverse users, including those with disabilities can use. Accessibility ensures that all users can interact with, navigate, perceive, comprehend and contribute to the Web without encountering barriers [22], [23]. Web accessibility also benefits other people, including elderly people whose abilities have declined as a result of age [13]. Web developers must ensure that their websites are accessible according to WCAG for people with different range of abilities. These guidelines aim to eliminate barriers for individuals with visual, auditory, physical, speech, cognitive, and neurological disabilities, making web content accessible to everyone. Accessibility is aimed at eliminating any form of limitations encountered by people having any type of disability [24].

Usability, on the other hand, refers to the quality of users' experience when interacting with a system. Consequently, it is an important consideration in the design and development of websites. Website usability refers to the extent to which a website can be used to achieve specific goals, such as service delivery and user engagement, with effectiveness, efficiency, and satisfaction [23], [24], [25] . Poor usability can significantly deter user engagement, reducing satisfaction and potentially leading users to abandon a site [26]. A system design with diminished usability reduces the effectiveness of users and effectiveness, thereby diminishing the overall value of the website. Factors such as loading speed, page size, broken links, and accessibility errors can all impact usability Therefore, it is imperative for web designers to create websites that are effective, efficient, and usable by everyone, including individuals with disabilities. Failure to do so can negatively impact users' attitudes, satisfaction, and trust in the information and services offered on the website. Difficult to use websites are not appealing to users as it will often be abandoned by them. Thus, usability and accessibility are crucial components of political website design.

Political party websites are intended to be accessed and used by a diverse population, including both able-bodied users and users with disabilities. Henceforth, there is need to incorporate usability and accessibility guidelines in the design of the website. These guidelines are critical for developers to ensure that websites are accessible to all users. For example, a person with visual impairments may require screen reader technology, while color contrast is

crucial for users with blurred vision [27]. Various online accessibility tools such as AChecker and WAVE are available to evaluate website accessibility and usability [28]. These tools assist Human Computer Interaction (HCI) researchers in adhering to Web Content Accessibility Guidelines (WCAG) guidelines [29]. World Wide Web (W3C) has developed WCAG to make the web accessible to people with disabilities. The WCAG 2.0 was approved as an ISO standard. The WCAG 2.0 outlines four key principles known as POUR: Perceivable, Operable, Understandable, and Robust. These principles are categorized into three levels of conformance: Level A (basic), Level AA, and Level AAA (highest). Unfortunately, many websites are not designed and developed in accordance with these accessibility standards. Hence, it is difficult for people with some kind of disability to get complete access to the information and services available on the website [30]

The quality assessment of websites has attracted the attention of many researchers globally. Hence, numerous researchers are focused on evaluating the usability and accessibility of websites across various domains so that that digital content and services are accessible to everyone [31], [32], [33]. Some of these studies are presented in the next section

B. *Review of Related Works*

As pointed out earlier, only few studies have focused on evaluation of political parties' websites evaluation till date and they are presented as follow. Lappas and Yannas [34] proposed a framework for evaluating political party websites so that the parties will upgrade their political activities and facilitate their operation in e-democracy. It focusses on four main maturity stages comprising online presence, interaction e-democracy and integration stages. The main approach is to see how evaluation of political websites could be done but automated tools were not used in the study.

Study by Schweitzer [35] focused on evaluation of German political parties during the election year of 2002. The study provided an integrated qualitative analysis of content related, functional and formal aspect of the websites. Four parameters of accessibility, navigation, design and updating were considered in the study. It was discovered that the websites only offered informational content while neglecting interface features. However, today's websites should focus more on the ability of the users to navigate easily the information provided on the websites which can only be made possible via good interfaces.

Abdillah et al.,[36] evaluated the seventeen official political websites in Indonesia. Website evaluation questionnaire (WEQ) and GTMetrix were used for the manual and automated evaluation of the websites. Results obtained from GTMetrix revealed that a greater percentage of the websites (41.2%) got grade D, while other scores E, C, F and B with corresponding percentages of 23.5%, 17.6%11.8% and 5.9% in descending order of performance ratings respectively. This shows that most party websites need enhancements in both performance and structure in order to improve their scores.

Another study by King and Youngblood [37] focused on analysing the content and quality of county voting and election website in Alabama. A combination of automated tools (AChecker), heuristics and manual inspection were used to test for the website's usability, accessibility and mobile friendliness. The study found out that limited voting and election information are offered by these websites and more importantly, none of them are in full compliance with accessibility, usability and mobile friendliness guidelines. Crossland and Chigona [38] carried out an exploratory study that is sought to investigate the use of the Internet by political parties South Africa as well as the functionality and design of these websites. The study noted that in spite of the lower Internet accessibility rates in Africa, 72.5% of the political parties contesting the 2009 elections have websites. Also, the large political parties' websites have more functionality in their websites than the smaller ones and most of the websites lack navigation tips.

Study by Kaushik [39] which was on content analysis of political websites in India reveals that the majority of political party websites in India are building on different patterns, options and designs and have diverse information regarding their respective party history, constitution, logo, manifesto, photo gallery, current events, speeches, contact, and web 2.0 applications. Further results obtained from the study confirmed that some political party websites do not have essential contents such as date of updating, feedback/suggestion option, party membership information, option for disabled people, mobile applications, public forum, and soon. 8 political parties.

Another study Mochla and Tsourvakaswas [2] conducted to investigate the quality dimensions of the Greek Parliaments' political parties' websites in an off-campaign period, one year before the legislative elections of 2019. The hypotheses used in the research were based on political marketing background theory with emphasis on variables from different eservice quality models and effective web-design dimensions. Using a sample of 400 voters to make the survey, it was observed that the useful content and aesthetic design were the most significant dimensions of the voters' perception about the political websites' quality.

Political parties use websites it for a variety of purposes, including relationships with its supporters and potential voters. Hence, the website of any political party is an important channel of communication. Such a website

https://doi.org/10.31436/ijpcc.v12i2.712

should be user-friendly to various audiences. Sajna-Kosobucka and Sajna-Kunowsky [40] in their study assessed the quality of information architecture (IA) on the websites of the two major American political parties: the Republican Party and the Democratic Party. By using triangulation methods; a qualitative-heuristic assessment of the IA on the websites, expert assessment and a comparative analysis were conducted. It was observed that there was noticeable lack of some important components of the IA on the observed websites. Also, the website of the Republican Party is slightly better in terms of quality than the website of Democratic Party by 14.09%.

Table 1
POLITICAL PARTY WEBSITE URLS

Political Party	Acrony	URL
All Progressives	APC	www.apc.com.ng
Peoples Democratic	PDP	https://peoplesdemocratic
New Nigeria	NNPP	https://nnpp.net.ng
All Progressives	APGA	https://apga.org.ng
Labour Party	LP	https://labourparty.com.ng
Social Democratic	SDP	https://sdpnigeria.org
Accord Party	ACCORD	https://accordnational.com.
African Democratic	ADC	https://adc.org.ng

As it could be observed from the review above, most of the studies did not involve the use of automated tools to carry out the evaluation of websites and hence research a gap is created which is the essence this paper.

III. METHODOLOGY

This study selected only political parties whose websites are accessible at the time of carrying out this research though the country has nineteen registered political parties. The name and URL of the Political Parties website used in the study are shown in Table 1. Eight political parties who have functional websites were used for the study and the evaluation was done using four online automated tools: AChecker, WAVE, TAW and GTMetrix. The use of multiple automatic tools is necessary so as to get comprehensive reports, remove biases and ensure accuracy and comprehensiveness [18], [41]. Also, some tools could substitute deficiencies of other tools, making the evaluation more reliable. The URL of each website is entered into webpage of each automated tool which eventually evaluates the site and generated results. The evaluation is done repeatedly about three times over several weeks to ensure correctness and consistency of the results obtained. Online automated tools are evaluation software used to assess and measure the usability and accessibility of a website. These tools are accessible resources for any developer who wants to ensure that the site they develop conform to established accessibility standards. Most of the online automated tools evaluate web sites against accessibility guidelines. These include the Section 508, WCAG 1.0 and 2.0, the Stanca Act (Italian Accessibility Law), W3C's, BITV (web accessibility test of the German BIK project), RGAA (French Accessibility Law), or a combination of them. Some tools also have their own accessibility guidelines while some require the URL. Others require upload of HTML file or pasting of the HTML code. The tools are all mostly web-based hence, no need for download or installation.

The AChecker tool is an open-source web accessibility evaluation tool that checks web pages for compliance with various accessibility standards, such as WCAG 2.0. It provides detailed reports on issues and suggests ways to fix them. The results of this tool identify three categories of accessibility problems at each level: known, likely, and potential. Known Problems are confirmed accessibility failures (must be fixed), Likely Problems are strong indicators of problems but need manual confirmation and Potential Problems are Possible issues flagged by tools, require human judgment [42].

Web Accessibility Evaluation Tool (WAVE) helps users identify accessibility issues on web pages. It visually displays the results directly on the webpage, highlighting areas that need improvement according to WCAG guidelines. The WAVE tool evaluation produces a report detailing errors, alerts, features, contrast errors, structural element issues, and the use of HTML5 and ARIA within a webpage, According to WCAG. The WAVE helps its users better evaluate the things that automatic tools cannot check [43]. TAW (Test de Accesibilidad Web) tool is a Spanish web accessibility evaluation tool that analyzes web pages for compliance with accessibility standards, mainly focusing on WCAG 2.0. It generates a summary of an error on the analyzed page with suggestions for improvements. The TAW clearly marks the accessibility violations that it discovers by providing an annotated version of the website as well as recommendations on how to resolve them. This tool uses four principles such as perceivable, operable, understandable, and robust to check a website. TAW makes sure that information and interface components are well presented to the user in ways they can perceive. It checks how operable user interface components and navigation are. It checks how understandable the information and operation of the user interface are and checks the content of the site to ensure it can be interpreted enough [44]

GTMetrix is a speed test and performance monitoring tool that analyzes a website's page speed (structure) and performance using Google's Lighthouse. The tool focuses on user-centric metric provides a full report on a site's performance. It analyses a site's load time, web page size, i.e., the overall page size, which includes images, videos, and

https://doi.org/10.31436/ijpcc.v12i2.712

HTML code and the number of pages in the website, and then generate a score with recommendations to improve it [45]

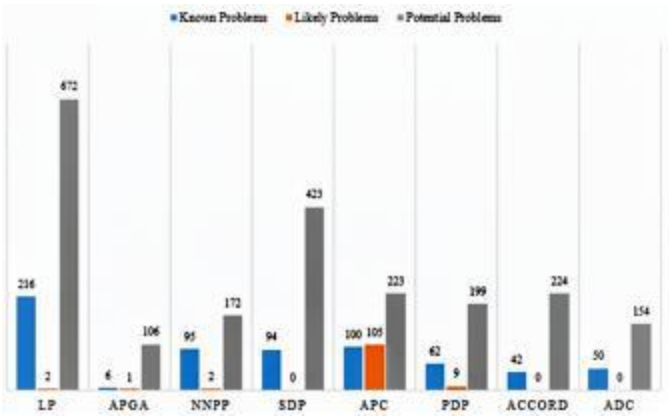


Fig 1: AChecker analysis and Results

IV. RESULTS AND DISCUSSION

The results of the accessibility and usability evaluation of the political websites in Nigeria are presented as follows.

A. *AChecker Results*

The analysis of the AChecker testing to the selected political websites against WCAG 2.0 is presented in Fig 1. It shows the known, likely, and potential errors represented as A, AA, and AAA conformance levels of WCAG 2.0.

From the results obtained, LP has a high number of known Problems which indicate a very high number of confirmed issues. The Likely Problems is almost negligible and the potential Problems is extremely high. The large number of known and potential problems suggests poor implementation or design flaws. It shows the site with ambitious accessibility intent but plagued by execution gaps. Users with disabilities may encounter severe barriers, reducing trust in the party's digital inclusivity.

APGA has few Known Problems which is very low confirmed issues, a negligible Likely Problems and low-to-moderate Potential Problems. Overall, it shows very poor ARIA/HTML5 adoption which implies lack of accessibility features rather than good design. The absence of known issues might reflect simplistic design rather than compliance. APGA's website risks being technically outdated and inaccessible, despite fewer flagged issues.

NNPP has high confirmed Known Problems issues, negligible Likely Problems and a moderate Potential Problems. NNPP faces substantial confirmed Problems, which directly affect users. With moderate potential issues, it shows structural weaknesses that may worsen over time if not fixed. Despite good visual readability (from first dataset), NNPP struggles with underlying technical issues.

SDP has a high confirmed Known Problems issue, no Likely Problems and very high potential problems. This implies SDP's site demonstrates systemic weaknesses: many confirmed and potential problems. It has combined with poor readability, this creates a very unfriendly user experience and the lack of "likely problems" suggests problems are either obvious (known) or deeply systemic (potential).

APC has a high confirmed issues Known Problems, a very high suspected issues of Likely Problems and a moderate Potential Problems. This implies that APC stands out because of very high "likely problems", indicating unstable or inconsistent implementation. This suggests sections of the site may behave unpredictably for users with assistive technologies. PDP has moderate confirm issues Known Problems, small but present Likely Problems and moderate Potential Problems. This shows that PDP's site has manageable issues compared to others, and combined with good visual readability, it may feel accessible to most users. The moderate problem levels suggest PDP is "average" but not forward-looking in accessibility.

ACCORD has relatively low Known Problems, no Likely Problems and moderate Potential Problems. ACCORD showing fewer known issues, suggesting a fairly clean website. However, potential problems are significant, meaning the site may not scale well or may fail under more complex accessibility audits. A risk-prone but functional site. ADC has few relatively Known Problems, 0 Likely Problems and moderate Potential Problems. This implies that ADC has a relatively clean profile, but the large number of potential problems suggests underlying structural weaknesses. When combined with poor contrast errors (from earlier dataset), the site still fails to deliver effective accessibility.

From the results above, the most problematic website is LP which has highest problems overall, especially potential problems (672), SDP with high confirmed and potential issues and APC is high across all categories, especially "likely problems." The relatively Cleaner Websites are APGA, ACCORD, ADC which fewer known issues, but this could reflect simplistic design rather than true accessibility and PDP with balanced, manageable issues. APC is imbalanced due to unusually high "likely problems," suggesting fragile accessibility. LP and SDP are structurally weak, both confirmed and potential. APGA has low issues but low accessibility adoption implies "false sense of compliance." Broader Implications is that of digital Inclusivity where high numbers of known and potential problems indicate structural neglect of accessibility. Voters with disabilities (visual, cognitive, motor) will face barriers

https://doi.org/10.31436/ijpcc.v12i2.712

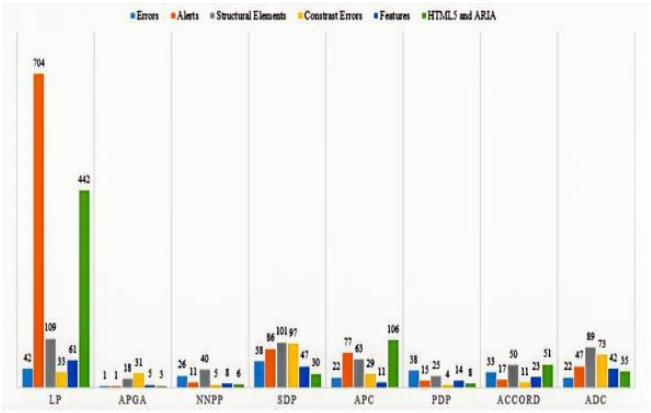


Fig 2 WAVE Analysis Tool Results

B. WAVE Results

WAVE compares the websites across six categories: Errors, Alerts, Structural elements, Contrast errors, Features and HTML5 & ARIA usage. The results are presented in Fig 2. Errors (Code/Validation Issues) is lowest in APGA. It shows almost no errors might reflect oversimplified design rather than quality. High error counts in LP, SDP, PDP suggest poorly validated HTML/CSS. This may cause usability and rendering inconsistencies. Alerts (Warnings for Accessibility/Code Quality) is highest for LP (704) which is abnormally high and moderately high for SDP (86), APC (77), ADC (47). It is Lowest for APGA (1). Alerts are not always fatal but indicate risky design practices. LP's 704 suggests very poor accessibility coding, while APGA's "1" may reflect minimal coding effort.

Structural Elements (Headings, Landmarks, etc.) is highest in LP (109), SDP (101), ADC (89) and lowest for APGA (18), PDP (25). A higher number usually means better content structuring (use of <header>, <main>, <nav>, <section>). LP and SDP look strong here, while APGA and PDP are structurally weak. Contrast Errors (Readability Issues) is highest for SDP (97) and ADC (73), lowest for PDP (4) and NNPP (5). High contrast errors make text difficult to read, especially for visually impaired users. PDP and NNPP websites are most readable, while SDP and ADC fail in basic accessibility. Features (Accessibility Enhancements) is highest for LP (61), Moderately high for SDP (47), ADC (42) and Lowest for APGA (5), NNPP (8). It implies that LP invested most in accessibility features, while APGA and NNPP neglected them.

HTML5 & ARIA (Modern Accessibility Standards) is highest in LP (442), APC (106), Moderate in ACCORD (51), ADC (35) and Lowest APGA (3), NNPP (6), PDP (8). Interpretation: LP and APC show strong commitment to modern accessibility coding. APGA, NNPP, and PDP websites are critically deficient in ARIA integration, meaning their sites are less accessible to assistive technologies like screen readers.

The best technical accessibility is LP (high ARIA adoption), though undone by errors, most balanced overall is APC (strong ARIA + moderate readability), Most readable are PDP and NNPP (very low contrast errors), the weakest is APGA (minimal structure, ARIA, and features which make it inaccessible) and worst readability are SDP and ADC (contrast failures make text hard to read). Table 2 shows the ranking of the websites based on the WAVE tools analysis.

C. *TAW Results*

The results for this are presented as problems, warnings and not reviewed. These are presented in figures 3, 4 and 5. Based on the results obtained, LP has 3 very low perceivable problems and 2 minimal warnings. LP performs well in WCAG compliance despite earlier datasets showing many technical issues. APGA shows minimal ARIA/HTML5 adoption. The low problems may reflect oversimplified site design rather than accessibility leadership. Risks being functionally inaccessible to screen-reader users despite low WCAG issues.

NNPP faces real accessibility failures across all WCAG categories. Users with disabilities will face barriers in both perception (contrast, images) and operation (keyboard/mouse use). This contradicts its fast performance earlier, showing speed does not equal inclusivity. SDP is one of the worst performers as users cannot easily see or operate the content. This aligns with earlier poor contrast and stability findings. It represents systemic exclusion of disabled users.

TABLE II
Final Ranking Based on Wave

Rank	Party website	Strengths	Weaknesses	Verdict
1st	**APC**	Strong ARIA (106), good structure	Moderate alerts, contrast errors	Best overall balance
2nd	**LP**	Highest ARIA (442), best structure	Too many alerts/errors	Ambitious but flawed
3rd	**PDP**	Most readable (contrast = 4)	Weak ARIA & structure	Readable but weak technically
4th	**ACCORD**	Moderate ARIA & features	Errors & contrast issues	Mid-level performer
5th	**NNPP**	Very readable (contrast = 5)	Very poor ARIA	Minimalist but accessible visually
6th	**ADC**	Strong structure (89)	Poor readability, many alerts	Structured but unreadable
7th	**SDP**	Strong structure (101)	Worst contrast (97), many errors	Organized but inaccessible
8th	**APGA**	Few errors/alerts	Almost no accessibility features	Oversimplified & exclusionary

https://doi.org/10.31436/ijpcc.v12i2.712

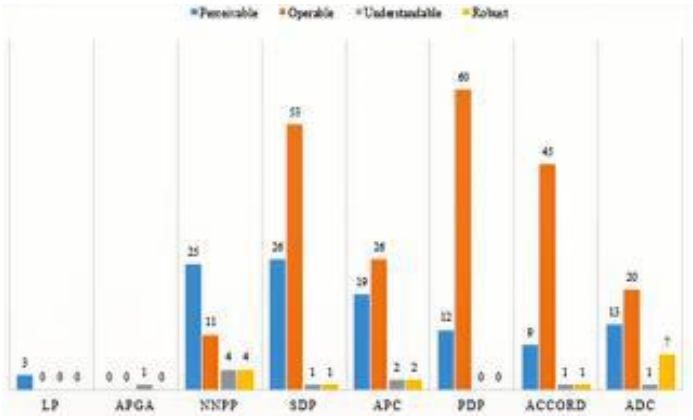


Fig 3 TAW Problems Analysis

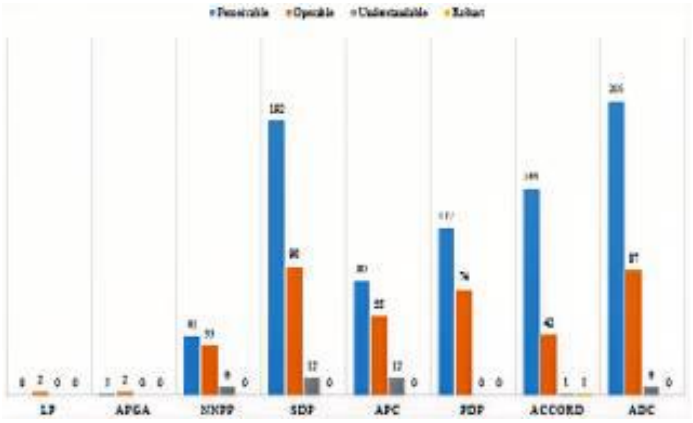


Fig 4: TAW Warning Analysis

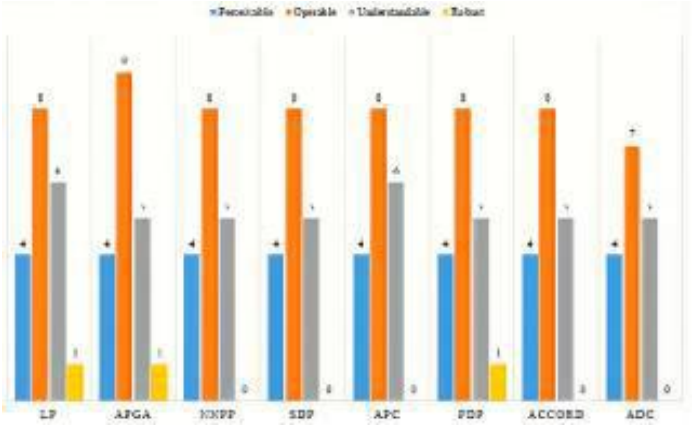


Fig 5: TAW Not reviewed Analysis

APC shows considerable accessibility warnings, meaning potential risks that could affect inclusivity. It reflects a site that tries to implement ARIA/HTML5 but struggles with execution quality. Must prioritize operability improvements (keyboard support, navigation). PDP is heavily operability-impaired (keyboard/mouse navigation issues). Users with motor impairments or those who rely on keyboard-only navigation will be excluded. It strongly reflects design without inclusivity testing.

ACCORD is visually unfriendly (contrast, images) and difficult to operate. Even though known problems are fewer than PDP/SDP, warnings suggest hidden risks. It needs urgent redesign for readability and navigation. ADC fails in both visual accessibility and operation. Robust issues mean it is also weak with assistive technologies which implies that screen readers may not work well. This further confirm ADC's poor performance and instability from earlier datasets.

LP and APGA are the best in WCAG Compliance (low problems) SDP, PDP ACCORD and NPPP are the worst in Problems (confirmed issues), while ADC, SDP and PDP are the worst in Warnings (potential risks) as the systemic risks that could worsen under different devices. Also, LP is technically compliant but undermined by earlier speed/stability issues. APGA has "low problems" due to simple design, not inclusivity leadership. SDP, PDP, ACCORD and ADC have serious accessibility barriers which implies that these are digitally exclusive websites.

APC is a moderate performer but must fix operability warnings to remain credible. Other observation is that in perceivability issues (SDP, ADC, ACCORD) visually impaired users cannot access text, images, or video. PDP, ACCORD and APC have Operability issues so keyboard-only users and those with motor impairments are excluded. NNPP, APC and ADC have understandability issues leading to confusing navigation and unclear content affect cognitive accessibility. ADC and NNPP also have Robustness issues which leads to poor compatibility with screen readers reduces inclusivity for disabled citizens.

For political parties, this is not just technical — it has democratic consequences as it can lead to exclusion of vulnerable groups from political participation. These

https://doi.org/10.31436/ijpcc.v12i2.712

TABLE III
FINAL RANKINGS FROM GTMETRIX RESULTS

Rank	Party	LCP (s)	TBT (ms)	CLS	Grade	Verdict
1st	**APGA**	1.1	0	0	A	Fastest and most stable site
2nd	**NNPP**	1.5	1000	0.07	B	Quick load, minor blocking
3rd	**ADC**	2.8	409,000	0.16	F	Structure fair, but catastrophic blocking time
4th	**PDP**	3.3	27,000	0.01	D	Stable layout, but very slow and unresponsive
5th	**APC**	3.2	1,150	0.31	F	Reasonable LCP, but high CLS and poor grade
6th	**SDP**	2.4	7,110	0.33	F	Structured but unstable and unresponsive
7th	**ACCORD**	11.5	1,500	0.42	F	Very slow load and unstable
8th	**LP**	17.7	2,250	0.44	F	Catastrophic load time, poorest perf

defects also show digital incompetence and unprofessionalism. Also, it limits global credibility as accessibility standards (WCAG, ADA, EU directives) are becoming universal benchmarks.

C. GTMetrix Results

In GTMetrix evaluation, the following metrics are used. Largest Contentful Paint (LCP) which measures how long it takes for the largest content element (i.e., a hero image or heading text) on the page to become visible within the visitors' viewport. It measures how long it takes for the main page to load. An LCP of 1.2 seconds or less is the best. LCD < 2.5s is good, while LCP of between 2.5 and 4s is moderate while LCP > 4s is poor.

Total Blocking Time (TBT) in seconds is a measure of how long scripts block user interaction. The lower the better, good value is usually < 0.2s. TBT indicate how much time is blocked by scripts during the page loading process. A TBT of 150 milliseconds or less is fine.

Cumulative Layout Shift (CLS) measures visual stability of the website while loading. That is how much layout shift is experienced as the web pages load. CLS of < 0.1 is good. CLS of 0.1–0.25 means improvement is needed while CLS > 0.25 is poor. Structure Score is the proprietary assessment of Lighthouse and Custom GTMetrix audits. It represents how well the page is built for optimal performance.

From the results shown in Table 3, LP with LCP of 17.7s is extremely poor (main content very slow to load). TBT of 2.25s implies high script blocking, poor interactivity, CLS of 0.44 means unstable visuals (content shifts) and Implications are that users face long waits, instability, and unresponsive pages which leads to very poor user experience. Accessibility suffers since long delays frustrate users with disabilities. It also damages credibility and engagement as people may abandon the site. APGA, LCP1.1s is excellent, TBT of 0 means no blocking scripts, CLS of 0 is perfect visual stability. APGA's site is highly optimized for speed and stability. However, earlier analyses showed lack of HTML5/ARIA support, meaning performance is great but accessibility features are missing. Good technical performance but excludes disabled users.

NNPP with LCP of 1.5s implies excellent, TBT of 1s means a bit high, but manageable and CLS of 0.07 is good stability. It is a fast, stable and user-friendly site. Slightly heavier scripts affect interactivity, but overall, a good experience. Better balance between performance and usability than most parties. SDP is LCP 2.4s is acceptable, TBT is 7.11s is very high script blocking which implies delays interactivity. CLS of 0.33 means poor stability. Page loads content at first but becomes unresponsive for several seconds. Users may experience click delays and shifting layouts thus leading to a frustrating experience. It also shows weak optimization and poor-quality assurance.

APC is LCP is 3.2s is moderate which is slightly slow, TBT of 1.15s high as it affects usability. But the CLS 0.31 is poor stability. Despite strong ARIA usage (from earlier analysis), APC's page performance is failing. This mismatch suggests heavy or unoptimized content/scripts. Accessibility improvements are undermined by poor speed and stability.

PDP LCP of 3.3s is moderate, TBT of 27s is extremely poor (page unresponsive for a long time). CLS of 0.01 is excellent stability. The site is visually stable but unresponsive for interaction due to blocking scripts. Creates frustration for users on slower devices. PDP must optimize JavaScript for accessibility and performance. ACCORD LCP is 11.5s which is extremely poor, TBT of 1.5s is too high and CLS of 0.42 is poor stability. It is one of the slowest and most unstable sites. This means there is poor optimization which implies that many users (especially on mobile) will abandon it. Damages digital image and limits reach. ADC with LCP of 2.8s is moderate, TBT of 409s is catastrophic which make the site essentially unresponsive. With CLS of 0.16 an improvement is needed. The extremely high TBT makes the site practically unusable. Even if content loads, users cannot interact. Suggests deep technical flaws or server misconfigurations.

In overall APGA is the fastest and most stable. NNPP is also fast, with slight blocking. PDP has stable visuals but unresponsive. On the other hand, LP, SDP, APC, ACCORD,

ADC websites are slow, unstable, and unresponsive. LP and ACCORD are the worst and ADC is catastrophic. ADC is almost unusable due to interactivity failures. APGA is good speed but poor accessibility features (previous dataset). APC and LP are strong ARIA adoption but fail in performance, undermining inclusivity.

Sites with F grades risk very high bounce rates, especially on mobile which lead to users' abandonment. Also, slow, unstable sites make parties look unprofessional and unprepared for modern digital engagement. Poor performance disproportionately harms disabled users, older voters, and those on low-bandwidth networks. The overall results for this are shown in Table 3 with APGA topping the table and LP coming last under this tool. Generally, most Nigerian political party websites fail web performance standards, making them inaccessible and unengaging for large sections of the population.

These findings are consistent with prior research that stresses the incompleteness of any single automated tool and the need for complementary approaches. Vigo et al., [46] found that different automated accessibility checkers capture distinct subsets of WCAG success criteria, with low overlap. Similarly, Alajarmeh [9], Vigo and González-mora et al. [47] noted that automated evaluation must be supplemented with manual inspection to capture semantic, contextual, and assistive-technology issues that automated tools cannot detect.

V. Conclusion

The increasing reliance on web technologies has placed accessibility, usability, and performance at the forefront of political party communication. This study assesses the usability and accessibility of eight political party websites in Nigeria, employing four automated tools: AChecker, Web Accessibility Evaluation (WAVE), Test Accessible Web (TAW) and GTMetrix.

The tools used in the study provide complementary insights: AChecker offers systematic validation against WCAG guidelines with classifications of known, likely, and potential problems; WAVE provides in-page visual overlays that highlight accessibility features and violations; TAW generates detailed WCAG-based compliance reports; and GTMetrix (powered by Google Lighthouse) measures performance metrics such as Largest Contentful Paint (LCP), Total Blocking Time (TBT), and Cumulative Layout Shift (CLS) while flagging accessibility-related issues linked to responsiveness and stability

The findings reveal significant disparities across the websites. APC emerged as the most balanced, with strong HTML5/ARIA adoption and moderate structural soundness, though its performance was weakened by slow load speeds (LCP 3.2s, CLS 0.31). PDP excelled in readability, recording the lowest contrast errors (4) and stable layouts, but lacked accessibility features and suffered from responsiveness failures (TBT 27s). LP demonstrated the highest ambition in accessibility adoption (442 ARIA elements, 61 features) but was undermined by excessive alerts (704) and poor optimization (LCP 17.7s, F grade). NNPP and APGA prioritized speed and readability, with fast LCP values (1.5s and 1.1s respectively), but both offered minimal accessibility support (ARIA ≤ 6), excluding users who rely on assistive technologies. ACCORD and ADC performed poorly overall, with ADC's catastrophic TBT of 409 seconds rendering the site virtually unusable. SDP demonstrated solid structure (101) but had the worst contrast errors (97) and numerous WCAG failures, making it systemically inaccessible.

WebAIM [43] emphasizes that contrast errors and structural issues strongly correlate with user-reported barriers, while GTMetrix documentation highlights that performance bottlenecks (long TBT, unstable CLS) directly undermine accessibility by delaying interaction readiness. Overall, the evaluation shows that no Nigerian political party website fully meets modern accessibility and performance standards.

ACKNOWLEDGMENT

The authors appreciate the efforts of the Federal University of Technology Minna for creating enabling to carry out this research

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHOR(S) CONTRIBUTION STATEMENT

All the authors contributed one way or the other in the preparation, drafting, reading, analysis, writing and proof reading the manuscript.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ETHICS STATEMENT

This study did not require ethical approval.

REFERENCES

[1] E. N. Asiimwe and N. Lim, "Usability of Government Websites in Uganda," Electronic Journal of e-Government, vol. 8, no. 1, pp. 1–12, 2010.

[2] V. Mochla and G. Tsourvakas, "Quality Dimensions of Political Parties' Website Services That Satisfy Voters in Political Marketing," Journal of Political Marketing, vol. 19, no. 0, pp. 1–13, 2020, doi: 10.1080/15377857.2020.1790470.

[3] P. Correa, J. Rodríguez-Teruel, and O. Barberà, "The Use of Party Websites for Political Information Among Spanish Party Activists," Party Politics, vol. 31, no. 2, pp. 239–250, 2025, doi: 10.1177/13540688241306729.

[4] Statista, "Internet user reach in Nigeria." [Online]. Available: https://www.statista.com/. [Accessed: Sept . 10, 2025].
[5] Nigerian Communications Commission, "Subscriber Statistics." [Online]. Available: https://www.ncc.gov.ng/stakeholder/statistics-reports/subscriber-data#quarterly-subscriber-operator-data. [Accessed: June. 17, 2026].
[6] Independent National Electoral Commission, "List of Political Parties." [Online]. Available: https://www.inecnigeria.org/list-of-political-parties/. [Accessed: June. 17, 2026].
[7] F. Kiran, N. Z. Bawany, and M. Bukhari, "Usability and Accessibility Evaluation of Banking Websites," in Proc. 2020 International Conference on Advanced Computer Science and Information Systems (ICACSIS), 2020, pp. 247–256, doi: 10.1109/ICACSIS51025.2020.9263083.
[8] P. L. P. Rau, L. Zhou, N. Sun, and R. Zhong, "Evaluation of Web Accessibility in China: Changes From 2009 to 2013," Universal Access in the Information Society, vol. 15, no. 2, pp. 297–303, 2016, doi: 10.1007/s10209-014-0385-9.
[9] N. Alajarmeh, "Evaluating the Accessibility of Public Health Websites: An Exploratory Cross-Country Study," Universal Access in the Information Society, vol. 21, no. 3, pp. 771–789, 2022, doi: 10.1007/s10209-020-00788-7.
[10] R. K. Gibson, M. Margolis, D. Resnick, and S. J. Ward, "Election Campaigning on the WWW in the USA and UK: A Comparative Analysis," Party Politics, vol. 9, no. 1, pp. 47–75, 2003, doi: 10.1177/135406880391004.
[11] M. K. Nasir, A. Awais, and B. Muhammad, "Accessibility Evaluation of Banks' Websites in Pakistan for Disabled Users," in Proc. 1st International Conference on Technology Helping People with Special Needs (ICTHP), Al-Imam University, Saudi Arabia, 2013. [Online]. Available: https://www.researchgate.net/publication/276063204_Accessibility_Evaluation_of_banks_website.
[12] S. Kruikemeier, A. P. Aparaschivei, H. Boomgaarden, G. van Noort, and R. Vliegenthart, "Party and Candidate Websites: A Comparative Explanatory Analysis," Mass Communication and Society, vol. 18, no. 6, pp. 1–22, 2015, doi: 10.1080/15205436.2015.1051233.
[13] M. Campoverde-Molina, S. Luján-Mora, and L. V. García, "Empirical Studies on Web Accessibility of Educational Websites: A Systematic Literature Review," IEEE Access, vol. 8, pp. 91676–91700, 2020, doi: 10.1109/ACCESS.2020.2994288.
[14] S. Gopinath, V. Senthooran, N. Lojenaa, and T. Kartheeswaran, "Usability and Accessibility Analysis of Selected Government Websites in Sri Lanka," in Proc. IEEE Region 10 Symposium (TENSYMP), 2016, doi: 10.1109/TENCONSpring.2016.7519439.
[15] B. Wentz, D. Pham, E. Feaser, D. Smith, J. Smith, and A. Wilson, "Documenting the Accessibility of 100 U.S. Bank and Finance Websites," Universal Access in the Information Society, vol. 18, no. 4, pp. 1–15, 2019, doi: 10.1007/s10209-018-0616-6.
[16] M. Bilal, Z. Yu, S. Song, and C. Wang, "Evaluate Accessibility and Usability Issues of Particular China and Pakistan Government Websites," in Proc. 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD), Chengdu, China, 2019, pp. 316–322, doi: 10.1109/ICAIBD.2019.8836950.
[17] S. A. Adepoju, A. Ekundayo, A. Ojerinde, and R. Ahmed, "Usability and Accessibility Evaluation of Nigerian Mobile Network Operators' Websites," in Proc. 2nd International Conference of the IEEE Nigeria Computer Chapter (NigeriaComputConf), Abuja, Nigeria, 2019, doi: 10.1109/NigeriaComputConf45974.2019.8949622.
[18] P. Dangol, "Website Accessibility Evaluation of the Federal Government of Nepal," Universal Access in the Information Society, vol. 23, pp. 1–21, 2024, doi: 10.1007/s10209-023-01076-w.
[19] L. Hasan and A. Morris, "Usability Problem Areas on Key International and Key Arab E-Commerce Websites," Journal of Internet Commerce, vol. 16, no. 1, pp. 80–103, 2017, doi: 10.1080/15332861.2017.1281706.
[20] Y. Bai and D. P. Bellisario, "The Relationship Between Website Accessibility and Usability: An Examination of U.S. County Government Online Portals," Electronic Journal of e-Government, vol. 17, no. 1, pp. 47–62, 2019. [Online]. Available: http://www.ejeg.com.
[21] S. Paul and S. Das, "Accessibility and Usability Analysis of Indian E-Government Websites," Universal Access in the Information Society, vol. 19, no. 4, pp. 949–957, 2020, doi: 10.1007/s10209-019-00704-8.
[22] S. Kumar, J. Shree D. V., and P. Biswas, "Comparing Ten WCAG Tools for Accessibility Evaluation of Websites," Technology and Disability, vol. 33, no. 3, pp. 163–185, 2021, doi: 10.3233/TAD-210329.
[23] I. O. Oyefolahan, A. A. Sule, S. A. Adepoju, and F. Babakano, "Keeping with the Global Trends: An Evaluation of Accessibility and Usability of Nigerian Banks' Websites," International Journal of Information Engineering and Electronic Business, vol. 11, no. 2, pp. 44–53, 2019, doi: 10.5815/ijieeb.2019.02.06.
[24] S. G. Abduganiev, "Towards Automated Web Accessibility Evaluation: A Comparative Study," International Journal of Information Technology and Computer Science, vol. 9, no. 3, pp. 18–44, 2017, doi: 10.5815/ijitcs.2017.09.03.
[25] V. Venkatesh, H. Hoehle, and R. Aljafari, "A Usability Study of the Obamacare Website: Evaluation and Recommendations," Government Information Quarterly, vol. 34, no. 2, pp. 203–221, 2017, doi: 10.1016/j.giq.2017.01.003.
[26] B. Csontos and I. Heckl, "Accessibility, Usability, and Security Evaluation of Hungarian Government Websites," Universal Access in the Information Society, vol. 20, no. 1, pp. 139–156, 2021, doi: 10.1007/s10209-020-00716-9.
[27] S. A. Adepoju and I. S. Shehu, "Usability Evaluation of Academic Websites Using Automated Tools," in Proc. 3rd International Conference on User Science and Engineering (i-USEr), Shah Alam, Malaysia, 2014, pp. 26–31, doi: 10.1109/IUSER.2014.7002700.
[28] J. Abascal, M. Arrue, and X. Valencia, "Tools for Web Accessibility Evaluation," in Web Accessibility: A Foundation for Research, London, U.K.: Springer, 2019, doi: 10.1007/978-1-4471-7440-0.
[29] A. Alsaeedi, "Comparing Web Accessibility Evaluation Tools and Evaluating the Accessibility of Webpages: Proposed Frameworks," Information, vol. 11, no. 1, Art. no. 40, 2020, doi: 10.3390/info11010040.
[30] Z. Jano, A. R. Abdul Ghani, M. M. Atan, et al., "Website Usability and Cultural Dimensions in Malaysian and Australian Universities," Asian Social Science, vol. 11, no. 9, pp. 1–10, 2015, doi: 10.5539/ass.v11n9p1.
[31] G. Agrawal, D. Kumar, and M. Singh, "Assessing the usability, accessibility, and mobile readiness of e-government websites: A case study in India," Universal Access in the Information Society, vol. 21, no. 3, pp. 737–748, 2022, doi: 10.1007/s10209-021-00800-8.
[32] H. Nazar, M. S. Sarfraz, and U. Shoaib, "Web accessibility evaluation of banking websites in Pakistan," International Journal of Computer Science and Information Security, vol. 15, no. 1, pp. 642–650, 2017. [Online]. Available: https://www.academia.edu/download/55897434/Journal_of_Computer_Science_IJCSIS_January_2017_Part_II.pdf
[33] J. Ara, C. Sik, and L. Arpad, "Accessibility engineering in web evaluation process: A systematic literature review," Universal Access in the Information Society, vol. 23, no. 2, pp. 1–23, 2024, doi: 10.1007/s10209-023-00967-2.
[34] G. Lappas and P. Yannas, "A framework to evaluate political party websites," in Proc. 4th International Conference on Politics and Information Systems: Technologies and Applications, 2006, pp. 226–231.
[35] E. J. Schweitzer, "Election campaigning online: German party websites in the 2002 national elections," European Journal of Communication, vol. 20, no. 3, pp. 327–351, 2005, doi: 10.1177/0267323105055261.
[36] R. Abdillah, R. Riski, and U. Asfari, "Indonesian political party website portal evaluation rank based on manual and automated usability

score," SISTEM: Jurnal Sistem Informasi, vol. 14, no. 1, pp. 54–65, 2025.

[37] B. A. King and N. E. Youngblood, "E-government in Alabama: An analysis of county voting and election website content, usability, accessibility, and mobile readiness," Government Information Quarterly, vol. 33, no. 4, pp. 715–726, 2016, doi: 10.1016/j.giq.2016.09.001.

[38] M. Crossland and W. Chigona, "An evaluation of the functionality and delivery of websites of political parties in South Africa," South African Journal of Information Management, vol. 12, no. 1, pp. 1–7, 2010, doi: 10.4102/sajim.v12i1.453.

[39] A. Kaushik, "Content analysis of political party websites in India," World Digital Libraries: An International Journal, vol. 8, no. 1, pp. 13–24, 2015, doi: 10.18329/09757597/2015/8102.

[40] P. Sajna-Kosobucka and R. Sajna-Kunowsky, "Information architecture of the websites of major American political parties," Journal of Information Science, 2023, doi: 10.1177/01655515231163367.

[41] P. Bokingkito, J. Beloy, J. J. Ecleo, A. R. Alce, N. Borinaga, and A. Galido, "Are we inclusive? Accessibility challenges in Philippine e-government websites," Informatics, vol. 12, no. 2, Art. no. 41, 2025, doi: 10.3390/informatics12020041.

[42] AChecker, "AChecker+ Web Accessibility Checker." [Online]. Available: https://achecks.org/achecker. [Accessed: Jul. 17, 2026].

[43] WebAIM, "WAVE: Web Accessibility Evaluation Tool Documentation." [Online]. Available: https://wave.webaim.org. [Accessed: Jul. 17, 2026].

[44] TAW, "Web Accessibility and W3C Standardization Services." [Online]. Available: https://www.tawdis.net/. [Accessed: Jul. 17, 2026].

[45] GTmetrix, "GTmetrix Performance and Lighthouse Metrics." [Online]. Available: https://gtmetrix.com/. [Accessed: Jul. 17, 2026].

[46] M. Vigo, J. Brown, and V. Conway, "Benchmarking Web Accessibility Evaluation Tools : Measuring the Harm of Sole Reliance on Automated Tests Categories and Subject Descriptors," in Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility - W4A '13, 2013, p. 1. doi: 10.1145/2461121.2461124.18.

[47] C. González-Mora, I. Garrigós, S. Casteleyn, and S. Firmenich, "A web augmentation framework for accessibility based on voice interaction," in *Proc. Int. Conf. Web Engineering (ICWE)*, Cham, Switzerland: Springer International Publishing, Jun. 2020, pp. 547–550.

International Journal on Perceptive and Cognitive Computing (IJPCC) Vol 12, Issue 2 (2026)
https://doi.org/10.31436/ijpcc.v12i2.711

Revisiting Layer 2 in Traditional DPI with Realist CFMO Thematic Synthesis

Layer 2 Header field usage in DPI

Shafana M.S[1,2], Adamu Abubakar[2]

[1]Dept. of Information and Communication Technology, South Eastern University of Sri Lanka, Sri Lanka.
[2]*Dept. of Computer Science, International Islamic University Malaysia, Kuala Lumpur, Malaysia.*

*Corresponding Author: zainashareef@gmail.com
(Received: 4th May 2026; Accepted: 12th June, 2026; Published on-line: 30th July, 2026)

Abstract— Traditional Deep Packet Inspection (DPI) systems rarely emphasise data-link layer fields for anomaly detection, despite their foundational role in Ethernet-based communications. Evidence on the security contributions of individual Layer 2 fields in traditional DPI therefore remains limited and fragmented. To examine this gap, this thematic synthesis consolidates research published from 2020 to 2025 on how traditional DPI pipelines utilise Destination and Source Media Access Control (MAC) addresses, EtherType/Length, and Frame Check Sequence (FCS) for security-relevant tasks. The review aims to identify their usage patterns and map the fields to their contexts, mechanisms, outcomes, benefits, and limitations. Eight studies were selected from 2,301 records identified across eight databases through PRISMA-based screening and eligibility assessment. Evidence was extracted manually and analysed using a realist Context–Feature–Mechanism–Outcome (CFMO) lens, while ATLAS.ti 25 supported code management, thematic synthesis, visualisation, and traceability. The findings show that Layer 2 field use remains limited but purposeful. Destination and Source MAC addresses were each reported in six studies, mainly for device distinction, sender identification, session-based flow grouping, traffic-direction analysis, and spoofing detection. EtherType/Length appeared in five studies and facilitated protocol-level parsing and discrimination, while FCS contributed to frame-integrity validation in one study. The Virtual Local Area Network (VLAN) Tag (802.1Q) was absent from the reviewed corpus, indicating broader underutilisation of Layer 2 fields in mainstream DPI research. Mechanisms were largely deterministic and deployed in context-specific scenarios. Overall, Layer 2 fields provide complementary support to higher-layer inspection, but the available evidence does not independently establish high-performance anomaly detection. Future research should examine multi-layer hybrid strategies, VLAN-aware segmentation, and integration of link-layer parsing into DPI-based anomaly-detection pipelines.

Keywords— Anomaly Detection, Data Link Layer, EtherType, Thematic Analysis, Packet Inspection

I. Introduction

The data-link layer header fields, such as frame type and MAC address, provide only basic network information and offer limited insight into actual network activity [1]. When deep packet inspection (DPI) incorporates both these headers and the packet payload, its capacity for threat detection and network traffic profiling is significantly enhanced. This supports both security goals and regulatory compliance. DPI enables analysts to detect anomalies and potential attacks more effectively [2]. Collaborative efforts and the sharing of detection rules through open-source platforms like Snort and Suricata ensure that practitioners remain informed about emerging threats and evolving network protocols [3]. By integrating data-link header checks into DPI, organisations can strengthen network security, meet regulatory requirements, and maintain operational efficiency [4].

A. Traditional Deep Packet Inspection (DPI) and Data Link Layer Field Usage

Layer 2 fields are attributes of the Ethernet data-link frame used for local delivery, protocol identification, network segmentation, and integrity checking. In this review, they include Destination and Source Media Access Control (MAC) addresses, EtherType/Length, the optional IEEE 802.1Q Virtual Local Area Network (VLAN) tag, and the Frame Check Sequence (FCS), which is located in the frame trailer rather than the header. These fields support traffic analysis and anomaly detection within traditional DPI systems. In this review, traditional deep packet inspection (DPI) refers to deterministic, rule-, signature-, or parser-based inspection that examines packet headers and/or payload content to classify, filter, or identify anomalous traffic, without a learned model serving as the primary

inspection decision mechanism. Studies combining DPI with machine learning were retained only when they described an explicit deterministic DPI component that used one or more Layer 2 fields; only evidence related to that component was extracted and coded. While early DPI implementations primarily focused on network and transport layer attributes, recent advancements have shown that including data-link layer information significantly improves visibility and the accuracy of security monitoring. These enhancements support critical functions such as device identification, session grouping, traffic classification, and the simulation of realistic network environments. They are also useful in detecting and investigating abnormal behaviours such as spoofing and unauthorised access. As security challenges have evolved, data-link layer header analysis has been incorporated into modern DPI tools and workflows. This provides foundational support for operational management and research-driven anomaly detection across varied network environments [5], [6].

B. Data Link Layer Field Usage in DPI-Based Anomaly Detection

In traditional DPI systems, data-link layer headers including Source and Destination Media Access Control (MAC) addresses, EtherType, and Virtual Local Area Network (VLAN) tags play an important role in supporting anomaly detection. These fields assist with device identification, maintaining an accurate inventory of network assets, grouping sessions, and analysing user or endpoint behaviour. This helps uncover unusual communication patterns that may signal potential compromise. The EtherType field allows DPI systems to classify traffic by protocol efficiently, which improves the targeting and precision of anomaly detection. VLAN tags provide information about network segmentation and can indicate policy violations or unauthorised access between network zones. Current DPI methods are capable of identifying spoofing attempts, monitoring topology changes, and offering multi-layered protection against advanced threats by using the unique characteristics of data-link layer fields [5–8].

Although existing DPI research broadly examines packet inspection and whole-system performance, evidence on how individual Ethernet Layer 2 fields contribute to deterministic anomaly-detection mechanisms remains fragmented. The contexts in which these fields are used, the mechanisms they enable, and their reported benefits and limitations have not been clearly synthesised. Consequently, the operational role and research gaps of Layer 2 inspection within traditional DPI remain unclear.

C. Objective of the Thematic Review

This thematic review examines the application of Layer 2 fields together with payload analysis for identifying network anomalies, based on research published from 2020 to 2025. By reviewing deterministic, rule-based DPI methods and the deterministic DPI components of eligible hybrid approaches, the study aims to identify the Layer 2 fields used, map their Context–Feature–Mechanism–Outcome configurations, synthesise their usage patterns, benefits, and limitations, and identify remaining evidence gaps. The guiding question is: What patterns and trends characterise the use of data link layer header fields in traditional DPI for anomaly detection between 2020 and 2025?

II. Methods and Materials

A. Theoretical Framework

We use a Realist Context–Feature–Mechanism–Outcome (CFMO) lens grounded in Context–Mechanism–Outcome (CMO) logic [9] and the view that mechanisms involve resources and the responses they generate [10]. We therefore make the resource explicit as Feature, similar to Context–Intervention–Mechanism–Outcome (CIMO)-style extensions [11]. Features are Layer 2 fields such as Destination MAC, Source MAC, EtherType/Length, and Frame Check Sequence (FCS), which enable mechanisms in traditional DPI, including device identification and inventory, session grouping and ingress/egress parsing, protocol discrimination and parser branching, spoofing/redirection detection, and frame-integrity validation. Context comprises the reported deployment or testing environments, such as programmable switch pipelines, software-defined networking (SDN)/Mininet testbeds, packet-analysis tools, Layer 2 protocol settings, industrial frames, and spoofing scenarios. Outcomes summarise the reported usage patterns, benefits, and limitations associated with each Feature–Mechanism relationship. Because few studies attribute quantitative performance to individual Layer 2 fields, outcomes are synthesised qualitatively [9].

In this review, CFMO thematic synthesis refers to a realist-informed qualitative process in which each reported Layer 2 field use is coded as a Context–Feature–Mechanism–Outcome configuration and recurring configurations are consolidated into themes. The unit of analysis was each reported use of a Layer 2 field in an included study. The four CFMO components were applied as deductive parent codes, while specific field uses, mechanisms, contexts, benefits, and limitations were developed as inductive subcodes and consolidated into broader themes. ATLAS.ti 25 was used to manage the codes, conduct code–document and co-occurrence analyses, generate thematic visualisations, and maintain traceability to the source studies. Study selection,

https://doi.org/10.31436/ijpcc.v12i2.711

coding decisions, and interpretation were performed manually by the researchers.

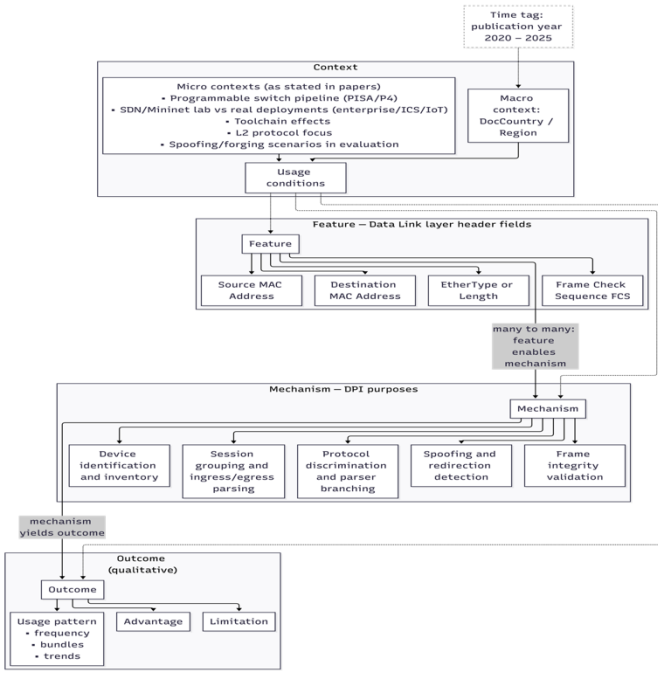


Fig. 1: CFMO-based theoretical framework used in the thematic synthesis.

B. *Rationale for Employing a Thematic Review*

Despite the fact that systematic reviews are consistently used as a means of aggregating primary findings and assessing research designs in a systematic and repeatable way [12], the current review is not just to summarise the literature related to the use of data link layer usage in deep packet inspection (DPI). Instead, the objective is to trace common themes and new directions on how data link layer header fields are utilised in traditional DPI to detect anomalies. Since the research question explores substantive issues, such as parsing of the header field, device and session attribution, protocol discrimination and operational issues of field-level inspection, a thematic review is a broader and flexible format. This method allows grouping different studies together based upon significant analytical themes, thus bringing out patterns and nuances that may be missed by purely quantitative reviews.

C. *Search Strategy*

This literature review aimed to capture recent developments in how data link layer header fields are used within deep packet inspection (DPI) systems for anomaly detection. The review prioritised studies that discuss the implementation, performance, and technical limitations of DPI techniques analysing both header fields and payload content.

Variation in terminology across sources posed a challenge in ensuring comprehensive coverage. To address this, the search terms were deliberately broad, with the understanding that preliminary results would be carefully refined using exclusion criteria. Notably, we did not include specific data link layer field keywords (such as "MAC Address" or "VLAN") in the search string at this stage. This approach was chosen to ensure a broader collection of articles relevant to traditional DPI methods, after which studies that explicitly addressed the use of data link layer header fields could be identified and retained during the screening and filtering process. Throughout the review, technically substantive articles were prioritised, specifically those that demonstrated practical use of DPI such as filtering, pattern matching, or protocol decoding. Theoretical works or papers with only superficial mentions of DPI or data link layer analysis were systematically excluded.

A Boolean search string was applied uniformly across databases:

("Deep Packet Inspection" OR "Traffic Inspection" OR "Packet Filtering" OR "Packet Inspection" OR "Payload Inspection" OR "Encryption Inspection") AND "Anomaly Detection".

The review was limited to peer-reviewed journal articles and conference papers published in English between 2020 and 2025. Sources included Scopus, IEEE Xplore, ACM Digital Library, ScienceDirect, ProQuest, Emerald Insight, Wiley Online Library, and Taylor & Francis Online.

D. *Inclusion and Exclusion Strategy*

This review used a systematic inclusion and exclusion approach to protect the rigour of the methodology and relevance of the thematic content. Specifically, it focused on research articles that covered practical DPI mechanisms involving the application of both data-link layer fields and payload analysis in the context of anomaly detection and network security. Only peer-reviewed journal articles or conference papers published in English between 2020 and 2025, inclusive, were included. Studies combining deterministic DPI with machine learning were included only when they described an explicit rule-, signature-, or parser-based DPI component involving one or more Layer 2 fields. In such cases, only evidence relating to the deterministic DPI and Layer 2 field-processing components was extracted and included in the thematic synthesis. Studies relying exclusively on machine learning without an identifiable deterministic DPI component were excluded. Articles were also excluded if they were published prior to 2020, were not published in English, did not discuss both data-link layer fields and payload inspection, were based on non-peer-reviewed materials, including blogs or white papers, or were retracted.

https://doi.org/10.31436/ijpcc.v12i2.711

The article selection involved four steps: title screening to remove irrelevant articles; abstract screening to exclude off-topic and inadequately detailed studies; diagonal (skim) reading to verify DPI relevance and inspection depth; and full-text assessment to confirm practical DPI application and analysis of data-link layer fields and payloads in accordance with the objectives of the review. Research focusing on DPI techniques, including filtering, protocol decoding, pattern matching, and content inspection involving data-link layer fields, was prioritised. A study was classified as having only a superficial mention when DPI or Layer 2 terminology appeared only in its title, abstract, keywords, introduction, or related work, without describing a practical packet-inspection operation, the use of a specific Layer 2 field, its security or anomaly-detection function, and sufficient evidence to code at least one CFMO configuration. Research focused on high-level monitoring or protocol tuning without explicitly considering packet-level inspection was excluded.

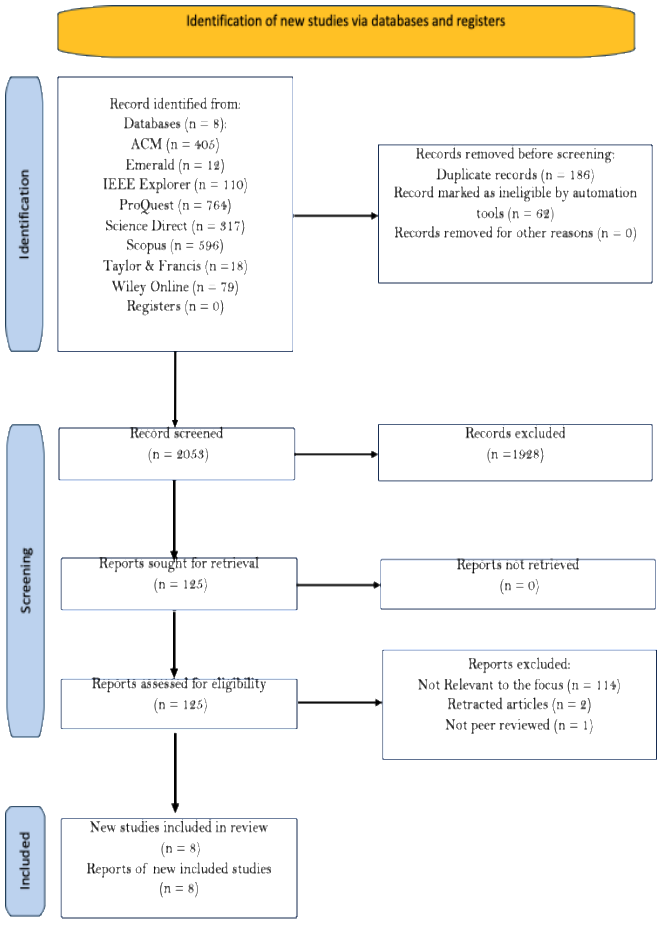


Figure 2: PRISMA Flow Diagram for Article Selection Process

The entire process of selection and screening is depicted in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram (Figure 2). Out of 2,301 records identified across eight major databases, 186 duplicate records were removed, while 62 records were excluded during Rayyan's automated record-processing stage. The remaining 2,053 records underwent manual title and abstract screening, during which 1,928 records were excluded. The full texts of 125 studies were then assessed against the inclusion criteria, including practical DPI operation, explicit use of at least one Layer 2 field, a security or anomaly-detection function, and sufficient technical detail to code at least one CFMO configuration. Following the full-text assessment, 117 studies were excluded, and eight studies were included in the qualitative synthesis and thematic analysis. The manual screening and eligibility assessment were conducted by the authors over approximately eight months.

E. The Thematic Review Process

The review process began with Rayyan, a web-based tool used to filter titles and abstracts effectively and to eliminate duplicate records automatically [13]. Next, the eight studies that met the inclusion criteria were exported to Mendeley, where they were cleaned and organised for final metadata. Each included full text was manually read, and relevant evidence was recorded in a structured extraction form covering bibliographic details, context, Layer 2 feature, mechanism, outcome, benefit, and limitation. A concise evidence-based synopsis of each study was then prepared by the authors. These summaries, rather than the complete articles, were imported into and coded in ATLAS.ti 25 according to the steps outlined by [14] and [15], while the original full texts were consulted when further clarification was required. In ATLAS.ti, the documents were classified according to the year of publication (2020–2025) as well as the main analytical approach. Codes were assigned for country of publication, year, and thematic focus, enabling analysis both chronologically and by research theme. Other code groups, such as author and venue, were based on the metadata and refined during close reading. Thereafter, the coded sections were combined into larger categories, which were further distilled into the principal themes presented in this review. This structured, layered process enabled the authors to trace the evolution of DPI-based anomaly detection using data-link layer fields and identify recurring design patterns throughout the literature during 2020–2025.

Based on the ATLAS.ti-generated codes, the eight articles selected for our review were gathered onto a single thematic level that is directly related to the major question of our research. The theme labels were based on the most common header fields and analytic strategies discussed in the articles (MAC address analysis, EtherType recognition, VLAN tagging, and Frame Check Sequence parsing). The field-level themes represent important parsing logic and attribute usage across referenced studies, consistent with

https://doi.org/10.31436/ijpcc.v12i2.711

well-established Ethernet frame structures, namely, destination and source MAC addresses, EtherType, optional VLAN tag, and the FCS trailer as detailed in foundational Ethernet documentation.

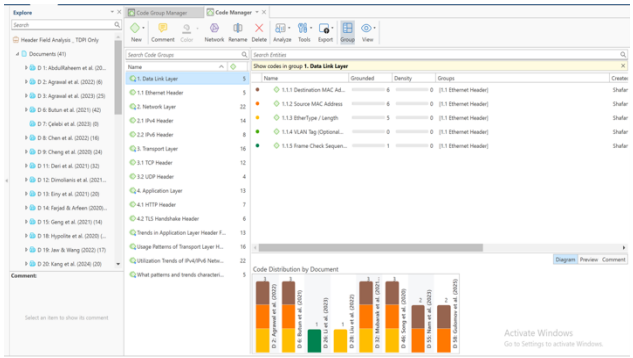


Figure 3: Established Code Group in ATLAS.ti

Notably, although this review is a subset of a broader qualitative study that will examine the utilisation of the header field framework across all four protocol layers in traditional DPI, the present analysis deals specifically with the data link layer usage, as shown in the depicted taxonomy in Figure 4.

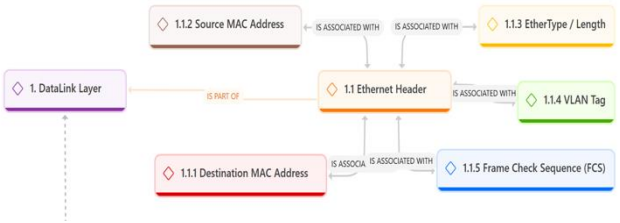


Figure 4: Thematic Code layer Taxonomy

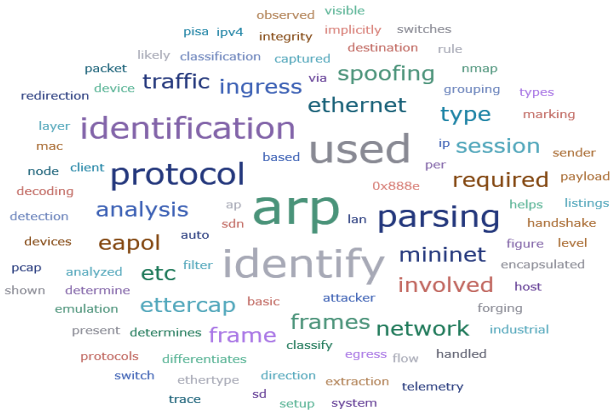


Figure 5: Word Cloud Generated from Articles

The visual summary of the field provided in Figure 5 is presented in the form of a word cloud that has been generated from the most significant terms that occur within the reviewed articles. This visualisation was created before formal coding in ATLAS.ti, and it became apparent that the terms "arp," "identify," "protocol," "frames," "spoofing," "session," and "ethernet" appeared frequently in the literature. The prevalence of the words "identification," "parsing," "ingress," "analysis," "used," and "mininet" highlights the central focus on protocol processing, device identification, Address Resolution Protocol (ARP) analysis, and Ethernet frame parsing fundamental operations in the application of data link layer fields within traditional DPI for anomaly detection. The appearance of the terms "ettercap," "session," and "spoofing" further emphasises the significance of practical tools and attack scenarios discussed in this field of research.

III. Results and Discussion

The main results of this thematic review can be found in two sections. The quantitative analysis summarises the number of studies meeting each identified theme, annual trends from 2020 to 2025, and maps publication activity by country and continent. Each theme is given a focused discussion through the qualitative analysis based on how the chosen studies implement the traditional DPI methods of parsing, protocol identification, frame analysis, and ARP spoofing detection based on data link layer header fields used as indicators of anomaly detection. Combined, these quantitative and qualitative views emphasise not only the scale and the geographic coverage of existing studies, but also the underlying analytic techniques and emerging trends that characterise data link layer field usage in conventional DPI to detect anomalies.

A. Quantitative Findings

Eight papers were found to specifically discuss the application of data link layer fields in conventional DPI methods for anomaly detection within the six-year span of 2020 to 2025 (see Table 2). These papers appeared in an assortment of reputable journals and conference proceedings, among them ACM Conference, IEEE Conference, Sensors, Connection Science, and Electronics (Switzerland). The relatively small number of works to consider, spread over a number of locations and years, points to both the pioneering essence and the niche focus of this research subject. Among the eight chosen articles, five were published in peer-reviewed journals and three in conference proceedings, which reflects the current interest in the topic but also indicates the relatively limited scope of work specifically focused on data link layer field analysis in rule-based DPI anomaly detection during this period.

https://doi.org/10.31436/ijpcc.v12i2.711

TABLE 1
ARTICLES REVIEWED BASED ON TYPES OF JOURNALS

Journal/Conference Name	2020	2021	2022	2023	2025
ACM Conference			1		
Connection Science			1		
Electronics (Switzerland)				1	
IEEE Conference		1			1
IEEE/ACM Transactions on Networking				1	
Sensors		1			
Sensors (Switzerland)	1				

As publication activity depicted in Figure 6, scholarly publication operations covering fields in data link layer in their application in conventional DPI-based anomaly detection were small over the years. In 2020, there was one relevant article, but in 2021, there were two articles, and from 2021 through 2023, the number remained stable at two per year. No study was found for 2024. A single article in 2025 suggests a minimal resurgence after the 2024 gap. In general, this trend indicates that the field of collecting and analysing data-link layer fields in the realm of classical DPI-based anomaly detection has received steady but low scholarly attention, and the lack of publications in 2024 may be either a momentary weakness or may positively indicate that other protocol layers are being addressed or alternative methods of detection are being pursued.

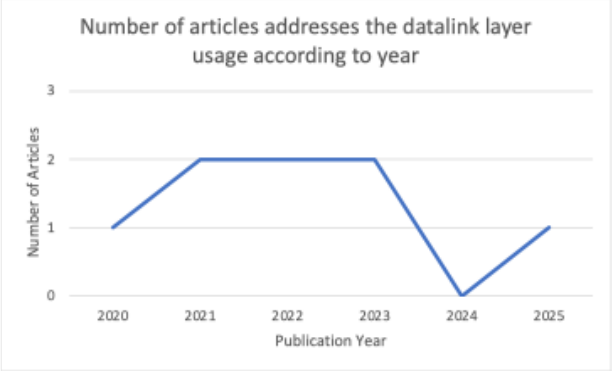


Figure 6: Breakdown of published articles according to the Year of Publication

As Table 2 illustrates, there is broad, though uneven, participation in the international research on the usage of a data-link layer field in traditional DPI-based anomaly detection. China shows sustained participation, with the additions of publications in 2020, 2022, and 2023, which indicates a long-term interest in conducting research in this field. Malaysia is also active with subsequent issues in 2021 and 2022 which mark a trend in the region. India contributes in 2022, reflecting continued interest. A number of other countries such as Poland, South Korea, Sweden, Turkey and Ukraine each contributed a single study within the same time period, which represents less habitual but significant use of this specialised topic. Notably, South Korea's publication in 2023 and Uzbekistan's entry in 2025, along with scattered contributions from Europe and Eastern Europe, signal the gradually diversifying geographic reach of research efforts.

TABLE II
COUNTRY AFFILIATIONS REPRESENTED IN THE EIGHT INCLUDED STUDIES BY YEAR

Country	2020	2021	2022	2023	2025
India			1		
Sweden		1			
Turkey		1			
China	1		1	1	
Ukraine	1				
Republic of Korea				1	
Malaysia		1			
Poland	1				
Uzbekistan					1

The yearly change is noticeable and the fact that there are no new articles as of 2024 indicates that there is either a transient gap or change of focus in the research. Altogether, although no single country can be considered a leader in the data-link layer sphere in the traditional DPI research, the involvement of many countries supports the idea of the growing international scope of this domain.

As Figure 7 and the table of country summative results indicate, the data-link layer field utilisation in conventional DPI-based anomaly detection is spread amongst the few national communities. China emerges as the leading contributor, with three publications over the review period, while Malaysia produced two studies. India, along with Poland, South Korea, Sweden, Turkey, and Ukraine, each contributed a single publication. This spread indicates that there is an established research presence in East and Southeast Asia, with meaningful yet less frequent engagement from European and Eurasian countries.

Figure 7: Breakdown of Geographical map of authors

https://doi.org/10.31436/ijpcc.v12i2.711

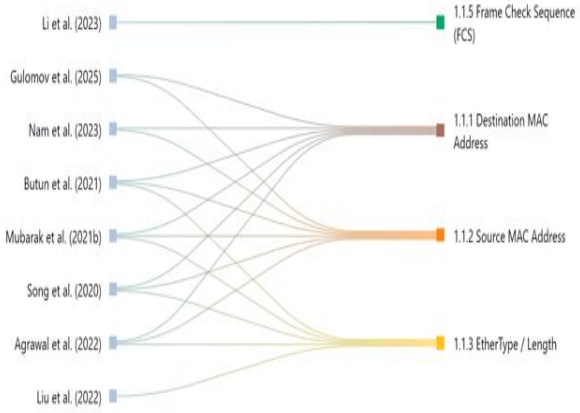


Figure 8: Thematic Mapping of the Publications

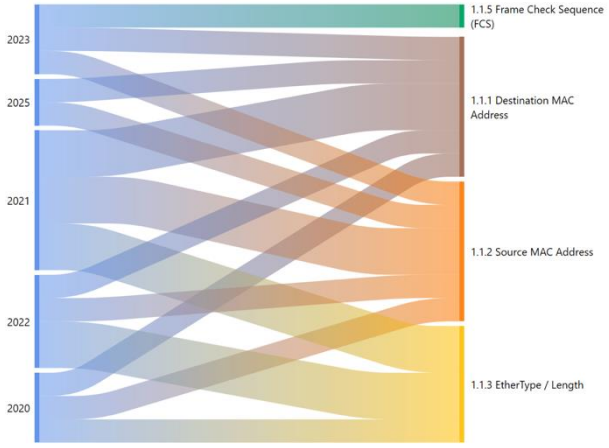


Figure 9: Sankey diagram of the Themes According to Year.

Note: The author–year labels in this ATLAS.ti network visualisations are software-generated document identifiers used for coding traceability. They are not formal in-text citations. The corresponding works are cited in the manuscript text and listed in the reference list according to IEEE style

As illustrated in Figure 8 and 9, the review of traditional DPI-based anomaly detection (n = 8 articles, 2020–2025) indicates that the studies related to data-link layer fields are centred around a limited number of fundamental attributes of Ethernet. Destination and Source MAC Address were addressed in six articles each, and EtherType/Length analysis was found in five research works. The Frame Check Sequence (FCS) was given very little focus because it was examined in only one article, meaning that integrity-check fields are rarely incorporated into DPI-based anomaly detection pipelines.

The trends in the number of publications annually suggest a steady but modest scholarly output, with publication activity peaking in 2021 and gradually declining in subsequent years. In the literature considered, it is shown that the field of data-link layer undergoes extensive inspection, and the majority of works use the combination of one or several types of fields [7], [5], [16], which indicates the further relevance of multi-attribute parsing strategies. Such focused reliance on common Ethernet frame fields highlights the technical consistency of the domain, and the lack of more varied or probabilistic data-link layer mechanisms indicates that there is a prevailing methodological preference for deterministic, header-driven inspection in traditional DPI-based anomaly detection.

B. *Qualitative Findings*

In this part, the qualitative findings of the traditional DPI techniques based on the data link layer field usage are synthesised. Every article was coded as per the thematic taxonomy, including the main areas, which are field-level parsing, device identification, and protocol differentiation. Each theme has been summarised in terms of core methods, deployment contexts, evaluation strategies, and limitations and how later studies build on earlier work. Figure 10 shows the entire polyline representation of all thematic linkages of the data link layer corpus, which assists in unambiguously visualising the clusters of techniques and their relationships. The numbers in the quote identifiers are those used as page numbers of our internal summary documents and not the original articles.

1) *Polyline View of Data Link Layer Themes*

Figure 10 shows a polyline thematic map of the utilisation of data-link layer header fields in conventional DPI-based anomaly detection (2020–2025). This mapping shows how the major Ethernet frame fields, namely, Destination MAC Address, Source MAC Address, EtherType/Length, Frame Check Sequence (FCS), are linked to the research studies implementing the said fields in their practical anomaly detection systems. Both similarities and functional overlaps in roles are depicted by documented use cases across the reviewed literature representing each of the themes. It is noteworthy that the VLAN Tag (802.1Q) field is presented as a stand-alone node that lacks corresponding studies, which indicates a lack of adoption or a research gap regarding the use of VLAN-based classification in the conventional DPI pipelines. The map emphasises a methodological preference for deterministic, field-based inspection, while highlighting the constant emphasis on standard Layer 2 characteristics.

https://doi.org/10.31436/ijpcc.v12i2.711

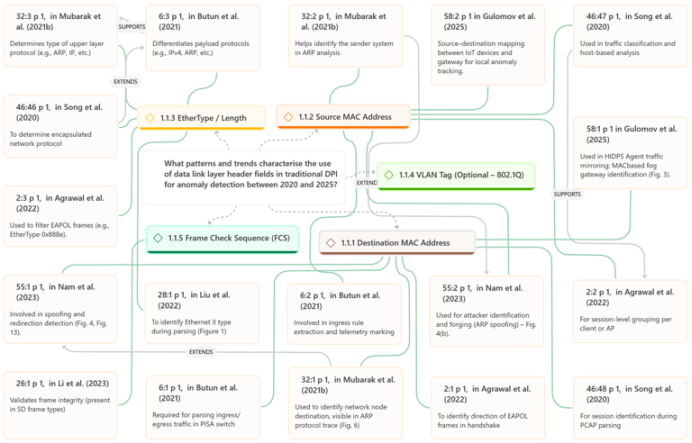


Figure 10: Polyline view of the Themes

2) *Theme 1: Destination MAC Address*

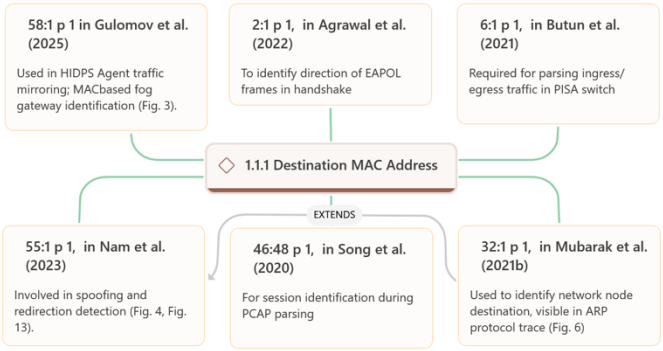


Figure 11: Network View of the research articles on the "Destination MAC Address" Theme

All of them, together, in Figure 11 demonstrate that the Destination MAC Address is a fundamental field for data link layer parsing and anomaly detection in the traditional DPI pipelines. Across recent studies, this area is actively used for key network operations, including identifying sessions, parsing ingress/egress traffic and detecting spoofing as well as device-level classification.

Agrawal et al. [7] show how the Destination MAC field could be used to identify the directionality of Extensible Authentication Protocol over LAN (EAPOL) frames during handshake-based authentication, and thus enabling fine-grained session tracking. According to [5], it is essential in parsing incoming and outgoing traffic in the Protocol-Independent Switch Architecture (PISA) switch environment, which facilitates detailed telemetry. Nam et al. in [17] use Destination MAC parsing to identify spoofing and redirection attacks, especially in particularly within hostile or redirected traffic scenarios. Mubarak et al. in [16] also use the field to determine the destinations of network nodes that can be viewed in ARP traces, which helps to attribute low-level traffic. Song et al. in [8] apply the field to session identification tasks when parsing packet capture (PCAP) sessions, grounding behavioural analysis on MAC-based session granularity. The authors of [18] capitalise on the Destination MAC in host-based intrusion detection and prevention system (HIDPS) agent configurations for fog-based gateway identification via traffic mirroring techniques.

3) *Theme 2: Source MAC Address*

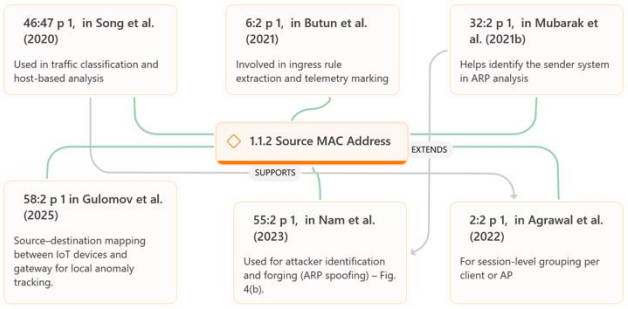


Figure 12: Network View of the research articles on the "Source MAC Address" Theme

The Source MAC Address is essential in tracing the identity of the sender and enforcing security in traditional DPI-based anomaly detection. Recent research uses this field for a broad range of functions, such as ingress rule extraction, telemetry marking, attacker identification, session grouping, host-based analysis, and ARP spoofing as shown in Figure 12.

Butun et al. [5] utilise the Source MAC Address for ingress rule extraction and telemetry marking in programmable network switches. Mubarak et al. [16] use it in the analysis of ARP to detect sender systems that are caught in cyber attack traces. Nam et al. [17] show how it can be applied to attacker identification and MAC address forging detection, particularly to identify ARP spoofing attacks. Song et al. in [8] use it in the field of host-based traffic classification. Agrawal et al. in [7] adopt the use of Source MAC for session-level tracking of clients or access points. In the Internet of Things (IoT) context, [18] apply it to source–destination mapping to support local anomaly tracking.

Taken together, these studies prove the versatility and security relevance of Source MAC Address inspection in different DPI applications. Nevertheless, certain obstacles still remain inconsistent traceability of spoofing, reliance on tool-related results, and insufficient attention to computational efficiency. Improving these limitations will be important in standardising Source MAC Address analysis as a foundational practice in DPI-based anomaly detection systems.

https://doi.org/10.31436/ijpcc.v12i2.711

4) Theme 3: EtherType / Length

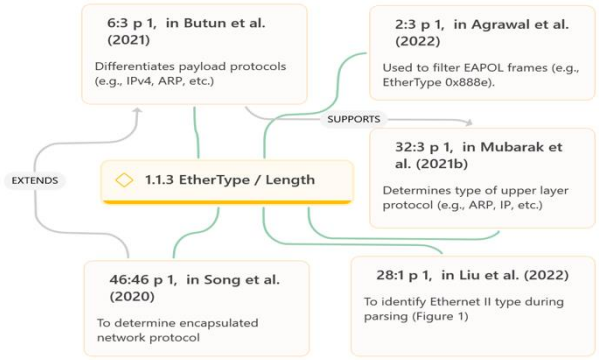


Figure 13: Network View of the research articles on the "EtherType/Length" Theme

The EtherType/Length field is a critical discriminator in traditional DPI-based anomaly detection, enabling precise protocol identification and targeted packet filtering at the data link layer. Recent work employs this field to distinguish between encapsulated payload types (as shown in Figure 13) like IPv4, ARP, and EAPOL, support upper-layer protocol recognition and apply selective filtering for downstream analysis.

Butun et al. [5] use EtherType to differentiate various payload protocols during parsing, ensuring DPI elements are fed with protocol-specific data. Authors of [7] use EtherType filtering to isolate EAPOL and better reflect the authentication traffic segmentation. Liu et al. [19] emphasise that the EtherType is used to determine the types of Ethernet II frames during detailed packet parsing. Mubarak et al. [16] use this area to identify the upper-layer protocol nature in both ARP as well as IP-based flows, contributing to effective protocol dissection in ARP and IP-based flows. EtherType analysis is also used by authors of [8] to detect encapsulated protocols for flow labelling and session classification. Through these contributions, it is clear that EtherType/Length is a pivotal element in facilitating proper differentiation of protocols and interpretation of layers of traffic through DPI pipelines. However, to fully harness its potential, the way forward is to ensure its integration with dynamic protocol stacks and the definition of consistent evaluation criteria.

5) Theme 5: Frame Check Sequence (FCS)

The Frame Check Sequence (FCS) field is a specialised field with a critical role in the analysis of the data link layer in traditional DPI-based anomaly detection, especially in contexts that demand robust frame integrity verification. As Figure 14 indicates, the application of its role is not common in the reviewed literature but remains vital where deployed. The authors of [20] use the FCS field to validate the integrity of SD (Service Data) frame types, making sure that only valid, uncorrupted frames are sent to the further inspection layer. This check is vital to prevent false anomalies due to frame-level transmission errors or tampering at the physical or link layer.

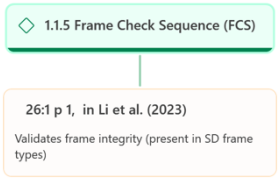


Figure 14: Network View of the research articles on the "Frame Check Sequence (FCS)" Theme

This method strengthens the integrity of subsequent DPI modules by eliminating corrupted frames early in the processing pipeline. Even though the FCS field appears in only one study within the review period, its practical use proves a strategic application of low-level error detection to enhance the accuracy of DPI. With DPI systems also contending with multi-layered threats, such as low-level evasion techniques and manipulation of protocols, a helpful way of providing such defence is to integrate integrity checks such as FCS.

C. Discussion

The subsequent findings are the summary of the principal themes identified in the thematic analysis of the use of data link layer fields in traditional DPI anomaly detection literature from 2020 to 2025.

In all the papers under review, scholars have always pointed out the enduring relevance of data link layer header fields in conventional DPI pipelines in detecting anomalies. Through the fundamental fields of Destination and Source MAC Address, EtherType/Length, and Frame Check Sequence (FCS), these studies can provide detailed traffic visibility and enable device identification, protocol differentiation, and integrity verification, which do not fully scale to flow-based and machine-learning-based methods.

This value is motivated by two factors: (i) the direct interpretability and determinism of header-level parsing, enabling security mechanisms to detect anomalies at line rate quickly and accurately, and (ii) the interplay between this set of fields with effective pre-filtering or session classification strategies, which provide yet further optimisation to DPI performance. Taken together, the evidence demonstrates that data link layer inspection remains a fundamental, low-latency foundation for high-fidelity anomaly detection, particularly in networks where context-aware packet-level scrutiny is essential.

The present paper discusses the application trend and thematic patterns of data link layer header fields in traditional DPI-based anomaly detection from 2020 to 2025. Code-to-document analysis in ATLAS.ti was able to identify recurring technical themes, however, one of the most notable findings is quantitative: of the six years represented in the ATLAS.ti data, just eight works had explicitly utilized data link layer fields in traditional DPI-based anomaly detection pipelines, and a portion of those also included the inspection of payloads. Such a low rate of adoption is in sharp contrast to the increased utilisation of network, transport, and application layer fields seen in the larger DPI literature. Although core data link layer fields such Destination MAC Address, Source MAC Address, EtherType/Length, and Frame Check Sequence (FCS) demonstrate practical utility for functions such as session identification, protocol parsing, and integrity validation, it can be argued that these fields remain underexplored or are perceived as less effective than more upper-layer elements.

The review claims that link-layer inspection is still applicable in specific cases, such as spoofing detection, device fingerprinting, and frame integrity verification. A relative lack of use may reflect technical constraints, such as reduced visibility in encrypted or tunnelled packets, reduced reliability due to MAC spoofing, or lack of data link parsing tools in most DPI frameworks. In addition, the fact that the VLAN Tag (802.1Q) has not been mentioned in any of the reviewed studies underscores the marginal role currently played by optional data link extensions. It is unclear whether this omission reflects oversight, irrelevance, or implementation difficulty, but it leaves open the potential for further research. Future studies that consider fields of the data link layer as part of multi-layer DPI pipelines and document their relative efficacy can help surface underutilised capabilities. Regardless of the existing constraints, data link field analysis can contribute to better anomaly detection, especially when paired with protocol-specific understanding and hardware-level visibility.

IV. Implications and Paths for Future Research

According to the survey, data-link layer fields have the potential to support the identification of spoofing, device identification and frame-integrity checks, however, their role in the traditional DPI-based anomaly detection has not been fully explored as only eight related studies have been reported in the six-year period. Moreover, the literature also does not provide abundant reporting on detection-error values, standardised performance baselines or real-world deployment evidence, which would facilitate a reliable appraisal of the practical viability of Layer 2 features. More importantly, the performance improvements reported cannot be attributed to data-link fields only, as other confounding variables such as characteristics of datasets, feature-engineering choices, classifier parameters, traffic mix, hardware acceleration and data pipeline optimisation can equally promote performance gains. In turn, the cross-study performance comparisons are inherently difficult and, in most cases, methodologically unreliable. Future investigations are hence necessary to include more rigorous and transparent evaluation designs, including ablation experiments that disaggregate the incremental worth of data-link fields, cross-dataset validation, and experimentation under a variety of operational circumstances.

The research gap is highly pronounced, as there is no usage of VLAN Tag (802.1Q) at all, despite the widespread use of segmented, virtualised, and multi-tenant networks. This gap highlights opportunities for future research on VLAN-aware anomaly detection strategies. Also, the analysis of data link layer may be combined with multi-layer or machine-learning–assisted DPI to detect encrypted or evasive traffic. Overall, to enhance the utility of data link layer fields in DPI-based security, better reporting practices, extended experimental validation, and hybrid inspection strategies are justified.

V. Contributions and Benefits of the Study

The present paper is a systematic analysis of the use of data link layer fields in traditional DPI-based anomaly detection between 2020 and 2025. It reports common patterns of application, identifies dominant patterns involving Destination and Source MAC addresses, EtherType/Length, and frame-integrity checks, and highlights significant gaps, particularly limited performance reporting and the complete absence of VLAN Tag analysis. The results provide researchers with concrete directions for future investigation and guide practitioners in understanding which security functions the reviewed evidence associates with each Layer 2 field, including device and session identification, spoofing detection, protocol discrimination, and frame-integrity validation. However, the available evidence is insufficient to determine the independent efficiency contribution of individual Layer 2 fields within traditional DPI pipelines.

ACKNOWLEDGMENT

The authors would like to thank colleagues who provided feedback on the study
design and manuscript.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

AUTHOR(S) CONTRIBUTION STATEMENT

All authors contributed equally to the study design, methodology, software development, data analysis, and manuscript preparation. All authors reviewed and approved the final manuscript.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ETHICS STATEMENT

Ethical approval is not required for the publication of the paper.

REFERENCES

[1] M. Çelebi, A. Özbilen, and U. Yavanoğlu, "A comprehensive survey on deep packet inspection for network security," NOHU Journal of Engineering Sciences, vol. 12, no. 1, pp. 1–29, 2023.

[2] F. A. Khan and A. A. Ibrahim, "Network traffic classification analysis on differentiated services code point using deep learning models for efficient deep packet inspection," International Journal of Innovative Computing, vol. 14, no. 2, pp. 15–24, 2024, doi: 10.11113/ijic.v14n2.438.

[3] Snort Project, "What are community rules?" Snort.org FAQ, 2025. [Online]. Available: https://www.snort.org

[4] M. Çelebi and U. Yavanoğlu, "Accelerating pattern matching using a novel multi-pattern-matching algorithm on GPU," Applied Sciences, vol. 13, no. 14, Art. no. 8104, 2023, doi: 10.3390/app13148104.

[5] I. Butun, Y. K. Tuncel, and K. Oztoprak, "Application layer packet processing using PISA switches," Sensors, vol. 21, no. 23, Art. no. 8010, 2021, doi: 10.3390/s21238010.

[6] S. Mubarak, M. H. Habaebi, M. R. Islam, F. D. Abdul Rahman, and M. Tahir, "Anomaly detection in ICS datasets with machine learning algorithms," Computer Systems Science & Engineering, vol. 37, no. 1, pp. 33–46, 2021, doi: 10.32604/csse.2021.014384.

[7] A. Agrawal, U. Chatterjee, and R. R. Maiti, "kTRACKER: Passively tracking KRACK using ML model," in Proc. 12th ACM Conf. Data and Application Security and Privacy (CODASPY '22), 2022, pp. 1–12, doi: 10.1145/3508398.3519360.

[8] W. Song, M. Beshley, K. Przystupa, H. Beshley, O. Kochan, A. Pryslupskyi, D. Pieniak, and J. Su, "A software deep packet inspection system for network traffic analysis and anomaly detection," Sensors, vol. 20, no. 6, Art. no. 1637, 2020, doi: 10.3390/s20061637.

[9] R. Pawson, T. Greenhalgh, G. Harvey, and K. Walshe, "Realist review—A new method of systematic review designed for complex policy interventions," Journal of Health Services Research & Policy, vol. 10, suppl. 1, pp. 21–34, 2005, doi: 10.1258/1355819054308530.

[10] S. M. Dalkin, J. Greenhalgh, D. Jones, B. Cunningham, and M. Lhussier, "What's in a mechanism? Development of a key concept in realist evaluation," Implementation Science, vol. 10, Art. no. 49, 2015, doi: 10.1186/s13012-015-0237-x.

[11] D. Denyer, D. Tranfield, and J. E. Van Aken, "Developing design propositions through research synthesis," Organization Studies, vol. 29, no. 3, pp. 393–413, 2008, doi: 10.1177/0170840607088020.

[12] J. Noyes, J. Popay, A. Pearson, K. Hannes, and A. Booth, "Qualitative research and Cochrane reviews," in Cochrane Handbook for Systematic Reviews of Interventions, vol. 5.0.1, J. Higgins and S. Green, Eds. Wiley, 2008, pp. 1–18.

[13] M. Ouzzani, H. Hammady, Z. Fedorowicz, and A. Elmagarmid, "Rayyan—A web and mobile app for systematic reviews," Systematic Reviews, vol. 5, Art. no. 210, 2016, doi: 10.1186/s13643-016-0384-4.

[14] A. F. Musfira, N. Ibrahim, and H. Harun, "A thematic review on digital storytelling (DST) in social media," The Qualitative Report, vol. 27, no. 8, pp. 1590–1620, 2022, doi: 10.46743/2160-3715/2022.5383.

[15] M. Zairul, "A thematic review on student-centred learning in the studio education," Journal of Critical Reviews, vol. 7, no. 2, 2020, doi: 10.31838/jcr.07.02.95.

[16] S. Mubarak, M. H. Habaebi, M. R. Islam, and S. Khan, "ICS cyber-attack detection with ensemble machine learning and DPI using cyber-kit datasets," in Proc. IEEE Conf., 2021, pp. 1–6, ISBN: 978-1-7281-1065-3.

[17] J. Nam, S. Lee, P. Porras, V. Yegneswaran, and S. Shin, "Secure inter-container communications using XDP/eBPF," IEEE/ACM Transactions on Networking, vol. 31, no. 2, pp. 672–685, 2023, doi: 10.1109/TNET.2022.3206781.

[18] S. R. Gulomov, T. R. Khudayberganov, T. T. Turdiyev, A. B. Xasanov, and R. R. Raximov, "Hybrid traffic filtering and anomaly detection model for next-generation networks," in Proc. 2025 IEEE Ural-Siberian Conf. Biomedical Engineering, Radioelectronics and Information Technology (USBEREIT), 2025, doi: 10.1109/USBEREIT65494.2025.11054242.

[19] X. Liu, Z. Liu, Y. Zhang, W. Zhang, D. Lv, and Q. Zhou, "TCN enhanced novel malicious traffic detection for IoT devices," Connection Science, vol. 34, no. 1, pp. 1322–1341, 2022, doi: 10.1080/09540091.2022.2067124.

[20] Z. Li, Q. Wei, R. Ma, Y. Geng, Y. Yang, and Z. Lv, "DpGuard: A lightweight attack detection method for an industrial bus network," Electronics, vol. 12, no. 5, Art. no. 1121, 2023, doi: 10.3390/electronics12051121.

International Journal on Perceptive and Cognitive Computing (IJPCC) Vol 12, Issue 2 (2026)
https://doi.org/10.31436/ijpcc.v12i2.700

A Survey of Surveys (SoS) on Sentiment Analysis using Machine Learning

[1]Sanah Abdullahi Muaz, [2]Usman Ali, [3]Abdulsalam Ya'u Gital
[1]Software Engineering Department, faculty of Computing, Bayero University Kano, Nigeria
[1,2,3]Computer Science Department Abubakar Tafawa Balewa University, Nigeria

Corresponding Author: samaaz.cs@buk.edu.ng
(Received: 30th March 2026; Accepted: 29th April, 2026; Published on-line: 30th July, 2026)

Abstract— This paper presents a comprehensive Survey of Surveys (SoS) on sentiment analysis using machine and deep techniques. Although several survey studies have been conducted in this domain, the absence of a centralized and structured synthesis makes it difficult for researchers to identify trends challenges, and research opportunities. To address this gap, this study systematically reviewed 70 papers categorizing them based on application domain such as fake news detection, hate speech detection, sarcasm analysis, context based, and content-based sentiment analysis. Unlike other surveys, the work provides a meta level analysis by comparing existing surveys, identifying methodological patterns, and highlighting their limitation. We present a structured taxonomy of sentiment analysis approaches, alongside an examination of the evolution from classical machine learning to deep learning techniques. In addition, recent advances such as transformer-based architecture and large language models are discussed to provide updated context. In addition, this study identifies key challenges, including data imbalance, domain adoption, multimodal complexity, and lack of standardized evaluation frameworks. The SoS serves as a comprehensive reference for researchers by consolidating existing knowledge, identifying research gaps and suggesting future research directions.

Keywords— Natural Language Processing, Machine Learning, Deep Learning, Sentiment Analysis, Multimodal, and Multi-task learning.

I. Introduction

The rapid growth of digital communication platforms has led to an unprecedented increase in user generated textual data. This development has intensified the importance of Natural language processing (NLP) as a key area within artificial intelligence for extracting meaningful insights from unstructured data ([1], [3]-[13]). Among the various application of NLP, sentiment analysis has emerged as a critical technique for identifying and interpreting opinions, emotions, and attitudes expressed in both textual and visual content [6]. Sentiment analysis, also refers as opinion mining plays a vital role in numerous domain such as business intelligence, social media analytics, political forecasting, and customer relationship management [11]. It enables organization and researchers to analyze user feedback, predict trends, and support decision making processes [15]. Over the years, machine learning, deep learning, and transformer-based techniques have significantly enhanced the performance of sentiment analysis systems, enabling more accurate and scalable solutions.

Due to the rapid advancement of this field, a substantial number of survey and review papers have been published each focusing on a specific aspects such as algorithms, applications, or data set. While these surveys provide valuable insights, the increasing volume of review studies has created a new challenge: the difficulty of identifying, comparing, and synthesizing survey contributions. Despite the availability of numerous surveys on sentiment analysis, there is currently a lack of a comprehensive meta-level study that systematically analyses and compares these surveys papers. Existing works primarily focus on the individual techniques or domain, without providing a unified perspective on the overall research landscape. This create a gap for a higher-level synthesis that can consolidate findings, identify trends, and highlight open challenges.

To address this gap, this paper presents a Survey of Surveys (SoS) on sentiment analysis using machine learning, unlike traditional surveys, this study focuses on reviewing and analysing existing survey papers and structured overview of prior survey efforts, enabling researchers better to understand the evolution, strength, and limitation of the field. The main contribution of this paper is as follows: I. A systematic collection and analysis of survey of surveys on

https://doi.org/10.31436/ijpcc.v12i2.700

sentiment analysis using machine learning, II. A structured taxonomy of sentiment analysis approaches based on application domain methodologies, III. A comparative analysis highlighting trends, strengths, and limitations of existing surveys and IV. Identification of key research challenges, and future directions, including emerging approaches such as transformer- based models and large language models. Recent advancement in large language models as GPT-4 [97] and LLaMA [27] have significantly transformed sentiment analysis research, enabling improved contextual understanding and performance across various NLP task.

The remainder of the papers is organized as follows: Section 2 describes the research methodology adopted in this study. Section 3 presents machine learning techniques used in sentiment analysis. Section 4 discusses open challenges and future directions. Finally, section 5 concludes the paper. Due to the rapid advancement in this field, a substantial number of surveys on sentiment analysis, there is currently a lack of a comprehensive meta-level study.

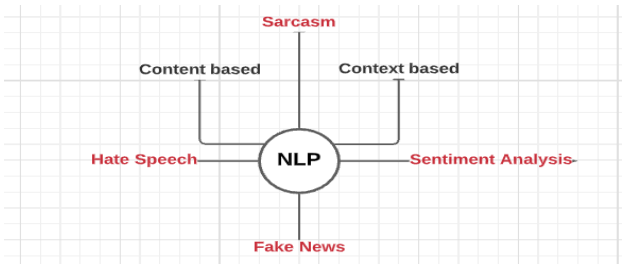


Fig. 1 Conceptual framework for NLP

Figure 1 presents the conceptual framework of natural language processing and its related components, including sentiment analysis, hate speech detection, sarcasm detection, and fake news identification. The figure highlights the interconnected nature of these domains, demonstrating how sentiment analysis operates as a central task within NLP ecosystem.

This framework provides a foundational understanding of how different analytical tasks contribute to extracting meaningful insights from textual data, thereby justifying the scope of this study. Unlike traditional survey compilations that primarily list and categorize prior studies, this work goes beyond descriptive summarization by providing a comparative and analytical synthesis of existing survey papers.

Specifically, the study evaluates methodological trends, identifies common limitations, and highlights research gaps across different survey contributions. This analytical perspective distinguishes the present work as comprehensive Survey of Surveys rather than a simple aggregation of prior studies

II. RESEARCH METHODOLOGY

This study adopts a structured review methodology to systematically identify, analyse, and synthesize existing survey papers on sentiment analysis using machine learning. The approach is inspired by established systematic literature review guidelines, including PRISMA principles, to ensure transparency, reproducibility, and minimization of bias. The review process consists of four main stages: (1) Search strategy, (2) study selection, (3) data extraction, and (4) analysis and synthesis. A comprehensive set of survey papers was systematically identified and analysed based on predefined inclusion and exclusion criteria. The survey adopted the procedures given by [81] for systematic literature review and surveys in computing, and [84] to serve as the research guide.

A. *Search Keywords*

The most keywords used for search benchmarks in this paper are (“Machine Learning”), (“Deep Learning”), (“Natural Language Processing”), (“Deep Learning” AND “Sentiment Analysis”), ((“Machine Learning”) AND (“Natural Language Processing”) AND (“Sentiment analysis” OR “Opinion Mining” OR “Sarcasm”)), ((“Fake news” OR “Hate speech” OR “Context based” OR “Content” AND “Natural Language processing”)). Many key words were initially generated to get insight about the research and were later concentrated in order to achieve the main purpose of the entire research.

B. *Search article results*

During the search era journal articles, edited books, and conference proceeding were search using the key words from the previous section 2.1, the initial search was conducted from 3rd July to 30th July 2021, the second search was conducted between October to November 2021, and the final search has been conducted in November 10, 2021. The nominated papers were scrutinized and also studied to (1) Define NLP and SA with regard to their types and distinction, (2) Identify and give details of NLP taxonomy, (3) presents the challenges of SA and recommend the alternative future direction.

C. *Academic databases*

The emphasis was on papers from reputable publishers such as IEEE, Elsevier, MDPI, Nature, ACM, and Springer. We have reviewed article journals on various SA, ML, NLP, and DL topics. The papers were engendered based on the keywords in the above section of 2.1; the goal of this review is to retrieve papers from reputable and reliable conference proceedings, journals, and edited book chapters indexed within reputable data bases. Table 1 shows the listed the academic data bases visited to source and retrieve articles.

https://doi.org/10.31436/ijpcc.v12i2.700

TABLE I
THE SOURCES AND DATABASE LINKS FOR PAPERS EXTRACTED FOR THIS STUDY

Database	Database link
IEEEXplore	http://ieeexplore.ieee.org/
ACM Digital Library	http://dl.acm.org/
Springer	http://www.link.springer.com/
ISI Web of Science	https://apps.webofknowledge.com/
Scopus	https://www.scopus.com/
Science Direct	http://www.sciencedirect.com/
DBLP	http://dblp.uni-trier.de/
Wiley Online Library	http://onlinelibrary.wiley.com/

D. Inclusion and exclusion measures

In order to have the rightful and eligible articles from this data bases, a certain criterion was set and observed (see Table 2).

Initial search of 1026 papers were retrieved, firstly we removed duplicate from the initial search generated from the academic bases to have the new total as 867, secondly abstract and conclusion were considered to remove and eliminate articles to directly related to the research. And finally, a total number of 74 references have succeeded in becoming eligible for the survey. Figure 2 represents and identified the process adopted in this literature for article selection

TABLE III
THE INCLUSION AND EXCLUSION MEASURES

Inclusion	Exclusion
We only consider papers written in English language.	We do not consider articles that are not written in English language
We considered papers published in reputable peer review journals, conferences and edited books.	Articles published as thesis, editorials, keynote speeches, abstract, and text book were not considered
We considered articles with the application of NLP to machine learning	NLP applications in other domains were not considered
The review focus on the survey papers in NLP	Any paper that does not cover NLP was eliminated

.

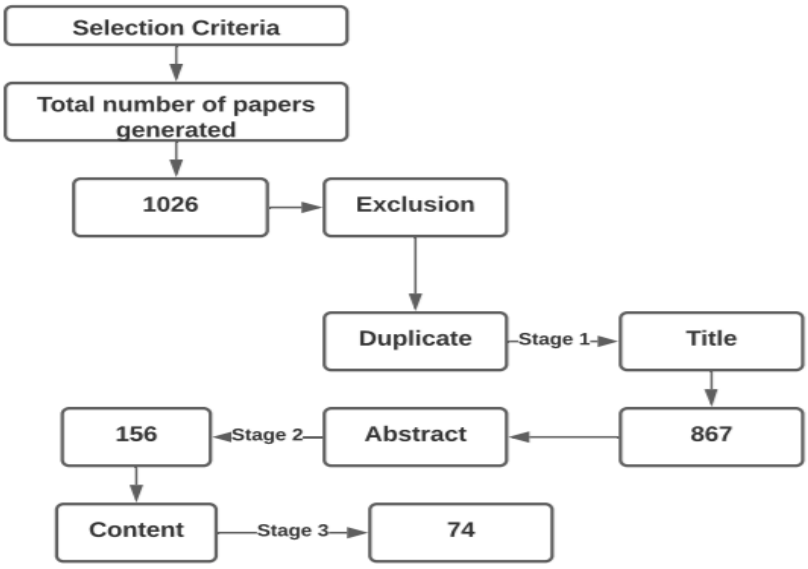


Fig. 2 Article exclusion procedures

III. MACHINE LEARNING ALGORITHMS

Machine learning plays a central role in sentiment analysis by enabling the automatic classification of textual data in to sentiment categories. Existing approaches can broadly be categorized in to shallow learning and deep learning techniques. While Shallow learning methods rely on handcrafted features, deep learning approaches automatically learn feature representations from data.

Over time, sentiment analysis has evolved from traditional machine learning models to deep learning and, and more recently large language models which offer improved contextual understanding and performance ([21]-[55]). This section provides a concise overview of these techniques, focusing on their relevance and application in sentiment analysis.

https://doi.org/10.31436/ijpcc.v12i2.700

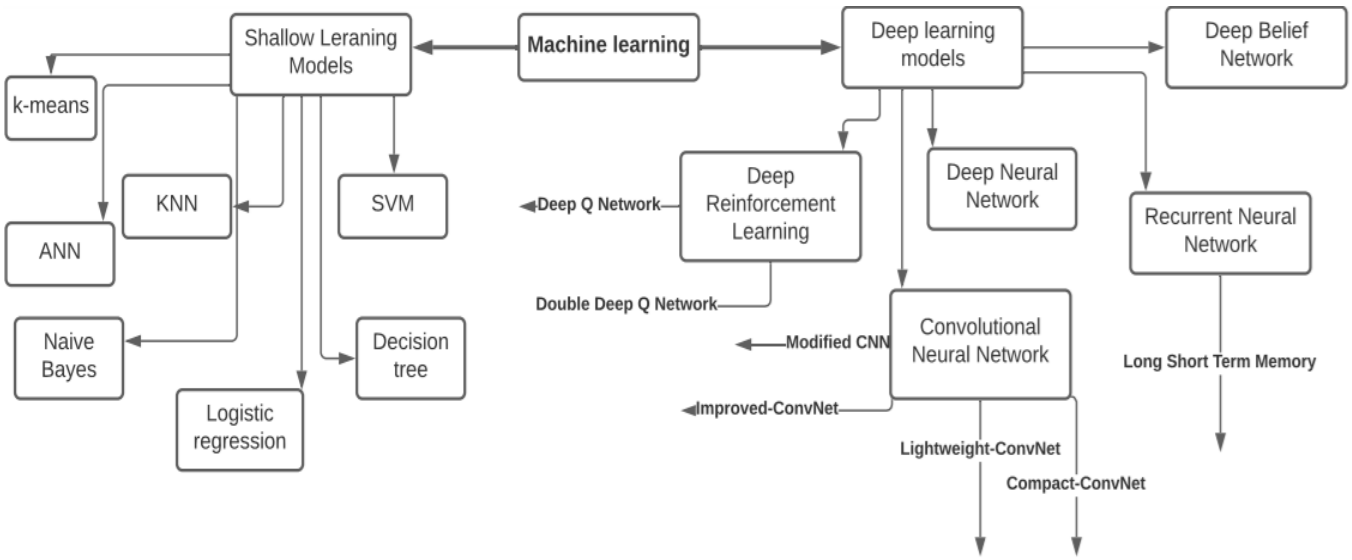


Fig. 3 Taxonomy of Machine learning architecture

Figure 3 illustrate the taxonomy of machine learning techniques in sentiment analysis, categorizing them in to shallow and deep learning approaches. This classification is significant as it highlights the evolution from traditional models such as Support Vector Machines and Naïve Bayes to more advanced deep learning architectures. The figure also emphasizes the increasing complexity and capability of models in handling large scale and unstructured data. A key insight from the figure 3 is the key insight is the increasing shift toward deep learning techniques, which offer improved performance due to their ability to capture contextual and semantic information. These trends reflect the growing demands for more robust and scalable sentiment analysis solutions

A. Shallow learning

Shallow learning methods were among the earliest approaches used in sentiment analysis and remain relevant due to their simplicity and computational efficiency. Common techniques include

K-nearest neighbour (KNN), neural network (ANN), support vector machine (SVM), logistic regression (LR), decision tree, clustering, naïve Bayes and joined hybridize approaches. Most of these approaches were considered some decades ago and their method is established [51]. These models rely heavily on feature engineering techniques such as n-gram, term frequency-inverse document frequency (TF-ID), and lexical features. Despite their limited effectiveness in structured environments, their performance is often limited handling large scale and context dependent data. Nevertheless, recent studies show that shallow models are still useful in low resources and domain specific application where computational constrains exist [42].

Artificial Neural Network (ANN) is an example of shallow learning algorithms and the intention is to reflect the way human being brain is working. ANN is made up several layers both input and output respectively. There is full connection between links at adjacent layers containing large amount of data that can tentatively estimate arbitrary meanings. Whenever it comes to nonlinear functions ANN poses resilient fitting capability. Out of the complexity in ANN structure, it ingests a lot of time while training data. It is worth noting that ANN model is being trained with back propagation there by making it impossible to be used in training deep networks algorithms.

K-Nearest Neighbour (KNN) is an easy and simple supervised shallow learning algorithm aimed at solving regression and classification related problems, the main focus of KNN is based on the multiple theory and hypothesis. The major problem of KNN is that it's very slow and insignificant with large amount data. Also, to note, that the easier and simpler the method is the infective and inefficient the fitting capability will be.

Support Vector Machine (SVM) is another example of supervised shallow learning algorithm used for the purpose of regression, classification and outliers' discovery. SVM memory is proficient in achieving acceptable results even with small amount of training sets. This is because splitting hyper plane can be determined by small amount of support vectors. It's worth noting that SVM can be used in solving nonlinear difficulties effectively

K-Mean clustering is a process of vector estimation, initially from indication process, that target to divide n observations into k clusters.

Decision Tree is to classify data using sequence of regulations, the technique is similar to a tree procession, and these make it to be understandable. The algorithm is

https://doi.org/10.31436/ijpcc.v12i2.700

capable of eliminating redundant and inappropriate features. Feature selection, tree generation, and tree snipping are amongst the learning process, appropriate features are selected independently before generating subsequent child nodes.

Naïve Bayes is an artless technique mainly used for classifiers construction. Naïve Bayes models is used to allocate class labels to difficult occurrences, characterized as vectors of feature principles, the Naïve Bayes classifier computes the provisional possibilities for dissimilar classes.

Logistic Regression the LR is another type of logarithm type of shallow algorithms with linear model. The LR algorithm calculates the chances of diverse classes via parametric logistic dissemination.

B. *Deep learning*

Deep learning approaches have significantly improved sentiment analysis by enabling automatic feature extraction from large and unstructured datasets. Prominent architectures include Convolutional Neutral Networks (CNN), Recurrent Neural Network (RNN), and Deep Neural Networks (DNN) (see Table 3).

TABLE III
COMPARISON BETWEEN SEVERAL DEEP LEARNING TECHNIQUES

Algorithms	**Function**	**Supervised/ Unsupervised**	**Data types**
CNN	Feature Classification and Extraction	Supervised	Feature vectors, Raw data and Matrices
DNN	Feature Classification and Extraction	Supervised	Feature vectors
RNN	Feature Classification and Extraction	Supervised	Feature vectors, Sequence data and Raw data
DBN	Feature Classification and Extraction	Supervised	Feature vectors
DRL	Feature classification and augmentation	Supervised	Feature vectors

Unlike shallow learning methods, deep learning models capture contextual and semantic relationships within text making them more effective for complex sentiment analysis task such as sarcasm detection and multimodal analysis. These approaches have demonstrated superior performance across various benchmark datasets ([41], [55] - [96]). However, deep learning models require large annotated datasets and substantial computational resources, which may limit their applicability in low resource settings. The evolution from shallow learning to deep learning techniques highlights the increasing capability of sentiment analysis models. Deep learning is agreed to have gained an ideal reputation because of the innovation in computational capability, with the advent of graphics processing unit (GPU), it reduces the amount spent in hardware and has enhanced and upgraded network connectivity with robustness in result.

C. *Transformer-Based Models in Sentiment analysis*

Recent advances in NLP have been significantly influenced by the introduction of the Transformer architecture, which enables efficient modelling of contextual relationships in textual and visual data. Transformer based models heavily rely on self-attention mechanisms to capture long range dependencies, making them highly effective for sentiment analysis task. Pre trained languages such as BERT have achieved state of the art performance across various NLP applications, including sentiment classification. These models leverage large-scale training data and transfer learning technique techniques to improve generalization and accuracy. The emergence of transformer-based approaches represents a significant shift in sentiment analysis research. Therefore, these models are highlighted as an important direction for future research and extension of existing survey of surveys. Recent studies demonstrate that large language models outperform traditional deep learning models in sentiment classification task providing improved contextual representation and generalization capabilities. This revision incorporates recent studies published between 2023 and 2025 to ensure the timeliness and relevance of the review.

Convolutional Neural Network (CNN) is among deep learning techniques examples developed for processing visual data such as videos and images among others. It is capable of handling visual data, audio data and text data respectively no matter the amount data [41]. CNN executes well whenever used in processing images more specific to image detection, classification recognition, and computer vision. In CNN input data must be transformed in to matrices for detection of attacks as CNN only works with two-dimensional data. Deep Neural Network (DNN) is a deep learning architecture generally adopted in dealing with classification and regression related problems. It is formed in hierarchical layers each layer having number of neurons [41].

Subsequent layers gradually learn any composite patterns from the received input in preceding layers. Every unit in DNN requires current and previous state in order to obtain contextual information, supervised and unsupervised learning techniques can be used to train data in DNN. Supervised learning methods uses labelled data to train DNN in order to learn for accurate prediction. While the unsupervised learning, method do not require labelled data to train DNN and are commonly used for clustering and future selection.

Recurrent Neural Network (RNN) is powerful algorithms ordinarily adopted and used in NLP for dealing with sequential and serial data problem that are normally characterized as contextual [85]. The sentence level representation can be used to create concluding classification for a particular input sentence. The vector of node "very interesting" is composed from the vectors of the node "very" and the node "interesting." Similarly, the node "is very interesting" is composed from the phrase node "very interesting" and the word node "is". The edifice of an RNN is depicted in

Deep Belief Network (DBN) is a multiplicative algorithm with quite a lot of layers. Every layer considers conspicuous units as an input and considers hidden units as an output. Each layer is coupled and connected to one another. Therefore, the detectible units are totally associated with the unseen units without any interconnection between the two [41]. In most cases upper layers have subsidiary contacts between them, however subordinate layers require straight and top-down assembly with the upper layers [6]. In general, there is direct connection between additional layers with the flow directing to the layer that is closer to the input. As a result, the connections between the two leading layers remain undirected.

Deep Reinforcement Learning (DRL) is a semi supervised algorithm that uses unlabelled data for input in order for the algorithm to study and acquire feature demonstration of the data. Cultured features will be later used in supervised learning task. The need for regulatory dataset is not required. For example, an agent will study and learn to finish a task through testing and getting errors. Thus, DRL uses examination and manipulation to take decision. In manipulation, action is taken based on recent best policy. Examination deals with taking action purposely to gain additional training data. The major challenge in DRL is trade-off has to be made between exploitation and exploration.

IV. Sentiment Analysis Epitome: Conception and Taxonomy

The rapid development of smart devices with the introduction of web 2.0 has led to the advent of social media network, forums, and blogs and has enabled users to take part in sharing their opinions explicitly over these platforms [29]. In consequence, the messages posted by users can reach to a giant spectator in a segment of minutes [72]. Sentiment analysis sometimes refers as Opinion mining is deliberated as an effective method or approach for categorizing and organizing user assessments into positive and negative polarities. However, sentiment analysis can be further classified into lexicon-based methods, machine learning methods and hybrid approaches. Furthermore, the main data source in sentiment analysis is from online social media, it generates considerable amount of data. Moreover, this kind of information must be considered and carefully measured using immense data approach in order to achieve effective and efficient data access, processing and storage for a better result.

A. *Taxonomy of sentiment analysis architecture*

The taxonomy of the sentiment analysis architectures has three architectures: Lexicon based, Machine Learning based, and Hybrid based approach. Several architectures have been proposed for the Sentiment analysis and the trending structures f will be discussed as follows:

Lexicon based is the first approach to be used in sentiment analysis, the approach is classified in two distinct categories namely corpus based and dictionary based that uses arithmetic or semantic procedures to discover sentiment polarity [60]. Formerly, sentiment classification is done through dictionary such as WordNet and SentiWordNet [29].

Machine learning based approach depends on the prominent ML algorithms in order to solve sentiment analysis as an ordinary text organization with difficulty using phonological features [85].

Hybrid based approach is the combination of the two approaches and demonstrates to corporate with sentiment lexicons playing a key role in the majority of methods. Therefore, [45] reports the hybridize experiment carried on Lithuanian data set using machine learning (Naïve Bayes Multinomial, Support Vector Machine) and deep learning (Long Short-Term Memory, Convolutional Neural Network) approaches. Similarly, the ordinary machine learning techniques were applied to feature based lexical, morphological, and character information and the deep learning approaches were applied to the word embedding (Vord2Vec continuous bag-of-words with negative sampling and FastText).

The taxonomy of sentiment analysis approaches, includes lexicon based, machine learning based, and hybrid methods. The taxonomy is important as it enables a comparative understanding of different methodological approaches and their respective strengths and limitations. The taxonomy also reveals a growing trend toward hybrid and data driven models, indicating a shift away from purely rule-based systems. This insight supports the need for more robust and integrated approaches in the future research.

1) *Aspect based sentiment analysis*

Aspect based sentiment analysis usually refers to feature based or entity-based sentiment analysis. It involves aspect identification in a given sentence for classifying its features to either negative or positive and will be later categorized in to diverse polarities ([4], [94]-[95]). Aspect-level SA is to categorize the sentiment with esteem to the precise aspects of entities. The first stage is to detect the entities with their

respective aspects. Moreover, aspect level classification is one of the areas that needs attention as far as NLP is concern ([16], [47], [94]-[95]). Thus, aspect level sentiment analysis is cantered on a short text semantic richness and has been reflected to be one of the most promising research directions in the future ([66], [69], [70]-[78]). However, abundant machine learning frame work has been developed to solve aspect-based problem but deep learning techniques could also be employed in order to have more accurate results ([48]-[75]).

Similarly, Deep learning has achieved quite number of success in the field of aspect based and have proved productivity in application fields such as economics, products, biomedicine, and in policy making by both private and public agencies ([40],[94]–[95]). Consequently, RNN achieved the highest result as compared to other techniques, therefore the performance of deep learning-based models on the aspect level sentiment classification is quite a challenge that needed much more efforts to be solved [54]. Whereas LSTM is one of the most commonly used tools for aspect based sentiment analysis [85], employing deep learning in aspect based sentiment analysis is still in the early stage.

Therefore, a better performance could be achieved by joining extraction and classification of aspect with classification and sentiment. Common approaches used in aspect sentiment analysis include CNN, LSTM, and GRU [31]. Finally, applying deep learning approaches in aspect based will provide a better result in getting the most comprehensive research in sentiment analysis [23]. The central objective of aspect-based classification is to discover the sentence and assume sentiment conveyed in each aspect

2) *Sentence based sentiment analysis*

Sentence based sentiment aims at classifying user's opinion conveyed in each and every sentence. The first stage is to identify if the sentence is subjective or objective. If the sentence happens to be subjective, therefore sentence based will determine if the sentence expresses positive or negative opinion ([58]-[85]). For example, *"I have a beautiful wallet"* indicates positive polarity on wallet and is measured as subjective statement and so can be categorized in to several polarities.

Therefore, genuine statements are considered as objectives, for example *"The laptop is silver in colour"* presents reality with no sentiment and is classified as objective statement. Furthermore, several approaches were proposed to solve sentence-based problem, thus the realization of these approaches depending on their applications and primary domain in which they were employed makes it very difficult to standardize and evaluate the methods in order to understand the best among them [67].

3) Document based sentiment analysis

Document-level SA organize an opinion in a document by expressing a positive or negative opinion polarities. Document level ruminates the entire document as a simple information unit. Therefore, it becomes difficult and tasking to extract different sentiment about dissimilar entities independently [85]. Several applications such as Recurrent Neural Network, Convolutional Neural Network and Hybrid Neural Network are widely used to solve the document sentence-based sentiment classification, and have performed well with document level sentence [54].

Hausa WordNet (HWN), a lexical resource for Hausa language was developed, which extracts knowledge from a conservative Hausa dictionary and accepts the structure of Hindi and English WordNet as it groups words based on different categories [2].Traditional approaches are insufficient and inadequate to solve the persistent problems of opinion mining, therefore, for effectiveness a proposed model has been proposed for contextual feature incorporated with n-gram graph and many process for enhancing efficiency [19].

4) Multimodal sentiment analysis

Multimodal this is one of the evolving research directions in sentiment analysis where sentiment from videos, text, images, audio data are mined using verbal and non-verbal manners. There are quite number of customary datasets for multimodal sentiment analysis such a, YouTube, ICT-MMMO, MOUD and CMU-MOSI. "Ref. [39]" performed a review on computational methods in multimodal sentiment analysis to outline a number of researches in the field and prove how multimodal approach (textual, visual, and audio data) behaves and performs better than the traditional unimodal approaches., Similarly, together with textual review, facial and vocal expressions of multimodal approach can assist in identifying sarcastic review. DAM approach was developed for forming feature level tasks with attention device for multi-domain classification, the model was made up of two distinct segments namely; sentiment section and domain section [85].

DAM model proves to perform well for multimodal sentiment analysis when compared to machine learning approaches. Box office prediction, Stock Market prediction, Business Analytics, TV Program and News summarization, Recommender systems, politically trend prediction are all among the different application of multimodal sentiment analysis. With the rapid development in technology, individuals are geared towards audio, images, videos to indicate their opinions [8]. Therefore, several researchers

https://doi.org/10.31436/ijpcc.v12i2.700

define and express opinions or sentiment in different customs. Formerly, the text has been measured as the major key to express view which is referred as unimodal. However, multimodal sentiment analysis is the process of adopting and using multiple techniques to detect the sentiment or opinion. Henceforward, researchers are therefore concentrating on this new direction for cultivating the sentiment analysis procedure.

5) Multi-domain sentiment analysis

Multi-domain sentiment analysis focuses on a set of different documents that can be related to a particular topic, the main aim of multi-domain is to transfer information from one domain to another domain. The model will be first trained in the initial or basic domain and will later be transferred to a different domain [85]. Several researches have been conducted on multi-domain sentiment analysis where [58] provided a report on sentiment analysis related fields such as emotion detection, building resources and transfer learning. "Ref. [78]" also performed a multi-class sentiment analysis with deep learning techniques (DAM) on amazon product review data. However, DAM was proposed for forming the feature-level tasks by means of attention device for multi-domain sentiment classification [85]. Finally, Transfer learning is a process of extracting information from secondary domain to increase the learning procedure of a targeted domain [58].

B. Sentiment datasets

Researches that employ sentiment analysis can either generate data by i. generating a virgin data that fits the problem formulation and ii. using available standard data set. The main problem is extracting a label data set which is a daunting and challenging as far as sentiment analysis is concern. Therefore, researchers need to judiciously choose data set that will be accepted by research community, data set that is accessible and available. This is a publication requirement in some journal act and regulation. It is worth noting that there is insufficient amount of benchmark data set to be used in the meadow of sentiment classification [58]. Example of some standard and frequently datasets used are SemEval 2014, SemEval 2015, SemEval 2016, Twitter, Sentihood, Mitchell [94].

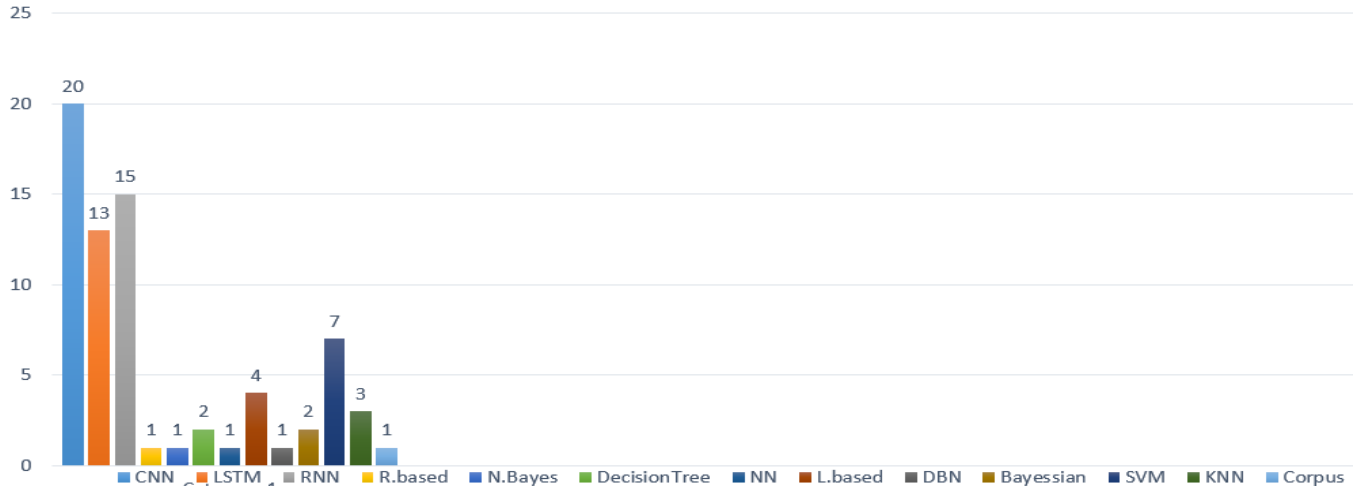


Fig. 4: Machine learning techniques used

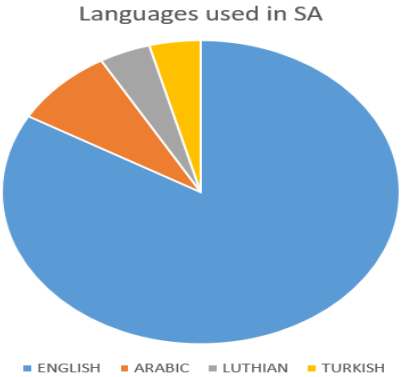


Fig 5: Language used

The Figure 4 above depicts the algorithms used and adopted in the survey paper reviewed, and Figure 5 depicts language representation of the survey conducted in this research. As such it is worth mentioning that most of the researches conducted is in English language, therefore it is important for researchers to consider other languages in order to have the best and exact sentiments of different personalities.

C. . Natural Language Processing Taxonomy

Question: What are researches conducted using machine learning to solve problem in NLP? We presented the different categories of NLP that employ machine learning algorithms to solve NLP problem. Figure:6 depicts the

https://doi.org/10.31436/ijpcc.v12i2.700

taxonomy of the natural language processing in machine language. The taxonomy presents NLP and it is categorization with respect to machine learning algorithm. The proceeding Figure 6 depicts the NLP taxonomy with several sub categories within the NLP family.

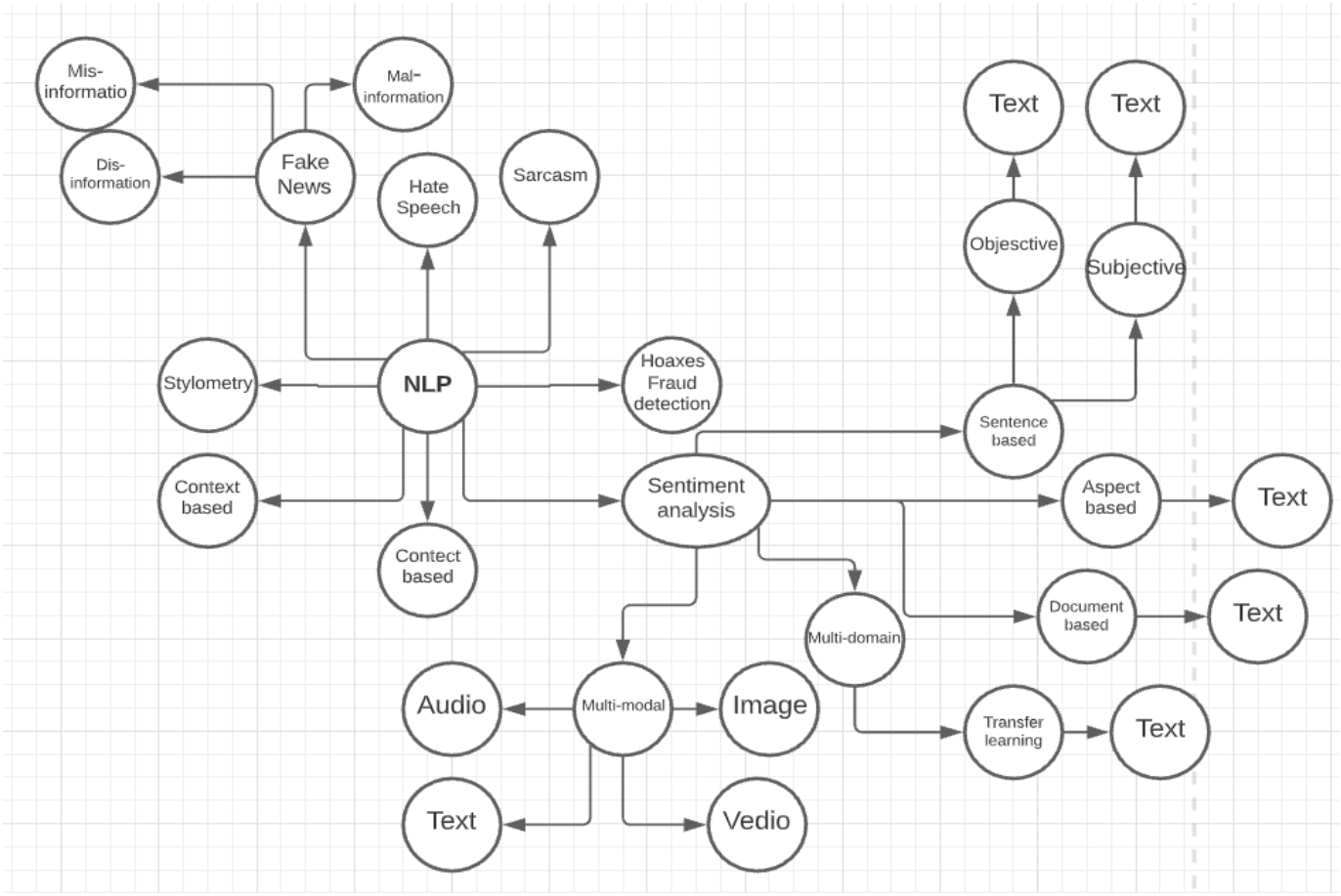


Fig 6: Taxonomy of the natural language processing in machine language

Figure 6 summarizes the relationships and trends observed across the reviewed survey papers. It provides a comparative overview of the focus areas, methodologies, and application domains addressed in existing studies. This figure supports the identification of research gaps and reinforces the need for a comprehensive Survey of Surveys. To ensure clarity and consistency throughout this study, standardized terminology is adopted for sentiment analysis categories. Specifically, "Content-based sentiment analysis" refers to approaches that analyze the textual content itself without external context. "Context-based sentiment analysis" considers surrounding contextual information such as behaviour or situation factors. "Multi-domain sentiment analysis" refers to models trained and evaluated across multiple application domain to improve generalization. This standardization ensures consistency in interpretation and improves the clarity of the taxonomy presented in this study.

1) Hate Speech

Online Hate speech is a communal way of expressing prejudice and aggression which is categorized in to three levels namely; religious, racist and gendered hate speech, in most cases tribe and racist have their own interpretation on cyber bullying, therefore each community is expected to work accordingly for intervention to eliminate such acts [9]. Also, a survey was conducted to presents a holistic view on the all the approaches, containing methods and basic algorithms used, the study also analyses the complication of conception in hate speech that has been define in several context and platforms, and offers cleaving definition [34]. Thus, social network platforms spend considerable amount of Euro to stop this act, but are still being condemn for not doing sufficient [93]. Additionally, corpus was introduced to be used for training and testing sentiment analysis systems, the researcher's with effort from academia and industries are needed be applied in order to achieve this goal, the current contribution in the recent researches is mainly in

designing whereas trying a novel schema for hate speech will be of great advantage [65]. Following the increasingly growing body of social media content the amount of online hate speech is growing on daily basis, this study is on NLP approaches focusing on feature extraction in particular in order to provide a holistic view to researchers that are new in the field of hate speech detection and want to have body knowledge in the state of the art in the trending and upcoming field [73]. Thus, most of the techniques are completely intended for common languages such as English, French, and Spanish and this serve as the major drawback of NLP nowadays, therefore, other language such as African languages and Arabic has a major challenge in using the available tools because of their peculiarities [9].

Furthermore, online hate speech is a common way of expressing prejudice and aggression among people [9]. Similarly, cyber hate is growing on a daily basis following the increasingly rising body of social media content [73]. Its growth has forced social media platforms to spend considerable amount of money to end the act, however their efforts are still being condemn as insufficient [93]. Also, survey conducted by [34] has concluded on the lack of studies and publication in automatic hate speech detection from a computing domain. Additionally, recent studies have looked into automatic hate speech detection and found out that: Deep Neural Network (DNN) structures are very effective feature extractors for detecting specific types of hate speech [93]. Finally, automatic cyber hate detection is challenging probably due to the fact that different tribe and racist have their own interpretation and characteristics of cyber hate [9].

2) *Fake News*

Fake news can be defined as an information that is deliberately false and has been broken down in to three distinct categories namely: mal-information, misinformation and disinformation [33]. Although, the enormous development of internet has urge for rapid spread of information of any kind, people are using this information without giving emphasis on it is credibility [22]. Globally, fake news dissemination is one of the major issues to both public and government sectors that raises incredible cause for alarm to researchers and academia, since people are using the internet space for disseminating information for one reason or the other [35].

Therefore, fake news detection is considered one of the most dangerous types of deception in recent time. The spread of fake news, fake reviews, fake advertisements, fake speech, misleading content and rumours has raised an important cause for alarm in the social media space. Since the use of social network has increased the spread of false content than mainstream media therefore, fake news identification and detection in social media platforms is among the most challenging task to be considered [71]. Thus, disseminating fake news and misinformation online is very easy and fast, the main challenge is in finding methods for identifying and detecting the difference between actual and false news online [64]. Moreover, inadequate consensual information can mislead user's opinion; therefore, machine learning and deep learning are referred as the best suited methods for handling problems of fake news, rumours, and misinformation [26].

Despite numerous prevailing computational solutions on fake news detection, however, lack of comprehensive remain a core barrier in the field [22]. However, with the recent development in technology, use of social media to gather, share and disseminate information has become a woven adore, people consume information from social media platform more than the traditional method as a result it has become an advanced repository of sharing and disseminating fake news and misinformation of any kind [74]. The effect of fake news has been extremely increasing, so often extending to the offline realm and intimidating public safety [63]. Social media platform has played a substantial role in portraying user's decisions and opinions in choosing or assessing a particular product, so often E-Comers customers' uses online review as their main source of getting insight about any product they intend to buy.

Therefore, product reviews and customers feedback are not the only ways opinion spam exist, misleading news and fake news is an additional form of opinion spam, the study proposed fake news detection approach using n-gram analysis via the lenses of possible feature extraction techniques, and examined two different feature extraction models with six diverse machine learning approaches [5]. Their model achieves its highest accuracy whenever uses linear SVM classifier and n-gram features.

Another study proposed end-to-end framework referred to as Event Adversarial Neural Networks (EANN) for fake news detection that is basically on multi-modal features and incorporate the event discriminator to predict the event auxiliary labels to training stage [83]. A new approach for fake news detection was proposed which incorporates sentiment as significant feature to improve the accurateness of existing models [22].

Rumours, misinformation, and fake news spread on various internet platforms just like pure and authenticated news, it's often difficult and tasking for online users to differentiate between real and fake news in such platforms [24]. Finally, the doubtful nature of news distributed by social networks, and the determination of judging its authenticity raises the need for developing computational tool to guide and assist in automatic identification of such news and information online [56].

3) Others in NLP

Deep learning and machine learning approaches especially neural network models have been identified to be one of the remarkable achievements in both sentence and document modeling in natural language processing ([46], [80]- [87]). Hybridization of some machine learning algorithm have harnessed the advantages of both algorithm as shown in the study of ([91], [92] -[93]). Also, language such as African and Arabic have challenges associated with the difficulty to adapt to available tools for feature extraction due to differences in dialect and acoustics among others.

Moreover, [2] developed a lexical resource for Hausa language which can extract knowledge from a conservative Hausa Language dictionary and accepts a structure of Hindi and English WordNet. So also [1] developed a Hausa Language Words embedding model hauWE which is bigger and better than the existing model, making them more useful in NLP tasks. Subsequently, a recent SLR conducted and highlighted that quite a number of studies exist in sentiment analysis but there is potential research gap in the domain of sarcasm identification [30].

4) Open Challenges and Future Directions

Despite significant progress in sentiment analysis, several challenges remain unresolved:

1. Data Imbalance: several datasets suffer from skewed class distributions, affecting model performance.
2. Domain Adaption: Models trained in one domain often fail to generalize effectively to other domains.
3. Multimodal Complexity: Integrating text, audio, and visual data remains a challenging task and many deep learning models lack interpretability, making their decision difficult to understand.
4. Low Resource Languages: There is limited research on languages such as Hausa, and other African languages, therefore these languages are underserved by technology.
5. Lack of Standard Evaluation Frameworks: Inconsistent evaluation metrics hinder fair comparison across studies.

Therefore, future research should focus on addressing these challenges by developing robust, interpretable and scalable models

V. . Conclusion

This paper presented a comprehensive Survey of Surveys (SoS) on sentiment analysis using machine learning. By analysing existing studies, this work provides a centralized perspective on the evolution, methodologies, and research trends in the field. The study introduced a structured taxonomy and highlighted key differences among survey papers, offering valuable insights for researchers. A cross analysis reveals that most existing surveys focus on specific techniques or domains with limited efforts toward integrating multiple approaches. Furthermore, earlier surveys emphasize algorithmic performance, while recent works increasingly incorporate deep learning and transformer-based methods. However, a lack of standardized evaluation frameworks remains a consistent limitation across studies. Additionally, several open challenges were identified, emphasizing the need for improved methodologies, standardized evaluation frameworks, and inclusion of emerging techniques such as transformer-based models.

This work serves as a foundational reference for both novice and experienced researchers, facilitating better understanding and guiding research directions in sentiment analysis. This study moves beyond descriptive aggregation by providing a structured comparative analysis of existing survey works, enabling the identification of research trends, methodological gaps, and future directions in sentiment analysis.

ACKNOWLEDGMENT

The authors would like to acknowledge the support provided by our respective institutions. Bayero University Kano and Abubakar Tafawa Balewa University.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHORS CONTRIBUTIONS

All authors contributed to the development of this manuscript

DATA AVAILABILITY

No new data were created or analyzed in this study. Data sharing is not applicable

ETHICS STATEMENT

This study does not involve humans or animal subjects

REFERENCES

[1] I. Abdulmumin and B. S. Galadanci, "HAUWE: Hausa words embedding for natural language processing," in Proc. 2nd Int. Conf. IEEE Nigeria Computer Chapter (NigeriaComputConf), 2019.

[2] A. I. Abubakar, A. Roko, A. Muhammad, and I. Saidu, "Hausa WordNet: An electronic lexical resource," Saudi J. Eng. Technol., vol. 4, no. 8, pp. 279–285, 2019.

[3] I. S. Ahmad, A. A. Bakar, and M. R. Yaakub, "A review of feature selection in sentiment analysis using information gain and domain-specific ontology," Int. J. Adv. Comput. Res., vol. 9, no. 44, pp. 283–292, 2019.

[4] I. S. Ahmad, A. A. Bakar, and M. R. Yaakub, "Movie revenue prediction based on purchase intention mining using YouTube trailer reviews," Inf. Process. Manage., vol. 57, no. 5, Art. no. 102278, 2020.
[5] H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques," in Proc. Int. Conf. Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments, 2017.
[6] A. Ahmet and T. Abdullah, "Recent trends and advances in deep learning-based sentiment analysis," in Deep Learning-Based Approaches for Sentiment Analysis. Cham, Switzerland: Springer, 2020, pp. 33–56.
[7] O. Ajao, D. Bhowmik, and S. Zargari, "Sentiment-aware fake news detection on online social networks," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2019.
[8] M. S. Akhtar, D. S. Chauhan, D. Ghosal, S. Poria, A. Ekbal, and P. Bhattacharyya, "Multi-task learning for multi-modal emotion recognition and sentiment analysis," arXiv preprint arXiv:1905.05812, 2019.
[9] A. Al-Hassan and H. Al-Dossari, "Detection of hate speech in social networks: A survey on multilingual corpus," in Proc. 6th Int. Conf. Computer Science and Information Technology, 2019.
[10] M. Aldwairi and A. Alwahedi, "Detecting fake news in social media networks," Procedia Comput. Sci., vol. 141, pp. 215–222, 2018.
[11] A. Alharbi, M. Taileb, and M. Kalkatawi, "Deep learning in Arabic sentiment analysis: An overview," J. Inf. Sci., vol. 47, no. 1, pp. 129–140, 2021.
[12] A. I. Alharbi and M. Lee, "Multi-task learning using a combination of contextualised and static word embeddings for Arabic sarcasm detection and sentiment analysis," in Proc. 6th Arabic Natural Language Processing Workshop, 2021.
[13] S. S. Aljameel et al., "A sentiment analysis approach to predict an individual's awareness of the precautionary procedures to prevent COVID-19 outbreaks in Saudi Arabia," Int. J. Environ. Res. Public Health, vol. 18, no. 1, Art. no. 218, 2021.
[14] A. Arango, J. Pérez, and B. Poblete, "Hate speech detection is not as easy as you may think: A closer look at model validation," in Proc. 42nd Int. ACM SIGIR Conf. Research and Development in Information Retrieval, 2019.
[15] M. Z. Asghar, S. Ahmad, A. Marwat, and F. M. Kundi, "Sentiment analysis on YouTube: A brief survey," arXiv preprint arXiv:1511.09142, 2015.
[16] E. Aydoğan and M. A. Akcayol, "A comprehensive survey for sentiment analysis tasks using machine learning techniques," in Proc. Int. Symp. Innovations in Intelligent Systems and Applications (INISTA), 2016.
[17] S. Afroz, M. Brennan, and R. Greenstadt, "Detecting hoaxes, frauds, and deception in writing style online," in Proc. IEEE Symp. Security and Privacy, 2012.
[18] V. Agarwal, H. P. Sultana, S. Malhotra, and A. Sarkar, "Analysis of classifiers for fake news detection," Procedia Comput. Sci., vol. 165, pp. 377–383, 2019.
[19] F. Aisopos, G. Papadakis, K. Tserpes, and T. Varvarigou, "Content vs. context for sentiment analysis: A comparative analysis over microblogs," in Proc. 23rd ACM Conf. Hypertext and Social Media, 2012.
[20] O. D. Apuke and B. Omar, "Fake news and COVID-19: Modelling the predictors of fake news sharing among social media users," Telematics Informatics, vol. 56, Art. no. 101475, 2021.
[21] B. Bharathi et al., "Machine learning and deep learning approaches for sentiment analysis," 2025.
[22] B. Bhutani, N. Rastogi, P. Sehgal, and A. Purwar, "Fake news detection using sentiment analysis," in Proc. 12th Int. Conf. Contemporary Computing (IC3), 2019.
[23] R. G. Bhati, "A survey on sentiment analysis algorithms and datasets," Rev. Comput. Eng. Res., vol. 6, no. 2, pp. 84–91, 2019.
[24] A. Bondielli and F. Marcelloni, "A survey on fake news and rumour detection techniques," Inf. Sci., vol. 497, pp. 38–55, 2019.
[25] Y. Cai, Q. Huang, Z. Lin, J. Xu, Z. Chen, and Q. Li, "Recurrent neural network with pooling operation and attention mechanism for sentiment analysis: A multi-task learning approach," Knowl.-Based Syst., vol. 203, Art. no. 105856, 2020.
[26] F. Cardoso Durier da Silva, R. Vieira, and A. C. Garcia, "Can machines learn to detect fake news? A survey focused on social media," in Proc. 52nd Hawaii Int. Conf. System Sciences, 2019.
[27] H. Touvron et al., "LLaMA: Open and efficient foundation language models," arXiv preprint arXiv:2302.13971, 2023.
[28] D. De Beer and M. Matthee, "Approaches to identify fake news: A systematic literature review," in Proc. Int. Conf. Integrated Science, 2020.
[29] N. C. Dang, M. N. Moreno-García, and F. De la Prieta, "Sentiment analysis based on deep learning: A comparative study," Electronics, vol. 9, no. 3, Art. no. 483, 2020.
[30] C. I. Eke, A. A. Norman, L. Shuib, and H. F. Nweke, "Sarcasm identification in textual data: Systematic review, research challenges and open directions," Artif. Intell. Rev., vol. 53, no. 6, pp. 4215–4258, 2020.
[31] H. H. Do, P. Prasad, A. Maag, and A. Alsadoon, "Deep learning for aspect-based sentiment analysis: A comparative review," Expert Syst. Appl., vol. 118, pp. 272–299, 2019.
[32] C. I. Eke, A. A. Norman, L. Shuib, and H. F. Nweke, "Sarcasm identification in textual data: Systematic review, research challenges and open directions," Artif. Intell. Rev., vol. 53, no. 6, pp. 4215–4258, 2020.
[33] M. K. Elhadad, K. F. Li, and F. Gebali, "Fake news detection on social media: A systematic survey," in Proc. IEEE Pacific Rim Conf. Communications, Computers and Signal Processing (PACRIM), 2019.
[34] P. Fortuna and S. Nunes, "A survey on automatic detection of hate speech in text," ACM Comput. Surv., vol. 51, no. 4, pp. 1–30, 2018.
[35] S. Girgis, E. Amer, and M. Gadallah, "Deep learning algorithms for detecting fake news in online text," in Proc. 13th Int. Conf. Computer Engineering and Systems (ICCES), 2018.
[36] G. Gravanis, A. Vakali, K. Diamantaras, and P. Karadais, "Behind the cues: A benchmarking study for fake news detection," Expert Syst. Appl., vol. 128, pp. 201–213, 2019.
[37] A. P. Garrido et al., "A systematic review of transformer models in sentiment analysis," 2025.
[38] F. Hemmatian and M. K. Sohrabi, "A survey on classification techniques for opinion mining and sentiment analysis," Artif. Intell. Rev., vol. 52, no. 3, pp. 1495–1545, 2019.
[39] M. G. Huddar, S. S. Sannakki, and V. S. Rajpurohit, "A survey of computational approaches and challenges in multimodal sentiment analysis," Int. J. Comput. Sci. Eng., vol. 7, no. 1, pp. 876–883, 2019.
[40] D. M. E.-D. M. Hussein, "A survey on sentiment analysis challenges," J. King Saud Univ.-Eng. Sci., vol. 30, no. 4, pp. 330–338, 2018.
[41] F. Jauro, H. Chiroma, A. Y. Gital, M. Almutairi, M. A. Shafi'i, and J. H. Abawajy, "Deep learning architectures in emerging cloud computing architectures: Recent development, challenges and next research trend," Appl. Soft Comput., vol. 96, Art. no. 106582, 2020.
[42] S. Kanta and G. Sidorov, "Machine learning approaches for sentiment analysis in code-mixed data," 2023.
[43] G. Kaur et al., "Comparative analysis of transformer models for sentiment classification," 2025.
[44] R. K. Kaliyar, A. Goswami, P. Narang, and S. Sinha, "FNDNet—A deep convolutional neural network for fake news detection," Cogn. Syst. Res., vol. 61, pp. 32–44, 2020.
[45] J. Kapočiūtė-Dzikienė, R. Damaševičius, and M. Woźniak, "Sentiment analysis of Lithuanian texts using traditional and deep learning approaches," Computers, vol. 8, no. 1, Art. no. 4, 2019.
[46] K. Kowsari, K. Jafari Meimandi, M. Heidarysafa, S. Mendu, L. Barnes, and D. Brown, "Text classification algorithms: A survey," Information, vol. 10, no. 4, Art. no. 150, 2019.

[47] A. Kumar and A. Jaiswal, "Systematic literature review of sentiment analysis on Twitter using soft computing techniques," Concurrency Comput. Pract. Exp., vol. 32, no. 1, Art. no. e5107, 2020.
[48] N. K. Laskari and S. K. Sanampudi, "Aspect-based sentiment analysis survey," IOSR J. Comput. Eng., vol. 18, no. 2, pp. 24–28, 2016.
[49] P. Lin and X. Luo, "A survey of sentiment analysis based on machine learning," in Proc. CCF Int. Conf. Natural Language Processing and Chinese Computing, 2020.
[50] H. Liu, I. Chatterjee, M. Zhou, X. S. Lu, and A. Abusorrah, "Aspect-based sentiment analysis: A survey of deep learning methods," IEEE Trans. Comput. Social Syst., 2020.
[51] H. Liu and B. Lang, "Machine learning and deep learning methods for intrusion detection systems: A survey," Appl. Sci., vol. 9, no. 20, Art. no. 4396, 2019.
[52] R. Liu, Y. Shi, C. Ji, and M. Jia, "A survey of sentiment analysis based on transfer learning," IEEE Access, vol. 7, pp. 85401–85412, 2019.
[53] M. M. Lopez and J. Kalita, "Deep learning applied to NLP," arXiv preprint arXiv:1703.03091, 2017.
[54] A. Mabrouk, R. P. D. Redondo, and M. Kayed, "Deep learning-based sentiment classification: A comparative survey," IEEE Access, vol. 8, pp. 85616–85638, 2020.
[55] S. Minaee et al., "Large language models in natural language processing: A survey," ACM Comput. Surv., 2023.
[56] C. V. Meneses Silva, R. Silva Fontes, and M. Colaço Júnior, "Intelligent fake news detection: A systematic mapping," J. Appl. Security Res., pp. 1–22, 2020.
[57] S. I. Manzoor and J. Singla, "Fake news detection using machine learning approaches: A systematic review," in Proc. 3rd Int. Conf. Trends in Electronics and Informatics (ICOEI), 2019.
[58] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: A survey," Ain Shams Eng. J., vol. 5, no. 4, pp. 1093–1113, 2014.
[59] A. B. Nassif, A. Elnagar, I. Shahin, and S. Henno, "Deep learning for Arabic subjective sentiment analysis: Challenges and research opportunities," Appl. Soft Comput., Art. no. 106836, 2020.
[60] A. K. Nassirtoussi, S. Aghabozorgi, T. Y. Wah, and D. C. L. Ngo, "Text mining for market prediction: A systematic review," Expert Syst. Appl., vol. 41, no. 16, pp. 7653–7670, 2014.
[61] J. A. Nasir, O. S. Khan, and I. Varlamis, "Fake news detection: A hybrid CNN-RNN-based deep learning approach," Int. J. Inf. Manage. Data Insights, vol. 1, no. 1, Art. no. 100007, 2021.
[62] G. Nguyen et al., "Machine learning and deep learning frameworks and libraries for large-scale data mining: A survey," Artif. Intell. Rev., vol. 52, no. 1, pp. 77–124, 2019.
[63] R. Oshikawa, J. Qian, and W. Y. Wang, "A survey on natural language processing for fake news detection," arXiv preprint arXiv:1811.00770, 2018.
[64] F. A. Ozbay and B. Alatas, "Fake news detection within online social media using supervised artificial intelligence algorithms," Physica A, Stat. Mech. Appl., vol. 540, Art. no. 123174, 2020.
[65] F. Poletto, M. Stranisci, M. Sanguinetti, V. Patti, and C. Bosco, "Hate speech annotation: Analysis of an Italian Twitter corpus," in Proc. 4th Italian Conf. Computational Linguistics (CLiC-it), 2017.
[66] S. Poria, D. Hazarika, N. Majumder, and R. Mihalcea, "Beneath the tip of the iceberg: Current challenges and new directions in sentiment analysis research," IEEE Trans. Affect. Comput., 2020.
[67] M. I. Prabha and G. U. Srikanth, "Survey of sentiment analysis using deep learning techniques," in Proc. 1st Int. Conf. Innovations in Information and Communication Technology (ICIICT), 2019.
[68] V. M. Pradhan, J. Vala, and P. Balani, "A survey on sentiment analysis algorithms for opinion mining," Int. J. Comput. Appl., vol. 133, no. 9, pp. 7–11, 2016.
[69] K. Ravi and V. Ravi, "A survey on opinion mining and sentiment analysis: Tasks, approaches and applications," Knowl.-Based Syst., vol. 89, pp. 14–46, 2015.
[70] D. R. Rice and C. Zorn, "Corpus-based dictionaries for sentiment analysis of specialized vocabularies," Political Sci. Res. Methods, vol. 9, no. 1, pp. 20–35, 2021.
[71] S. R. Sahoo and B. B. Gupta, "Multiple features-based approach for automatic fake news detection on social networks using deep learning," Appl. Soft Comput., vol. 100, Art. no. 106983, 2021.
[72] H. Sankar, V. Subramaniyaswamy, V. Vijayakumar, S. Arun Kumar, R. Logesh, and A. Umamakeswari, "Intelligent sentiment analysis approach using edge computing-based deep learning technique," Softw. Pract. Exp., vol. 50, no. 5, pp. 645–657, 2020.
[73] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in Proc. 5th Int. Workshop Natural Language Processing for Social Media, 2017.
[74] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," ACM SIGKDD Explor. Newsl., vol. 19, no. 1, pp. 22–36, 2017.
[75] P. Singhal and P. Bhattacharyya, "Sentiment analysis and deep learning: A survey," Center for Indian Language Technology, Indian Institute of Technology Bombay, Mumbai, India, 2016.
[76] R. Socher, Y. Bengio, and C. D. Manning, "Deep learning for NLP (without magic)," in Tutorial Abstracts of ACL 2012, 2012, p. 5.
[77] H.-C. Soong, N. B. A. Jalil, R. K. Ayyasamy, and R. Akbar, "The essential of sentiment analysis and opinion mining in social media: Introduction and survey of recent approaches and techniques," in Proc. IEEE 9th Symp. Computer Applications and Industrial Electronics (ISCAIE), 2019.
[78] S. Sukheja, S. Chopra, and M. Vijayalakshmi, "Sentiment analysis using deep learning—A survey," in Proc. Int. Conf. Computer Science, Engineering and Applications (ICCSEA), 2020.
[79] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, and H. Liu, "FakeNewsNet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media," Big Data, vol. 8, no. 3, pp. 171–188, 2020.
[80] S. Singhania, N. Fernandez, and S. Rao, "3HAN: A deep neural network for fake news detection," in Proc. Int. Conf. Neural Information Processing, 2017.
[81] M. Thilakaratne, K. Falkner, and T. Atapattu, "A systematic review on literature-based discovery workflow," PeerJ Comput. Sci., vol. 5, Art. no. e235, 2019.
[82] M. Umer, Z. Imtiaz, S. Ullah, A. Mehmood, G. S. Choi, and B.-W. On, "Fake news stance detection using deep learning architecture (CNN-LSTM)," IEEE Access, vol. 8, pp. 156695–156706, 2020.
[83] Y. Wang et al., "EANN: Event adversarial neural networks for multi-modal fake news detection," in Proc. 24th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2018.
[84] F. Weidt and R. Silva, "Systematic literature review in computer science—A practical guide," Relatórios Técnicos do DCC/UFJF, vol. 1, 2016.
[85] A. Yadav and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: A review," Artif. Intell. Rev., vol. 53, no. 6, pp. 4335–4385, 2020.
[86] P. Yang and Y. Chen, "A survey on sentiment analysis by using machine learning methods," in Proc. IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conf. (ITNEC), 2017.
[87] D. Yogatama, C. Dyer, W. Ling, and P. Blunsom, "Generative and discriminative text classification with recurrent neural networks," arXiv preprint arXiv:1703.01898, 2017.
[88] S. Yoo, J. Song, and O. Jeong, "Social media contents-based sentiment analysis and prediction system," Expert Syst. Appl., vol. 105, pp. 102–111, 2018.
[89] C. Zhang, A. Gupta, C. Kauten, A. V. Deokar, and X. Qin, "Detecting fake news for reducing misinformation risks using analytics approaches," Eur. J. Oper. Res., vol. 279, no. 3, pp. 1036–1052, 2019.

[90] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," WIREs Data Mining Knowl. Discovery, vol. 8, no. 4, Art. no. e1253, 2018.
[91] X.-D. Zhang, "Machine learning," in A Matrix Algebra Approach to Artificial Intelligence. Singapore: Springer, 2020, pp. 223–440.
[92] X. Zhang, F. Chen, and R. Huang, "A combination of RNN and CNN for attention-based relation classification," Procedia Comput. Sci., vol. 131, pp. 911–917, 2018.
[93] Z. Zhang and L. Luo, "Hate speech detection: A solved problem? The challenging case of long tail on Twitter," Semantic Web, vol. 10, no. 5, pp. 925–945, 2019.
[94] J. Zhou, J. X. Huang, Q. Chen, Q. V. Hu, T. Wang, and L. He, "Deep learning for aspect-level sentiment classification: Survey, vision, and challenges," IEEE Access, vol. 7, pp. 78454–78483, 2019.
[95] J. Zhou, J. X. Huang, Q. V. Hu, and L. He, "Is position important? Deep multi-task learning for aspect-based sentiment analysis," Appl. Intell., vol. 50, pp. 3367–3378, 2020.
[96] Z. Zhai et al., "A survey of sentiment analysis using deep learning," Artif. Intell. Rev., 2023.
[97] OpenAI, "GPT-4 technical report," arXiv preprint arXiv:2303.08774, 2023

International Journal on Perceptive and Cognitive Computing (IJPCC) Vol 12, Issue 2 (2026)
https://doi.org/10.31436/ijpcc.v12i2.673

A Decision Support System for Optimal Rice Cultivation Site Selection Using the Additive Ratio Assessment (ARAS) Method

April firman Daru[1], Rastri Prathivi[2*], Muhamad Amirudin Safa[3] , Paminto Agung Christianto[4],
[1,2,3]Faculty of Information and Communication Technology, Universitas Semarang, Tlogosari Kulon, 50196 Kota Semarang, Jawa Tengah, Indonesia
[4]Departmentof Informatics STMIK Widya Pratama ,Patriot No.25, Pekalongan, Jawa Tengah,Indonesia

*Corresponding author: vivi@usm.ac.id
(Received: 7th January 2026; Accepted: 14th April, 2026; Published on-line: 30th July, 2026)

Abstract— Optimal site selection for rice cultivation is a critical factor in enhancing agricultural productivity and ensuring regional food security, particularly in areas with heterogeneous agro-climatic conditions. However, decision-making related to rice planting locations is often constrained by fragmented data utilization and the absence of systematic multi-criteria evaluation frameworks. This study proposes a data-driven Decision Support System (DSS) based on the Additive Ratio Assessment (ARAS) method to identify optimal rice cultivation locations at a provincial scale in Central Java, Indonesia. The proposed framework evaluates 35 administrative regions (regencies and cities) using five key criteria: groundwater availability, irrigation infrastructure, harvested rice area, rice production, and rainfall intensity. Secondary data were obtained from the Central Java Statistics Bureau (BPS) and processed through a structured multi-criteria decision-making (MCDM) procedure, including decision matrix construction, normalization, weighted utility calculation, and ranking. To strengthen methodological robustness, the ARAS results were systematically compared with the MOOSRA method, and the consistency of rankings was statistically validated using Spearman's rank correlation. The results indicate that Cilacap Regency consistently achieved the highest utility score, identifying it as the most optimal location for rice cultivation, while Tegal City ranked lowest. The Spearman correlation coefficient of 0.922 demonstrates a very strong agreement between ARAS and MOOSRA rankings, confirming the reliability of the proposed approach. The novelty of this study lies in the application of an ARAS-based DSS for large-scale regional agricultural site selection, combined with cross-method validation to enhance decision reliability—an aspect rarely addressed in prior rice cultivation studies. The proposed framework provides a practical, transparent, and replicable decision-support tool for policymakers and agricultural planners, contributing to sustainable agricultural planning and evidence-based food security strategies.

Keywords— Decision Support System, ARAS, Multi-Criteria Decision Making, Rice Cultivation, Site Selection, Sustainable Agriculture.

I. INTRODUCTION

Agriculture remains a strategic sector vital for ensuring national food security and supporting economic stability, particularly in rural areas. As a staple food for the majority of the Indonesian population, rice (Oryza sativa) is a primary agricultural commodity whose productivity must be continuously enhanced to meet the demands of a growing population [1]. Central Java Province stands as one of Indonesia's primary agricultural centers, contributing significantly to national rice production. However, the optimization of rice cultivation locations within the province has not been fully addressed through systematic, data-driven decision-making approaches. Currently, land-use planning and agricultural development strategies often rely on conventional methods or partial analyses, which may not comprehensively consider the multifaceted factors influencing cultivation success [2](Connor et al., 2021). This gap highlights the need for a robust analytical framework to support more precise and effective agricultural planning. The research gap identified in this context is the limited application of structured Multi-Criteria Decision-Making (MCDM) methods within a Decision Support System (DSS) framework specifically for optimizing rice cultivation locations in Central Java. While several studies have employed MCDM techniques like Additive Ratio Assessment (ARAS) in agriculture—such as for selecting superior seeds [3] or fertilizers [4] their focus has not been on the macro-

scale spatial analysis of an entire province's regencies and cities for optimal rice planting. Furthermore, existing agricultural DSS models often lack validation through comparative analysis with other MCDM methods, limiting the robustness and reliability of their recommendations [5]. There is a distinct need for research that integrates geographical, environmental, and production data into a validated DSS to provide actionable insights for regional agricultural policy. The novelty of this study lies in its dedicated application of the ARAS method within a DSS to perform a comprehensive comparative analysis of all 35 regencies and cities in Central Java Province for optimal rice cultivation. This research introduces a validated multi-criteria model based on five key factors: groundwater availability, irrigation infrastructure, harvested rice area, rice production, and rainfall, using authoritative secondary data from the Central Java Bureau of Statistics. To strengthen the analytical rigor, the results are validated through a comparative ranking analysis with the MOOSRA method and statistically verified using the Spearman Rank Correlation test—a approach not commonly combined in previous regional agricultural planning studies in this context [6], [7]. Additionally, this study addresses a practical policy need by providing a ranked, evidence-based recommendation list to guide farmers, local governments, and stakeholders in formulating efficient and sustainable rice cultivation strategies, thereby directly contributing to regional food security and agricultural resilience [8]. By bridging the gap between theoretical MCDM application and practical agricultural land-use planning, this research offers a replicable model for data-driven agricultural decision-making in similar regional contexts.

II. Related Work

Research on decision support systems (DSS) and multi-criteria decision-making (MCDM) methods has expanded significantly in recent years to address complex selection problems in agriculture, land use, and resource management. Zavadskas and Turskis originally introduced the Additive Ratio Assessment (ARAS) method as a quantitative MCDM technique to evaluate alternatives by their overall utility relative to an ideal solution, facilitating transparent ranking in decision environments that involve multiple criteria and heterogeneous measurement units. This foundational work underpins the ARAS methodology used in the current study. [9]

In the agricultural domain, several studies have applied MCDM techniques to support cultivation decisions. For example, Suswanto et al. developed a decision support system for rice cultivation on acid sulfate soils to optimize lime and fertilizer application using experimental and economic modeling, highlighting the value of DSS for site-specific agronomic recommendations.[10]

Beyond rice cultivation, multi-criteria systems have been successfully adapted for agricultural supply chain optimization. In particular, Atlı's study integrated AHP-F (Analytic Hierarchy Process under fuzzy logic) with ARAS-F (fuzzy ARAS) to select sustainable fertilizer suppliers, illustrating how combined weighting and ranking methods can address uncertainty and sustainability concerns in agricultural resource decisions.[11]

Decision support systems utilizing ARAS have also been applied outside agriculture to demonstrate methodological versatility. For instance, Junior et al. implemented a web-based ARAS DSS for housing selection, showing how ARAS can rank alternatives effectively in complex decision contexts with diverse criteria[12] Similarly, Ernawati et al. combined entropy weighting with ARAS to evaluate and rank villages for governance outcomes, reinforcing the integration of different MCDM techniques to improve weighting robustness [13]. In the education sector, ARAS has been deployed to determine the best lecturers through ranking based on predefined performance criteria. This implementation demonstrates the flexibility of ARAS in human resource evaluation contexts[14].

Other application areas include recommender systems for customer service roles using ARAS for performance evaluation, which underscores the adaptability of ARAS in operational decision problems with multiple qualitative and quantitative criteria. While not directly ARAS-based, related MCDM research in agriculture includes decision support systems for rice seed selection using MOORA, which highlight the broader trend of employing ratio-based MCDM techniques for agronomic decisions [15]. These studies collectively show the prevalence of MCDM frameworks in agricultural DSS and their capacity to capture trade-offs between competing evaluation criteria.

Additionally, general reviews on DSS in agriculture emphasize the critical role of advanced decision systems in Agriculture 4.0, supporting precision management of resources such as soil, climate, and crop inputs. [16]. Despite extensive use of methods like AHP, MOORA, and hybrid approaches, research specifically combining ARAS with optimal site selection for rice cultivation remains sparse. This gap indicates a need for focused investigations that tailor ARAS-based DSS methodologies to agronomic site suitability problems, particularly under varying environmental and socioeconomic constraints.

III. RESEARCH METHODOLOGY

This research was conducted in several systematic stages to ensure a structured and objective analysis for determining optimal rice cultivation locations in Central Java

https://doi.org/10.31436/ijpcc.v12i2.673

Province. The overall research flow is presented in Figure 1. and detailed in the subsequent sections.

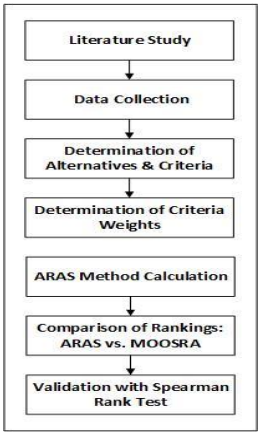


Fig 1. Research Stages

The research utilized secondary data for the year 2021, obtained from the official website of the Central Java Provincial Statistics Agency (Badan Pusat Statistik/BPS). The data comprises quantitative indicators for 35 regencies and cities (kabupaten/kota) across five primary criteria relevant to rice cultivation. The raw data is presented in Table 1.

TABLE I
RAW DATA FOR ALTERNATIVES AND CRITERIA

No.	Regency / City	Groundwater Obtainment (m^3) C1	Irrigation Channels (unit) C2	Harvested Rice Area (Ha) C3	Rice Production (Ton) C4	Daily Rainfall (mm) C5
1	Cilacap	8,017	164	1,696,712	9,618,657	3,033
2	Banyumas	7,251	282	115,664	739,140	2,696
3	Purbalingga	2,960	206	52,315	289,998	2,731
...	...	...	...	...	...	...
35	Kota Tegal	6,620	6	1,423	7,332	2,152

The alternatives consist of all 35 administrative regions (regencies and cities) in Central Java Province, coded from A1 to A35. A complete list is provided in Table 2.

TABLE II
LIST OF ALTERNATIVES

Alternative Code	Regency/City Name
A1	Cilacap
A2	Banyumas
A3	Purbalingga
A4	Banjarnegara
A5	Cilacap
A6	Purworejo
A7	Wonosobo
A8	Magelang
A9	Boyolali
A10	Klaten
A11	Sukorejo
A12	Wonogiri
A13	Karanganyar
A14	Sragen
A15	Grobokan
A16	Blora
A17	Rembang
A18	Pati
A19	Kudus
A20	Jepara
A21	Demak
A22	Semarang Regency
A23	Temanggung
A24	Kendal
A25	Batang
A26	Pekalongan
A27	Pemalang
A28	Tegal
A29	Brebes
A30	Magelang City
A31	Surakarta City
A32	Salatiga City
A33	Semarang City
A34	Pekalongan City
A35	Tegal City

Five key criteria were selected based on their direct influence on the success and productivity of rice cultivation. The criteria, along with their type (benefit or cost), are defined in Table 3.

TABLE III
CRITERIA (C_J) AND THEIR CLASSIFICATION

Criteria	Type	Description
Groundwater Obtainment	Benefit	Represents the availability of subsurface water resources, crucial for irrigation during dry periods. Higher values are better.
Irrigation Channels	Benefit	Indicates the availability of infrastructure for controlled water distribution. Higher values are better.
Harvested Rice Area	Benefit	Reflects the existing scale and land-use commitment for rice farming. Higher values are better.
Rice Production	Benefit	Represents the output and productivity efficiency of the existing rice cultivation

https://doi.org/10.31436/ijpcc.v12i2.673

Criteria	Type	Description
		system. Higher values are better.
Daily Rainfall	Cost	While necessary, excessive rainfall can cause flooding, disease, and soil waterlogging. Lower values are better.

The criteria weights (w_j) were assigned based on their relative importance in determining a suitable location for sustainable and productive rice farming. The weight values and their justification are summarized in Table 4.

TABLE 4.
CRITERIA WEIGHTS

Kode	Criteria	Weight (w_j)	Rationale
C1	Groundwater Obtainment	0.20	High importance for water security, especially in non-rainy seasons.
C2	Irrigation Channels	0.20	Critical for efficient and reliable water management and distribution.
C3	Harvested Rice Area	0.25	High weight as it directly indicates current agricultural land capacity and potential for scaling.
C4	Rice Production	0.25	High weight as it is the ultimate measure of agricultural success and output.
C5	Daily Rainfall	0.10	Classified as a cost criterion. Given its dual nature (necessary but risky in excess), it receives the lowest weight.
Sum		**1.00**	

The primary analytical framework for this study is the Additive Ratio Assessment (ARAS) method, a Multi-Criteria Decision-Making (MCDM) technique renowned for its systematic and rational approach to ranking alternatives. The procedure transforms the complex decision matrix of alternatives and criteria into a clear, quantitative ranking. The process begins with the formation of the initial decision matrix, X[m×n], incorporating all 35 alternatives (A1 to A35) and the five established criteria (C1 toC5). A critical step is the inclusion of an optimal, non-existent alternative (A0) as the first row of the matrix. The values for A0A0 are determined by selecting the maximum observed value for each benefit criterion (C1,C2,C3,C4) and the minimum observed value for the cost criterion (C5) from among the 35 real alternatives. This optimal alternative serves as the benchmark against which all others are compared.

Following the matrix construction, a normalization process is essential to eliminate the effects of differing units and scales among the criteria, rendering the values dimensionless and comparable. For criteria classified as benefit (where a higher value is preferable), each element x_{ij} is divided by the sum of all values in that criterion's column, including the value for A_o. For the cost criterion (where a lower value is preferable), a two-step transformation is applied. First, each value is inverted ($x'_{ij}=1/x_{ij}$) to convert the "lower is better" logic into a "higher is better" format consistent with benefit criteria. Subsequently, these inverted values are normalized using the same summation method applied to benefit criteria. This results in a normalized matrix where all values are proportionally scaled between 0 and 1, and the sum of each column equals unity.

The subsequent step involves weighting the normalized values to reflect the predetermined importance of each criterion. Each element in the normalized matrix is multiplied by its corresponding criterion weight (w_j), as defined in Table 4, yielding the weighted normalized matrix. This operation ensures that the more critical criteria exert a greater influence on the final score. The optimality function value (S_i) for each alternative (including A_o) is then calculated by summing the weighted normalized values across all criteria for that alternative. This S_i score represents the aggregate performance of the alternative.

The final and most crucial step in the ARAS method is the calculation of the utility degree (K_i) for each real alternative (i=1,...,m). The utility degree is derived by dividing the optimality function of the alternative (S_i) by the optimality function of the ideal reference alternative (S_o). Mathematically, $K_i=S_i/S_o$. Since S_o is derived from the optimal values, K_o equals 1. The resulting K_i values, which range between 0 and 1, quantify the relative closeness of each alternative to the theoretically optimal solution. The alternatives are then ranked in descending order of their K_i values; the alternative with the highest utility degree is deemed the most optimal location for rice cultivation in

https://doi.org/10.31436/ijpcc.v12i2.673

Central Java Province, providing a clear, data-driven hierarchy for decision-making.

IV. ARAS METHOD IMPLEMENTATION AND RESULTS

The implementation of the ARAS method commenced with the construction of the initial decision matrix (X). This matrix integrates the 35 real-world alternatives (A1 to A35), representing the regencies and cities of Central Java, with the five evaluation criteria (C1 to C5). A pivotal step in the ARAS framework is the introduction of an optimal reference alternative, designated as A0. The values for A0 for each criterion are derived from the dataset itself: for the benefit criteria (C1: Groundwater, C2: Irrigation, C3: Harvested Area, C4: Production), the value is the maximum observed among all 35 alternatives. For the cost criterion (C5: Rainfall), the value is the minimum observed. This constructs A0 as a hypothetical, ideal location possessing the best possible attribute values. The initial segment of this matrix is presented in Table 5.

TABLE IV.
SEGMENT OF THE INITIAL DECISION MATRIX (X)

Alt.	C1 (Benefit)	C2 (Benefit)	C3 (Benefit)	C4 (Benefit)	C5 (Cost)
A0	156,688	331	1,696,712	9,618,657	1,248
A1	8,017	164	1,696,712	9,618,657	3,033
A2	7,251	282	115,664	739,14	2,696
A3	2,96	206	52,315	289,998	2,731
...	...	...	...	...	...

The Additive Ratio Assessment (ARAS) method is a multi-criteria decision-making (MCDM) approach employed within a decision support system (DSS) framework to evaluate and rank alternatives based on their relative utility compared to an ideal solution. In this study, ARAS was applied to determine the most suitable locations for rice cultivation across 35 regencies and cities in Central Java Province, Indonesia. The analysis utilized secondary data from the Central Java Provincial Bureau of Statistics (BPS) for 2021, focusing on five key criteria: groundwater availability (C1, weight 0.20, benefit type), irrigation channels (C2, weight 0.20, benefit type), harvested rice area (C3, weight 0.25, benefit type), rice production (C4, weight 0.25, benefit type), and rainfall (C5, weight 0.10, cost type). These criteria were selected to reflect resource availability, infrastructure support, productivity potential, and environmental risks.

The initial step entailed forming a decision matrix (X) comprising m alternatives (35 regencies/cities) and n criteria (5 factors). Each entry x_{ij} represented the performance value of alternative i on criterion j, derived directly from the BPS dataset. An optimal row (X_0) was added as a reference, where for benefit criteria, $x_{0j} = \max(x_{ij})$, and for cost criteria, $x_{0j} = \min(x_{ij})$. This matrix served as the foundation for subsequent normalization, capturing raw values such as groundwater volumes, irrigation lengths, harvested areas in hectares, production in tons, and average daily rainfall in millimeters.

To standardize the matrix and enable comparability across diverse units, normalization was performed. For benefit criteria (C1–C4), the normalized value r_{ij} was calculated as $r_{ij} = x_{ij} / \Sigma x_{kj}$ (for k = 0 to m). For the cost criterion (C5), values were first inverted ($1 / x_{ij}$) to align with the "higher is better" logic, then normalized similarly: $r_{ij} = (1 / x_{ij}) / \Sigma (1 / x_{kj})$. This resulted in a normalized matrix where column sums equaled 1, eliminating scale discrepancies and facilitating weighted aggregation.

The normalized values were then multiplied by their respective criterion weights (w_j) to produce the weighted normalized matrix (V), with $v_{ij} = r_{ij} \times w_j$. Weights were predetermined based on expert judgment and criterion importance, emphasizing productivity (C3 and C4 at 0.25 each) while balancing resource and risk factors. This step integrated the relative significance of each criterion into the evaluation.

The optimality function S_i for each alternative i was computed by summing the weighted normalized values across all criteria: $S_i = \Sigma v_{ij}$ (for j = 1 to n). The ideal optimality value S_0 was derived from the optimal row. The utility degree K_i, representing the relative performance of each alternative against the ideal, was then obtained as $K_i = S_i / S_0$. Higher K_i values indicated greater suitability for rice cultivation.

Alternatives were ranked in descending order of K_i, yielding a prioritized list of locations. This ranking provided actionable insights for stakeholders, highlighting areas with optimal combinations of criteria.

The Additive Ratio Assessment (ARAS) method was applied to rank 35 regencies and cities in Central Java Province based on their suitability for rice cultivation. The evaluation incorporated five weighted criteria using 2021 BPS data. The final utility degree (K_i) reflects the relative performance of each alternative against the ideal solution, with higher values indicating greater optimality.

TABLE V
THE MOST AND LEAST OPTIMAL LOCATIONS

Rank	Regency/City	Utility Degree (K_i)
1	Cilacap Regency	0.808673
2	Kudus Regency	0.226189
3	Jepara Regency	0.150258

https://doi.org/10.31436/ijpcc.v12i2.673

Rank	Regency/City	Utility Degree (K_i)
4	Pati Regency	0.140171
5	Semarang Regency	0.135032
6–26	Other regencies/cities	0.005–0.130 (intermediate)
27–30	Other regencies/cities	<0.030
31	(Intermediate low)	~0.025
32	(Intermediate low)	~0.020
33	(Intermediate low)	~0.015
34	(Intermediate low)	~0.010
35	Tegal City	0.023481

Table 5 highlights the most and least optimal locations and summarizes intermediate performance. The high utility scores for the top-ranked regencies underscore their dominant role in regional rice production, driven by extensive irrigated land and high yields. Lower-ranked urban areas reflect inherent limitations of agricultural scale. The full ranking (35 alternatives) confirms Cilacap Regency as a benchmark for optimal rice cultivation potential in Central Java.

V. RESULT AND EVALUATION

This study investigates the identification of optimal sites for rice cultivation in Central Java Province, Indonesia, employing a decision support system (DSS) grounded in the Additive Ratio Assessment (ARAS) method. Utilizing secondary data from the Central Java Provincial Bureau of Statistics (BPS) for 2021, the analysis evaluated 35 administrative units (regencies and cities) against five weighted criteria: groundwater availability (weight: 0.20, benefit-oriented), irrigation infrastructure (weight: 0.20, benefit-oriented), harvested rice area (weight: 0.25, benefit-oriented), rice production volume (weight: 0.25, benefit-oriented), and rainfall intensity (weight: 0.10, cost-oriented). These criteria were chosen to encompass hydrological resources, infrastructural support, agronomic capacity, output efficiency, and climatic hazards, ensuring a holistic assessment of cultivation feasibility.

The ARAS methodology proceeded through decision matrix formulation, normalization to standardize disparate units, weighted aggregation to reflect criterion priorities, computation of optimality functions, and derivation of relative utility degrees (K_i) for ranking. The outcomes delineated a clear hierarchy of suitability. Cilacap Regency achieved the highest utility degree of 0.808673, signifying its preeminence due to abundant groundwater, robust irrigation networks, expansive harvested areas, and superior production yields, which collectively mitigate rainfall-related risks. Subsequent rankings included Kudus Regency (K_i = 0.226189), Jepara Regency (K_i = 0.150258), Pati Regency (K_i = 0.140171), and Semarang Regency (K_i = 0.135032), highlighting regions with strong agronomic profiles and resource balance. In contrast, Tegal City recorded the lowest utility degree of 0.023481, attributable to constrained arable land, inadequate irrigation, and urban-dominated landscapes that limit agricultural scalability. Intermediate positions, such as those occupied by Brebes and Grobogan Regencies, reflected moderate potentials influenced by variable resource distributions.

To assess the robustness and validity of the ranking generated by the ARAS method, a comparative analysis was conducted using an alternative Multi-Criteria Decision-Making (MCDM) technique: the Multi-Objective Optimization based on Simple Ratio Analysis (MOOSRA). This method provides a distinct computational approach for ranking alternatives, serving as an effective benchmark. While ARAS calculates utility by comparing the sum of weighted normalized values to an ideal solution, MOOSRA determines performance by calculating the ratio of the sum of normalized values for benefit criteria to the sum for cost criteria, each weighted appropriately. The comparative rankings from both methods are presented in Table 6.

TABLE VI
COMPARATIVE RANKING: ARAS VS. MOOSRA METHODS

Alternative	ARAS Rank	MOOSRA Rank	Rank Difference (d_i)
A1 (Cilacap)	1	1	0
A19 (Kudus)	2	3	-1
A20 (Jepara)	3	12	-9
A18 (Pati)	4	2	2
A22 (Semarang Regency.)	5	7	-2
A33 (Semarang City)	6	8	-2
A2 (Banyumas)	7	9	-2
A16 (Blora)	8	6	2
A15 (Grobogan)	9	5	4
A10 (Klaten)	10	11	-1
...	...	...	...
A35 (Tegal City)	35	35	0

The comparative analysis reveals a high degree of concordance between the two methods for the extreme ranks and several key positions, while highlighting interpretable differences for others. A strong consensus is evident at the top: Cilacap Regency (A1) is unanimously ranked first by both ARAS and MOOSRA, solidifying its status as the most optimal location. Similarly, Kota Tegal (A35) is consistently ranked last by both methods. This

https://doi.org/10.31436/ijpcc.v12i2.673

agreement on the best and worst performers provides strong initial validation for the ARAS outcome. Significant alignment is also observed for several other alternatives. For instance, Purbalingga (A3), Banjarnegara (A4), Purworejo (A6), Wonosobo (A7), and Karanganyar (A13) hold identical ranks in both lists. This consistency across different algorithmic logics suggests that the performance profiles of these regions are clear and unambiguous, leading to stable rankings regardless of the specific MCDM technique employed. However, notable rank differences exist for certain alternatives, which are not indicative of error but rather of the different sensitivities inherent to each method.

The most pronounced difference is for Jepara Regency (A20), ranked 3rd by ARAS but 12th by MOOSRA. This discrepancy can be attributed to Jepara's data profile: it scores very highly on the major benefit criteria of Groundwater (C1) and has moderate rainfall (C5), which is favorable in ARAS's additive utility model. However, its scores for Harvested Area (C3) and Production (C4) are relatively low. MOOSRA's ratio-based model (sum of benefits / sum of costs) may penalize this imbalance more severely, especially given the high weights on C3 and C4, causing its rank to drop. Similarly, Grobogan (A15) and Rembang (A17) are ranked higher by MOOSRA (5th and 4th) than by ARAS (9th and 11th). Both regions have strong production figures (C4) and, critically, the lowest or very low rainfall values (C5). MOOSRA's ratio formula significantly rewards a low cost criterion denominator, providing a substantial boost to their composite scores.

The Spearman Rank validation is applied to examine the consistency of ranking results produced by the Additive Ratio Assessment (ARAS) method by comparing them with the MOOSRA method. Spearman's rank correlation is appropriate in this context because both methods generate ordinal rankings of alternatives, rather than absolute measurements.

Spearman's Rank Correlation Coefficient (ρ) measures the strength and direction of the monotonic relationship between two ranking results. It is calculated based on the differences between the ranks assigned by two methods to the same alternatives. A coefficient value close to +1 indicates a very strong positive correlation, meaning that both methods produce highly similar rankings, whereas a value close to 0 indicates weak consistency. In this study, ARAS rankings for optimal rice planting locations in Central Java Province are validated against MOOSRA rankings using Spearman Rank analysis. This validation aims to ensure that the ARAS-based decision-making results are reliable and not method-dependent. The implementation compares the rank of each alternative generated by ARAS and MOOSRA, calculates the rank difference (d), and then computes d^2. A summary of this process is shown in Table 7.

TABLE VII
SPEARMAN RANK VALIDATION BETWEEN ARAS AND MOOSRA

Alternative	Rank (ARAS)	Rank (MOOSRA)	d	d^2
A1	1	1	0	0
A2	7	9	-2	4
...	...	...	...	...
A35	35	35	0	0
Σ d^2				**554**

Based on these results, the calculated Spearman Rank correlation coefficient is ρ = 0.92241 The obtained coefficient value of 0.92241 indicates a very strong positive correlation between the ARAS and MOOSRA ranking results. This demonstrates that both methods exhibit a high level of agreement in evaluating and ranking the optimal rice planting locations. Several alternatives, such as A1, A3, A4, A6, A7, A13, and A35, show identical or nearly identical ranks, further reinforcing this consistency. Although some alternatives display ranking differences, these variations do not significantly affect the overall correlation strength. Therefore, the Spearman Rank validation confirms that the ARAS method produces stable, consistent, and reliable ranking outcomes when compared with an established comparison method. Consequently, ARAS can be considered valid for multi-criteria decision-making in the context of determining optimal rice planting locations in Central Java Province

VI. Conclusions

This study employed the Additive Ratio Assessment (ARAS) method to identify optimal rice planting locations in Central Java Province based on 35 alternatives and five evaluation criteria. The results indicate that Cilacap Regency achieved the highest preference value (0.808673), followed by Kudus, Jepara, Pati, and Semarang Regencies, which can be prioritized for rice cultivation planning. In contrast, Tegal City exhibited the lowest suitability for rice planting. The robustness of the ARAS results was verified through comparison with the MOOSRA method. The analysis yielded a Spearman rank correlation coefficient of 0.92241, indicating a very strong positive correlation and a high level of consistency between both multicriteria decision-making approaches. This confirms the validity and reliability of ARAS for determining optimal rice planting locations. Future research should incorporate field-based observations to capture site-specific conditions not fully represented in secondary data, such as irrigation performance, soil quality, and socio-economic factors. Additionally, integrating GIS and satellite imagery with multicriteria decision-making

methods is recommended to enhance spatial accuracy and support sustainable agricultural planning in Central Java Province.

ACKNOWLEDGMENT

The authors would like to acknowledge the support provided by Universitas Semarang, Tlogosari Kulon, 50196 Kota Semarang, Jawa Tengah, Indonesia and STMIK Widya Pratama, Patriot. Indonesia.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHORS CONTRIBUTIONS

All authors contributed to the development of this paper

DATA AVAILABILITY

Data for this study are cited inside the paper

ETHICS STATEMENT

This study does not involve required ethical approval

REFERENCES

[1] B. Santoso, S. Sugihardjo, and Y. Widyastuti, "Peran sektor pertanian dalam perekonomian Indonesia: Pendekatan input-output," *Jurnal Ekonomi dan Kebijakan Pembangunan*, vol. 9, no. 1, pp. 1–19, 2020.

[2] M. Connor, A. de Guia, A. Pustika, Sudarmaji, M. Kobarsih, and J. Hellin, "Rice farming in central Java, Indonesia—adoption of sustainable farming practices, impacts and implications," *Agronomy*, vol. 11, no. 5, p. 881, 2021, doi: 10.3390/agronomy11050881.

[3] D. Zai, A. Nasyuha, and B. Andika, "Sistem Pendukung Keputusan Menentukan Bibit Padi Yang Terbaik Di Desa Gempola Tebing Tinggi Dengan Metode ARAS," *Jurnal Cyber Tech*, vol. 2, no. 8, 2019.

[4] M. Purba, V. Sihombing, and D. Irmayani, "Sistem Pendukung Keputusan Pemilihan Lokasi Pemasaran Laptop Bekas Menggunakan Metode ARAS," *Jurnal Teknik Informasi Dan Komputer (Tekinkom)*, vol. 7, no. 2, pp. 828–836, 2024.

[5] A. Kumar and S. Pant, "Analytical hierarchy process for sustainable agriculture: An overview," *MethodsX*, vol. 10, p. 101954, 2023, doi: 10.1016/j.mex.2022.101954.

[6] A. Mardani, A. Jusoh, K. Nor, Z. Khalifah, N. Zakwan, and A. Valipour, "Multiple criteria decision-making techniques and their applications – a review of the literature from 2000 to 2014," *Economic Research-Ekonomska Istraživanja*, vol. 28, no. 1, pp. 516–571, 2015, doi: 10.1080/1331677X.2015.1075139.

[7] E. Zavadskas, A. Mardani, Z. Turskis, A. Jusoh, and K. Nor, "Development of TOPSIS method to solve complicated decision-making problems — An overview on developments from 2000 to 2015," *International Journal of Information Technology & Decision Making*, vol. 15, no. 03, pp. 645–682, 2016, doi: 10.1142/S0219622016300019.

[8] R. K. Singh, R. Berkvens, and M. Weyn, "AgriFusion: An architecture for IoT and emerging technologies based on a precision agriculture survey," *IEEE Access*, vol. 9, pp. 136253–136283, 2021.

[9] B. Marr, *The Future of AI in Business: How Artificial Intelligence Will Transform Organizations*. London: Kogan Page, 2023.

[10] T. Suswanto, S. Jusop, S. R. S. Omar, and P. Mat, "A Decision Support System for Rice Cultivation on Acid Sulfate Soils in Malaysia," *Jurnal Ilmu Tanah dan Lingkungan*, vol. 7, no. 1, pp. 1–5, 2005, doi: 10.29244/jitl.7.1.1-5.

[11] H. F. Atlı, "Sustainable Supplier Selection Using Fuzzy AHP and Fuzzy ARAS Techniques for Fertilizer Supply in the Agricultural Supply Chain," *Turkish Journal of Agriculture-Food Science and Technology*, vol. 12, no. 8, pp. 1269–1280, 2023, doi: 10.24925/turjaf.v12i8.1269-1280.6568.

[12] F. Junior, R. W. Darmawan, and S. Petriyansyah, "Design of a Decision Support System for Housing Selection Using the Web-Based Additive Ratio Assessment (Aras) Method," *Technology and Society Perspectives (TACIT)*, vol. 2, no. 2, pp. 252–262, July 2024, doi: 10.61100/tacit.v2i2.205.

[13] A. Ernawati, A. Ofta Sari, S. N. Sofyan, A. Aulia, Z. Sitorus, and Khairul, "Implementasi Sistem Pendukung Keputusan dalam menentukan Kecamatan Terbaik Menggunakan Algoritma Entropy dan Additive Ratio Assessment (ARAS)," *Bulletin of Information Technology (BIT)*, vol. 4, no. 4, pp. 488–500, Dec. 2023, doi: 10.47065/bit.v4i4.1066.

[14] D. Aldo, Y. Nur, and A. Utami, "Implementation of Additive Ratio Assessment Method in Decision Making on Fertilizer Supplier Selection at BBTPH Banyumas," *Jurnal Ecotipe (Electronic, Control, Telecommunication, Information, and Power Engineering)*, vol. 10, no. 2, pp. 237–246, 2023.

[15] Y. Syawali, D. Y. Niska, M. H. H. Rangkuti, and K. A. Mayadi, "Decision Support System for Optimizing the Selection of the Best Rice Seeds Using the MOORA Method," *Jurnal CoSciTech (Computer Science and Information Technology)*, vol. 6, no. 1, pp. 45–54, May 2025, doi: 10.37859/coscitech. v6i1.9104.

[16] Z. Zhai, J. F. Martínez, V. Beltran, and N. L. Martínez, "Decision Support Systems for Agriculture 4.0: Survey and Challenges," *Computers and Electronics in Agriculture*, vol. 175, p. 105256, 2020, doi: 10.1016/j.compag.2020.105256.

International Journal on Perceptive and Cognitive Computing (IJPCC) Vol 12, Issue 2 (2026)
https://doi.org/10.31436/ijpcc.v12i2.717

Feasibility and Acceptance of Decision-Based Games for the Assessment of Anxiety and Depression

Suhaila Samsuri,* Zakaria Mohiuddin Khandaker, Syamil Rosli, Maria Wardah Mohd Ghazali, Ahmad Lutfi Ruhaimi

Kulliyyah of Information and Communication Technology, International Islamic University Malaysia, Gombak, Selangor

*Corresponding author: suhailasamsuri@iium.edu.my
(Received: 7th March 2026; Accepted: 15th April, 2026; Published on-line: 30th July, 2026)

Abstract — Traditional mental health assessments for anxiety and depression often face limitations due to their overtly clinical nature and susceptibility to self-report and social desirability biases. To overcome these barriers, this preliminary study explores the feasibility and user acceptance of decision-based games as immersive, early-screening tools specifically targeting university students. Employing a qualitative descriptive design, the research integrated a multimedia production methodology with semi-structured interviews and user acceptance testing (UAT) involving clinical psychologists, academic experts, and student volunteers. The findings indicate that decision-based games can successfully and practically integrate standard psychometric tools, such as the GAD-7 and PHQ-9, yielding high user acceptance. However, experts stressed that to maintain psychological safety and clinical validity, these games must be strictly classified as preliminary screening tools rather than diagnostic instruments, and they must retain the exact original wording of the assessments. Furthermore, targeted design mechanics, including culturally resonant scenarios, minimalist anxiety-sensitive visuals, and brief 5- to 10-minute gameplay sessions, significantly enhance user engagement. Ultimately, well-designed decision-based games offer a viable, engaging, and less stigmatizing alternative to traditional questionnaires for routine mental health screening.

Keywords — Decision-based games, Mental health screening, GAD-7, PHQ-9, University students, Gamification in healthcare

I. Introduction

Anxiety and depression remain among the most prevalent and disabling mental disorders worldwide, placing a substantial burden on individuals, families, and healthcare systems. Accurate, early, and routine assessment of these conditions is critical for timely intervention. However, traditional assessment methods —such as clinician-administered interviews and self-report questionnaires — are often constrained by limited accessibility, high costs, heavy time demands, and susceptibility to response biases or social desirability. These limitations have spurred growing interest in alternative digital approaches that can capture mental health constructs in naturalistic, engaging, and scalable ways [1][2][3].

Serious games and digital interventions have emerged as promising tools for mental health assessment and treatment. By offering immersive environments, game-based approaches lessen the evaluation anxiety and participant burden often triggered by conventional testing, while eliciting more authentic, spontaneous behaviour [2]. Recent research indicates that game-based tools demonstrate acceptable usability, feasibility, and sustained engagement among university students experiencing heightened anxiety and stress [4][5]. Nevertheless, while much of the existing literature evaluates digital games as therapeutic or psychoeducational interventions [6], comparatively less attention has been devoted to their role as objective assessment instruments — particularly those that rely on dynamic, decision-based mechanics to infer underlying affective and cognitive states [3].

Decision-based games, in which players make choices under conditions of uncertainty, risk, or competing goals, can operationalize cognitive and emotional processes central to depression and anxiety, such as threat sensitivity, reward responsiveness, hopelessness, and perseveration. By embedding diagnostic constructs into interactive decision scenarios, game-based assessments capture subtle behavioural markers that static questionnaires struggle to elicit [8]. Despite their theoretical appeal and potential to reduce test-taking anxiety, the real-world feasibility and clinical acceptability of decision-based games remain underexplored [1].

This study addresses this gap by examining the feasibility and acceptance of decision-based games for assessing anxiety and depression in university students. Specifically, we evaluate feasibility across implementation practicality, data quality, and participant adherence, alongside acceptance in terms of perceived usefulness, engagement,

https://doi.org/10.31436/ijpcc.v12i2.717

and trust from both end-users and clinical raters. By synthesizing the perspectives of clinical psychologists and diagnosed individuals, these findings aim to inform the development of scientifically rigorous and practically viable game-based instruments for primary care, digital mental health platforms, and academic environments.

II. Literature Review

A scoping review on the use of video games and virtual reality (VR) for assessing anxiety and depression reveals that emerging evidence is promising but remains limited [9]. Out of ten eligible studies, only two were clinical trials, and three utilized machine-learning techniques to derive assessment metrics [9]. Studies focusing on anxiety disorders predominantly relied on VR, whereas those targeting depression more frequently evaluated video games [9]. While most findings were encouraging — identifying potential digital biomarkers such as risk-taking patterns, persistence, error rates, and response timing — the limited number of studies highlights a need for extensive validation before games or VR can serve as reliable assessment tools [9]. Complementary research on digital mental health tools underscores that while technology will play a pivotal role in future psychiatric care, comprehensive appraisals of screening and diagnostic accuracy are required [10]. This requirement is particularly critical for decision-based games, which seek to infer psychological states from player choices rather than self-report scales [10].

Although systematic reviews and meta-analyses demonstrate that gaming interventions can effectively treat depression and anxiety in adolescents and young adults, most existing research focuses on therapeutic outcomes rather than clinical assessment [11]. Nevertheless, these interventions offer valuable precedent by showing that game mechanics can successfully engage users, elicit task performance, and generate behavioural data that correlates with symptom severity [11]. These findings support the feasibility of operationalizing core cognitive and emotional constructs — such as threat sensitivity, hopelessness, avoidance, and perseveration — into interactive assessment indicators [11]. Empirical support for game-based assessments also stems from studies in other domains, such as tablet-based games evaluating cognitive status in older adults and RPG-based behavioural health tools for adolescents, which consistently report high participation rates, strong usability, and positive user engagement [12].

The feasibility of implementing game-based screening depends heavily on the attitudes and acceptance of mental health professionals. Research indicates generally positive yet cautious perspectives among clinicians [13]. For example, a South African study found that while only 16% of therapists were familiar with serious games, 71% considered them a suitable adjunct treatment modality [13]. Earlier survey data similarly revealed that 90% of therapists could envision therapeutic uses for serious games, particularly targeting anxiety and affective disorders [13]. More recent qualitative work with Finnish mental health professionals highlights complex, ambivalent views; practitioners often distinguish between their own recreational gaming and client gaming, viewing the latter with caution regarding ethics, validity, and potential overuse [14]. Nonetheless, clinicians acknowledge that game-based tools can offer approachable, motivating alternatives — especially for youth, digitally fluent individuals, and withdrawn clients — provided ethical and validity concerns are thoroughly addressed [14].

Broader evaluations of serious games in mental health settings consistently report strong feasibility and user acceptability when applications feature intuitive interfaces, coherent narratives, and active service-user involvement [15]. Qualitative feedback confirms that patients are open to engaging with gamified platforms and view them as less intimidating than traditional methods [15]. Despite these encouraging trends, a significant gap remains in research specifically examining the feasibility and acceptability of decision-based games as assessment instruments for anxiety and depression from the direct perspective of practicing psychologists. Existing literature remains heavily skewed toward therapeutic interventions or cognitive assessments.

Collectively, current research indicates that while game-based approaches show preliminary viability for evaluating affective conditions, robust clinical evidence remains scarce [9]. Mental health professionals view these tools as promising adjuncts but express valid concerns regarding psychometric integrity, ethical boundaries, and practical workflow integration [16]. Feasibility and acceptability remain highest when games are well-designed, yet systematic evaluation of game-based assessment tools in clinical settings remains limited [17]. These findings justify a qualitative investigation directly exploring psychologists' perspectives to clarify perceived benefits, implementation barriers, and clinical conditions, ultimately informing the development of scientifically valid and practically viable game-based assessment tools.

III. Methodology

A. Study Design

This study employed a qualitative, descriptive design using semi-structured interviews to explore psychologists' perspectives on the feasibility of using decision-based games as mental health assessment tools for anxiety and depression. The qualitative approach was chosen to capture nuanced views on clinical utility, barriers, and conditions for implementation that are not readily measurable through

https://doi.org/10.31436/ijpcc.v12i2.717

quantitative surveys [18]. Refer to the Figure 1 below for the study design diagram.

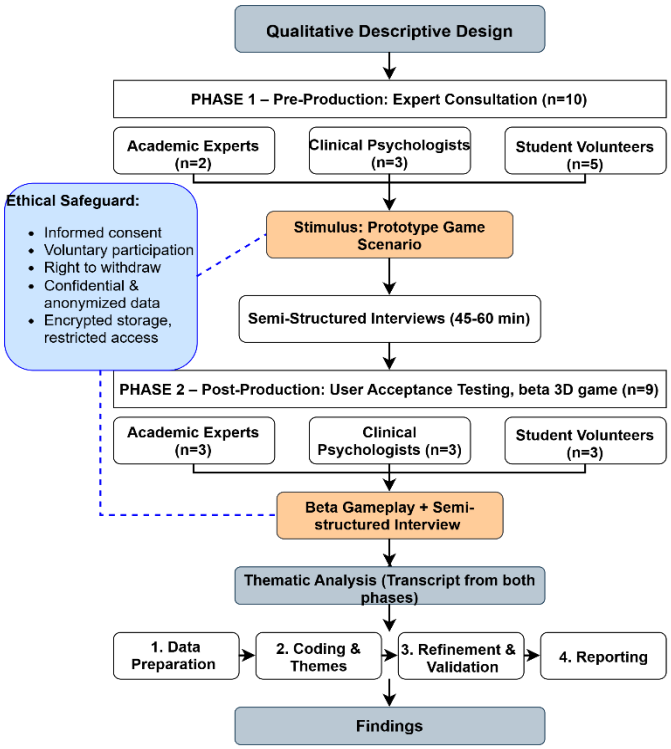


Fig. 1 Study Design Diagram

B. *Participants of Data Collections Across the Studies*

The data collection primarily utilized expert sampling for pre-production of game phase using semi-structured interviews, followed by post-production of game phase for user testing with the target demographic. The participants are detailed below by project:

Pre-production phase, this preliminary study utilized an expert consultation methodology, gathering data from a total of 10 participants divided into four groups.

1) Academic Experts (n=2): Specialists from the International Islamic University Malaysia (IIUM) with over 20 years of experience in fields such as game-based learning and mental health interventions. The experts gave important insights into demographic targeting, ethical considerations, the use of the Likert scale for anxiety detection, and appropriate game design.
2) Clinical Psychologists (n=3): A team of practicing mental health professionals recruited from university counselling centers, possessing over 10 years of experience in clinical psychology, anxiety assessment, and mental health screening.
3) Student Volunteers (n=5): A focus group of university students representing the target demographic. These participants were specifically recruited based on having either a prior anxiety diagnosis or self-reported anxiety symptoms.

Post-production (beta version of 3D game) for User Acceptance Test (UAT),

1) Academic Experts (n=3): Specialists from the International Islamic University Malaysia (IIUM) with over 20 years of experience in fields of mental health interventions.
2) Clinical Psychologists (n=3): A team of practicing mental health professionals recruited from university and private counselling centers, who conducted clinical psychology, anxiety assessment, and mental health screening.
3) University students (n=3), who were diagnosed and undiagnosed with anxiety and depression disorder, were selected from various faculties and academic years at IIUM. They played through the application to evaluate its usability, narrative flow, cultural appropriateness, and emotional impact.

C. *Intervention/Stimulus*

To ground participants' responses in a concrete example, all participants were briefly introduced to a prototype decision-based game scenario (either a short video demonstration, a written scenario, or a brief interactive demo) designed to capture decision patterns relevant to anxiety and depression (e.g., risk-taking, avoidance, persistence, hopelessness). The game was not administered for assessment purposes in this study; rather, it served as a stimulus to elicit discussion about feasibility and clinical relevance.

Descriptions of the game's purpose, mechanics, and intended assessment indicators were provided to ensure participants had a shared understanding of what was being discussed.

D. *Data Collection*

Data were collected through individual semi-structured interviews, conducted either:

1) face-to-face at participants' workplaces or a neutral location, or
2) via secure online video conferencing platforms.

Each interview lasted approximately 45–60 minutes and was conducted by researcher(s) with appropriate training in qualitative methods and mental health contexts. Interviews followed a structured interview protocol that included:

1) Background and context: Professional role, experience in assessment of anxiety/depression, and use of digital tools.
2) Perceived feasibility: Practicality of integrating decision-based games into current assessment

https://doi.org/10.31436/ijpcc.v12i2.717

workflows. Time, training, technical, and organizational requirements.

3) Perceived clinical utility: Relevance of game-derived indicators for anxiety and depression. Potential advantages over traditional questionnaires or interviews.
4) Barriers and concerns: Validity, reliability, interpretability, and ethical considerations (e.g., privacy, consent, security). Suitability across patient groups (e.g., age, cultural background, neurodiversity).
5) Conditions for implementation: Training needs, validation requirements, and partnerships with developers or researchers.

Recommendations: Design features, implementation strategies, and future research directions.

Interviews were audio-recorded with participants' consent. When the recording was not possible, detailed notes were taken immediately after each interview.

E. Ethical Considerations

Participants were provided with an information sheet describing the study's aims, procedures, and their rights. Informed consent was obtained prior to participation. Participants were informed that:

1) participation was voluntary,
2) they could decline to answer any question,
3) they could withdraw at any time without consequence, and
4) all data would be confidential and anonymized.

Audio recordings and transcripts were stored on encrypted devices, accessible only to the research team.

F. Data Analysis

Interview data were analysed using thematic analysis, following a structured approach consistent with established guidelines [19].

1) *Data preparation:*

- Audio recordings were transcribed verbatim.
- Transcripts were checked for accuracy and anonymized (removing names and identifiable information).

2) *Coding and theme development:*

- Initial familiarization with the data was conducted by reading all transcripts and noting preliminary ideas.
- A coding framework was developed, beginning with broad categories aligned with the study objectives (e.g., feasibility, clinical utility, barriers, implementation conditions).
- Transcripts were coded systematically using NVivo
- Codes were reviewed and refined, and grouped into candidate themes and subthemes.

3) *Theme refinement and validation:*

- Themes were reviewed against the coded extracts and the full dataset to ensure coherence and distinctiveness.
- Where possible, analytic memos and peer debriefing were used to enhance rigor and reduce researcher bias.
- Final themes were named and described, with illustrative quotes selected to represent each theme.

4) *Reporting:*

- Findings were organized around the main themes, with each theme linked to the study objectives.
- Quotes were anonymized and attributed only by participant identifier (e.g., P1, P2).

G. Data Limitations

The study is limited to psychologists' perspectives and does not include data from patients, clinicians from other disciplines, or developers. Findings may be influenced by participants' level of familiarity with digital tools and games. The qualitative design does not permit generalization in a statistical sense, but it provides in-depth insights that can inform future development and validation studies of decision-based games for mental health assessment.

IV. Findings

Based on the interview results across the two phases of pre-production and post-production of the game studies, expert consultations with clinical psychologists, academic experts, and student representatives provide a strong foundation for the feasibility and user acceptance of digital games as mental health screening tools. Refer to Figure 2 for the findings diagram.

A. Pre-Production of Game Criteria

1) *Clinical Feasibility and Integration of Psychometric Tools:* The interviews confirmed that integrating validated clinical assessment tools into decision-based games is highly feasible, provided strict developmental guidelines are followed:

Assessment Tool Validity: Experts strongly supported the integration of standardized tools like the GAD-7 (for anxiety) and PHQ-9 (for depression) within gamified environments. However, clinical psychologists emphasized that to maintain psychometric validity, developers must use the exact original wording and response scales of these tools rather than altering them to fit the game.

Reliability Metrics: When designing custom interactive scenarios or Likert-scale scoring systems, psychology experts advise aiming for a Cronbach's alpha reliability score of at least 0.8.

https://doi.org/10.31436/ijpcc.v12i2.717

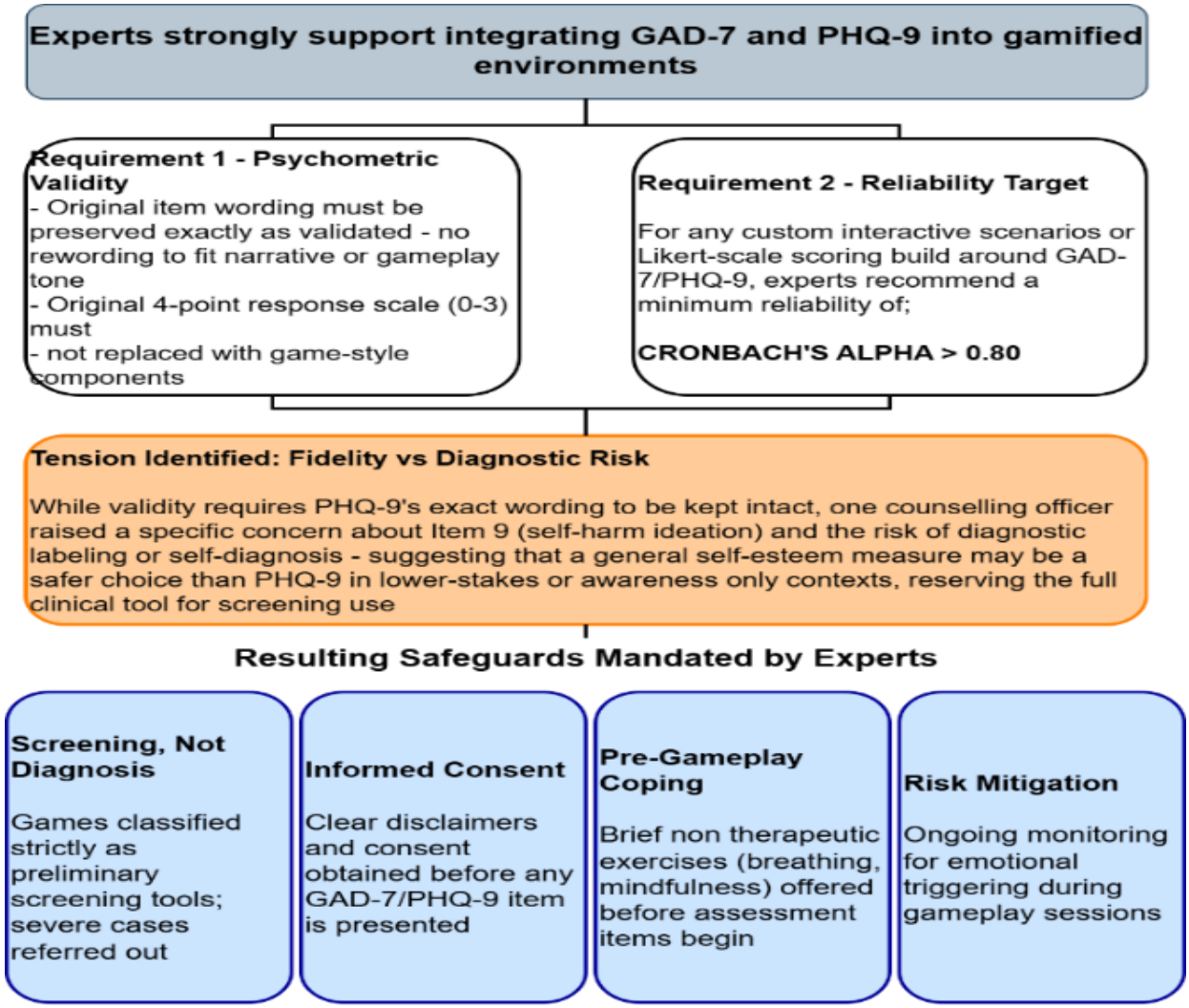


Fig. 2 Findings Diagram

2) Ethical Boundaries and Clinical Safety: The feasibility of these games relies heavily on maintaining strict ethical boundaries to ensure psychological safety:

Screening vs. Diagnosis: Experts unanimously agreed that decision-based games must be strictly classified as preliminary screening and awareness tools, not comprehensive diagnostic instruments. Extreme or severe cases detected by the game must be systematically referred to professional mental health services to avoid misdiagnosis.

Risk Mitigation: Clinical professionals highlighted the ethical risks of emotional triggering and self-diagnosis. In one instance, a counselling officer suggested that measuring general self-esteem might be safer than using clinical tools like the PHQ-9 to avoid diagnostic labelling. To mitigate risks, experts mandated the inclusion of clear disclaimers, informed consent, and non-therapeutic coping strategies, such as basic breathing or mindfulness exercises, prior to gameplay.

B. Post-production of game test

1) User Acceptance and Target Demographics: Interview data revealed strong user acceptance, particularly when the game is tailored to specific populations:

Focus on University Students: Experts advised exclusively targeting tertiary or university students rather than younger minors. This recommendation was based on both the distinct developmental behaviours of young adults and the legal/ethical restrictions regarding consent for minors.

Preference Over Traditional Methods: Student interviewees expressed a clear preference for digital gaming platforms over traditional pen-and-paper clinical questionnaires, noting that games are "unique, convenient,

https://doi.org/10.31436/ijpcc.v12i2.717

and more enjoyable," which helps bypass the stigma typically associated with mental health screening.

2) Design Mechanics for Enhanced Acceptance: Experts provided specific design requirements to maximize user acceptance and prevent the games from causing undue distress:

Anxiety-Sensitive Visuals: Multiple experts highlighted the necessity of using minimalist colour palettes to avoid sensory overstimulation, keeping players focused on the narrative without triggering anxiety symptoms.

Cultural and Narrative Resonance: To boost user acceptability, the games must embed local cultural contexts (such as Malaysian university scenarios) and familiar academic stressors. The tone of the game should prioritize positive reinforcement and a therapeutic feel, actively avoiding a patronizing or overly "clinical" atmosphere.

Pacing and Authenticity: To sustain engagement without causing fatigue, experts recommended short gameplay durations of 5 to 10 minutes. Furthermore, the use of professional human voice actors over AI-generated voices was recommended to provide authentic emotional resonance and increase player immersion.

V. Conclusion

This study demonstrates that decision-based games are a highly feasible and widely accepted alternative to traditional mental health assessments for detecting anxiety and depression, particularly among university students. By embedding validated psychometric tools — such as the PHQ-9 and GAD-7 — into immersive narratives, these gamified platforms effectively capture authentic behavioural responses while significantly reducing the social desirability bias, stigma, and intimidation often associated with conventional clinical questionnaires.

The findings emphasize that the viability of these tools depends heavily on specific design and ethical parameters:

1) Ethical and Clinical Boundaries: It is paramount that these games are strictly framed as preliminary screening and awareness tools rather than definitive diagnostic instruments. To ensure psychological safety, developers must integrate robust risk mitigation strategies, including clear disclaimers, informed consent, coping mechanisms, and direct referral pathways for users exhibiting severe symptoms.
2) Methodological Rigor: To maintain clinical validity, the exact wording and response scales of established psychometric tools must be preserved within the gameplay, and custom metrics should aim for high reliability (e.g., a Cronbach's alpha of at least 0.8).
3) Empathetic and Contextual Design: User acceptance is maximized when games incorporate culturally resonant scenarios, short pacing (5 to 10 minutes), professional voice acting, and anxiety-sensitive minimalist visuals that prevent sensory overstimulation.

Ultimately, when developed with a balance of scientific rigor and user-centric design, decision-based games offer a practically viable, scalable, and engaging solution for routine mental health screening. They present a unique opportunity to bridge the gap between vulnerable young adults and mental health services, promoting earlier detection and timely intervention in a naturalistic, non-threatening environment.

VI. Future Study

Future research should prioritize obtaining robust clinical evidence to validate decision-based games as reliable assessment tools. Because current approaches are under-validated in clinically representative samples, comprehensive appraisals of their screening and diagnostic accuracy are required before they can be deployed widely. The current study primarily reflects the perspectives of psychologists and small samples of student volunteers.

Future studies should incorporate data from a wider array of stakeholders, including patients, clinicians from other disciplines, and game developers. This will help account for varying levels of familiarity with digital tools and provide a more holistic view of the games' practical implementation. The qualitative design of the current study provides in-depth insights but does not permit statistical generalization. Therefore, future research should employ quantitative or mixed-methods approaches on larger, clinically representative populations to properly measure and generalize the feasibility, usability, and assessment accuracy of these games.

Mental health professionals still harbor concerns regarding the validity, ethics, and appropriate use of digital games. Future studies should specifically address these barriers to determine how these tools can be securely, ethically, and practically integrated into routine clinical workflows and primary care settings. While a large portion of existing literature evaluates games as therapeutic interventions, there is a notable gap regarding their use for diagnostics. Future research must continue to specifically examine them as assessment instruments that infer psychological states (like anxiety and depression severity) directly from a player's behavioural choices, rather than relying on self-reporting.

Acknowledgment

The authors acknowledge that the decision-based game platform (D3GAD) referenced in this study was developed

with support from the Sultan Ahmad Shah Medical Centre (SASMEC) grant.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORS CONTRIBUTION STATEMENT

S. Samsuri led and supervised the study, and was responsible for organizing and writing the compilation of findings from the anxiety and depression data collections. Z.M. Khandaker served as corresponding author and managed manuscript formatting and submission, and also contributed to the D3GAD component of the study. S. Rosli contributed to the development of the D3GAD game and user requirement data collection. M.W. Mohd Ghazali and A.L. Ruhaimi contributed to the adaptation and gamification of the PHQ-9 assessment instrument and to user requirement data collection. All authors read and approved the final version of the manuscript.

DATA AVAILABILITY STATEMENT

The interview transcripts and audio recordings analysed in this study are not publicly available due to participant confidentiality and the sensitive nature of the data. Anonymized data or thematic coding materials may be obtained from the corresponding author upon reasonable request, subject to participant confidentiality agreements.

ETHICS STATEMENT

Participants in this study were provided with an information sheet outlining the study's aims and procedures, and informed consent was obtained prior to participation. Participants were informed that participation was voluntary and that they could withdraw at any time without consequence. Audio recordings and transcripts were anonymized and stored on encrypted devices, accessible only to the research team, to maintain participant confidentiality.

REFERENCES

[1] K. Aneni, I. Gomati de la Vega, M. G. Jiao, M. C. Funaro, and L. E. Fiellin, "Evaluating the validity of game-based assessments measuring cognitive function among children and adolescents: A systematic review and meta-analysis," Progress in Brain Research, pp. 1–36, 2023, doi: 10.1016/bs.pbr.2023.02.002.

[2] F. Leutner, S.-C. Codreanu, S. Brink, and T. Bitsakis, "Game-based assessments of cognitive ability in recruitment: Validity, fairness and test-taking experience," Frontiers in Psychology, vol. 13, 2023, doi: 10.3389/fpsyg.2022.942662.

[3] C. Townsend, C. Humpston, J. Rogers, V. Goodyear, A. Lavis, and M. Michail, "The effectiveness of gaming interventions for depression and anxiety in young people: Systematic review and meta-analysis," BJPsych Open, vol. 8, no. 1, Art. no. e25, 2022, doi: 10.1192/bjo.2021.1078.

[4] O. A. David and C. Tomoiagă, "Game-based assessment of emotion regulation abilities in college students: A usability study of the REThink Life Game Assessment System," International Journal of Human–Computer Interaction, pp. 1–8, 2025, doi: 10.1080/10447318.2025.2462749.

[5] A. J. Williams et al., "Feasibility, perceived impact, and acceptability of a socially assistive robot to support emotion regulation with highly anxious university students: Mixed methods open trial," JMIR Mental Health, vol. 10, Art. no. e46826, 2023, doi: 10.2196/46826.

[6] S. Weller, P. A. Schroeder, and C. Plewnia, "Gamification improves antidepressant effects of cognitive control training—A pilot trial," Frontiers in Digital Health, vol. 4, 2022, doi: 10.3389/fdgth.2022.994484.

[7] E. Usta and M. Inozu, "The use of serious games in psychological interventions for anxiety disorders and related psychopathologies: A systematic review," Current Psychology, vol. 43, pp. 12610–12633, 2024, doi: 10.1007/s12144-023-05318-1.

[8] B. Yin and Y. Shen, "Compensatory beliefs in the internet gratification behavior: A study of game-based assessment," Frontiers in Public Health, vol. 11, 2023, doi: 10.3389/fpubh.2023.997108.

[9] V. Chitale, N. Baghaei, D. Playne, H. N. Liang, Y. Zhao, A. Erensoy, and Y. Ahmad, "The use of videogames and virtual reality for the assessment of anxiety and depression: A scoping review," Games for Health Journal, advance online publication, 2022, doi: 10.1089/g4h.2021.0227.

[10] N. A. Martin-Key, B. Spadaro, E. Funnell, E. J. Barker, T. S. Schei, J. Tomasik, and S. Bahn, "The current state and validity of digital assessment tools for psychiatry: Systematic review," JMIR Mental Health, vol. 9, no. 3, Art. no. e32824, 2022, doi: 10.2196/32824.

[11] M. Ruiz et al., "Winning the game against depression: A systematic review of video games for the treatment of depressive disorders," Current Psychiatry Reports, vol. 24, pp. 23–35, 2022, doi: 10.1007/s11920-022-01314-7.

[12] T. Tong, M. Chignell, M. C. Tierney, and J. Lee, "A serious game for clinical assessment of cognitive status: Validation study," JMIR Serious Games, vol. 4, no. 1, Art. no. e7, 2016, doi: 10.2196/games.5006.

[13] G. Spies, J. Huss, O. Motswage, S. Seedat, and C. Eichenberg, "Attitudes of patients and mental health professionals towards hypothetical use of serious games in psychotherapy," Health Informatics Journal, vol. 30, no. 2, Art. no. 14604582241259343, 2024, doi: 10.1177/14604582241259343.

[14] L. Lukka, V. M. Karhulahti, and J. M. Palva, "Mental health professionals' views on gaming to inform game-based interventions: Qualitative cross-sectional study," JMIR Serious Games, vol. 14, no. 1, Art. no. e69236, 2026.

[15] R. Tark, M. Metelitsa, K. Akkermann, K. Saks, S. Mikkel, and K. Haljas, "Usability, acceptability, feasibility, and effectiveness of a gamified mobile health intervention (Triumf) for pediatric patients: Qualitative study," JMIR Serious Games, vol. 7, no. 3, Art. no. e13776, 2019, doi: 10.2196/13776.

[16] C. Eichenberg, G. Grabmayer, and N. Green, "Acceptance of serious games in psychotherapy: An inquiry into the stance of therapists and patients," Telemedicine and e-Health, vol. 22, no. 11, pp. 945–951, 2016, doi: 10.1089/tmj.2016.0001.

[17] T. C. Gordon, A. H. Kemp, and D. J. Edwards, "Mixed-methods feasibility outcomes for a novel ACT-based video game 'ACTing Minds' to support mental health," BMJ Open, vol. 14, no. 3, Art. no. e080972, 2024, doi: 10.1136/bmjopen-2023-080972.

[18] University of Southern California, "Organizing your research paper methodology," USC Libraries, Jul. 27, 2021. [Online]. Available: https://libguides.usc.edu/writingguide/methodology

[19] R. Garg, "Methodology for research I," Indian Journal of Anaesthesia, vol. 60, no. 9, pp. 640–645, 2016, doi: 10.4103/0019-5049.190619.

International Journal on Perceptive and Cognitive Computing (IJPCC) Vol 12, Issue 2 (2026)
https://doi.org/10.31436/ijpcc.v12i2.673

Optimising L-Shaped Courtyard Building Design Using Prescriptive Analytics

Alea Syaffa Binti Mohd Sabri, Raini Binti Hassan*

Kulliyyah of Information and Communication Technology, International Islamic University Malaysia, Gombak, Malaysia

*Corresponding author: hrai@iium.edu.my
(Received: 17th March 2026; Accepted: 23rd June, 2026; Published on-line: 30th July, 2026)

Abstract— Buildings in tropical climates are predominantly cooling-dominated, making architectural design decisions an important factor in reducing operational energy demand. While machine learning has been widely applied to predict building energy performance, most existing studies focus primarily on predictive accuracy and provide limited guidance on translating predictions into practical design decisions. Consequently, architects and designers often know the expected energy performance of a design but remain uncertain about which configuration should be selected to achieve specific energy objectives. This study proposes a prescriptive analytics framework that extends machine learning beyond prediction to support architectural decision-making for L-shaped courtyard buildings in Malaysia's tropical climate. A Gradient Boosting (GB) model was identified as the best-performing predictive model, achieving R^2 values of 0.999684, 0.998808, and 0.999695 for cooling load, heating load, and total load prediction, respectively. The trained model was subsequently integrated into a prescriptive optimisation engine to identify energy-efficient architectural configurations. The optimal cooling- and total-load designs converged towards compact building forms with a window-to-wall ratio (WWR) of 0.10, producing predicted cooling and total energy loads of 43.81 kWh/m² and 45.55 kWh/m², respectively. In contrast, the heating-load optimum was achieved with a WWR of 0.40, resulting in a predicted heating load of 0.68 kWh/m². A scenario-based "what-if" analysis further confirmed that the cooling-oriented design package consistently achieved the lowest overall energy demand. The results identify the window-to-wall ratio as the most influential architectural parameter affecting building energy performance and demonstrate that cooling and heating objectives cannot be optimised simultaneously. The proposed framework transforms machine learning from a predictive model into a practical decision-support tool that provides interpretable and actionable design guidance for architects during the early stages of building design.

Keywords— Prescriptive Analytics, Machine Learning, Building Energy Prediction, Gradient Boosting, Courtyard Buildings

I. Introduction

Buildings account for almost 40% of global energy demand, making the built environment one of the largest contributors to energy consumption and greenhouse gas emissions worldwide [1]. In Malaysia, the building sector consumes nearly half of the national electricity supply, with energy demand continuing to increase because of rapid urbanisation and growing cooling requirements under the country's hot-humid tropical climate [17][14]. These challenges highlight the importance of improving building energy efficiency through architectural design decisions made during the early stages of project development. Passive design strategies, particularly those related to building geometry and envelope characteristics, offer significant opportunities to reduce cooling demand before mechanical systems are considered [1] [15].

Among passive design strategies, courtyard buildings have long been recognised for their ability to improve natural ventilation, daylight distribution, and thermal performance. In Malaysia, courtyard configurations have evolved from traditional shophouses to contemporary educational, residential, and commercial developments, demonstrating their continued relevance in climate-responsive architecture [11] [23]. Previous studies have shown that courtyard geometry, orientation, and envelope characteristics influence solar exposure, airflow, and building energy performance [15] [10] [2]. However, evaluating these design alternatives using conventional simulation tools remains computationally expensive and time-consuming, limiting their practicality during conceptual design where multiple design options must be explored efficiently [6].

Machine learning has emerged as an effective surrogate modelling approach capable of predicting building energy performance from architectural design variables with substantially lower computational cost than conventional simulation methods [20][21]. Algorithms such as Random Forest (RF), Gradient Boosting (GB), and Artificial Neural Networks (ANN) have demonstrated strong predictive capabilities in estimating cooling load, heating load, and

total energy consumption based on building geometry, orientation, and façade characteristics [6] [7] [18]. Nevertheless, most existing studies focus primarily on improving predictive accuracy, while providing limited support for practical architectural decision-making. Although designers can estimate the expected performance of a building, predictive models alone do not indicate which design configuration should be selected to achieve specific energy objectives [8].

This limitation highlights an important gap between predictive analytics and design practice. Predictive analytics answers the question of what will happen, whereas architects require guidance on what should be done to achieve better building performance. Prescriptive analytics addresses this limitation by combining predictive models with optimisation techniques capable of recommending design solutions based on predefined objectives [22] [13]. Despite its potential, the application of prescriptive analytics for courtyard building design remains limited, particularly within Malaysia's tropical context where cooling demand dominates operational energy consumption.

This study proposes a prescriptive analytics framework that extends machine learning from prediction to architectural recommendation. Building upon the best-performing Gradient Boosting model, the framework integrates an optimisation engine capable of evaluating thousands of potential L-shaped courtyard configurations to identify designs that minimise cooling load, heating load, and total energy load. In addition, a scenario-based "what-if" analysis is introduced through three representative design packages to provide architects with interpretable and practical design alternatives. Rather than simply predicting energy performance, the proposed framework supports evidence-based decision-making by translating machine learning outputs into actionable design recommendations for early-stage architectural design [3].

II. LITERATURE REVIEW

A. *Machine Learning for Building Energy Prediction*

Building energy prediction has traditionally relied on analytical and simulation-based methods, such as the Cooling Load Temperature Difference (CLTD), Radiant Time Series Method (RTSM), Total Equivalent Temperature Difference (TETD), and dynamic building simulation software. Although these methods provide reliable estimates of cooling and heating loads, they require detailed building information, extensive computational resources, and considerable modelling effort, making them less suitable for rapid evaluation during the conceptual design stage [5]. Furthermore, traditional approaches operate as deterministic models that do not learn from data, cannot capture complex nonlinear relationships, and are unable to generate prescriptive recommendations for design optimisation.

Recent advances in machine learning have provided an alternative approach for predicting building energy performance by modelling the nonlinear relationship between architectural variables and energy loads. Ensemble learning algorithms, particularly Random Forest (RF) and Gradient Boosting (GB), together with Artificial Neural Networks (ANN), have demonstrated high predictive accuracy in estimating cooling load, heating load, and total energy consumption based on geometric and envelope characteristics [6][7][18]. These surrogate models significantly reduce computational time while maintaining prediction accuracy, enabling designers to evaluate multiple design alternatives more efficiently than conventional simulation-based workflow [21] [20].

Despite these advances, most existing machine learning studies remain focused on predictive performance. Model evaluation is typically based on statistical indicators such as Root Mean Square Error (RMSE) and the coefficient of determination (R^2), with limited attention given to how prediction results can support architectural decision-making during the early design stage [7]. Consequently, machine learning is often used to answer what will happen rather than what designers should do.

B. *Prescriptive Analytics for Building Design Optimisation*

Prescriptive analytics extends predictive modelling by combining machine learning with optimisation techniques to recommend the most suitable decisions for achieving predefined objectives. Recent building energy studies have begun integrating machine learning surrogate models with optimisation and scenario evaluation to support early-stage design decision-making [13] [22]. Rather than predicting energy performance alone, prescriptive analytics identifies architectural configurations that minimise energy demand while satisfying specified design constraints.

Recent studies also emphasise that optimisation should acknowledge trade-offs between competing design objectives rather than seeking a single universal optimum. Building energy optimisation frequently requires balancing cooling demand, heating demand, thermal comfort, daylight availability, and other performance criteria, resulting in objective-specific solutions rather than one optimal design [9] [16]. This perspective is particularly relevant in tropical climates, where cooling demand dominates annual energy consumption and envelope characteristics, especially the window-to-wall ratio, play a decisive role in determining operational energy use. Consequently, prescriptive recommendations should be climate-specific and aligned with the primary energy objective of the building.

However, despite the growing application of machine learning in building energy prediction, limited research has

integrated predictive models with prescriptive optimisation for courtyard buildings. Existing studies rarely combine multi-output machine learning, optimisation search, and scenario-based what-if analysis within a unified framework capable of producing actionable architectural recommendations, particularly in Malaysia's tropical climate [5].

Therefore, this study addresses this research gap by integrating Gradient Boosting with a prescriptive optimisation framework to identify energy-efficient L-shaped courtyard building configurations. Unlike conventional predictive models that estimate energy performance, the proposed framework systematically evaluates multiple architectural configurations and translates prediction results into practical design recommendations through optimisation and scenario-based analysis, thereby supporting evidence-based architectural decision-making.

C. *Architectural Variables Affecting Building Energy Performance*

Previous studies consistently demonstrate that architectural and envelope characteristics significantly influence building energy performance. Among these variables, the window-to-wall ratio (WWR) has been identified as one of the most influential parameters affecting cooling demand in tropical climates because larger glazed areas increase solar heat gain and consequently cooling energy consumption [12]. Similarly, building geometry, including form factor and surface-to-volume ratio, influences heat transfer between the building envelope and the external environment, affecting both cooling and heating loads [13].

Courtyard geometry further contributes to building energy performance by influencing solar exposure, shading, and natural ventilation. Previous studies have reported that courtyard dimensions, orientation, and geometric proportions affect thermal behaviour and energy demand, particularly in hot-humid climates [15] [10] [2]. These findings indicate that optimising architectural variables rather than relying solely on mechanical systems can substantially improve building energy efficiency during the early stages of design.

D. *Optimisation Techniques for Building Energy Design*

Building energy optimisation has gained increasing attention as researchers seek to identify architectural configurations that minimise energy consumption while satisfying multiple design objectives. Conventional optimisation approaches are commonly integrated with building energy simulation to evaluate alternative design solutions. However, simulation-based optimisation is computationally intensive because each design iteration requires repeated energy simulations, limiting its practicality during the conceptual design stage when numerous design alternatives must be explored [21] [20].

The emergence of machine learning surrogate models has significantly improved optimisation efficiency by replacing repeated simulations with rapid prediction models. Once trained, surrogate models can estimate building energy performance almost instantaneously, allowing optimisation frameworks to evaluate a substantially larger number of design alternatives within a shorter computational time. This capability enables designers to explore the influence of architectural variables on energy performance more effectively during the early stages of design [21] [20].

Despite these developments, most optimisation studies *focus on identifying optimal numerical solutions and rarely translate the optimisation outcomes into interpretable recommendations for architectural practice [19]. Designers often receive optimal parameter values without understanding how different design strategies perform under varying objectives. Consequently, optimisation results remain difficult to apply in practical decision-making. This limitation highlights the need for prescriptive frameworks that integrate optimisation with scenario-based analysis to provide architects with actionable and interpretable design recommendations.*

E. *Research Gap*

Although machine learning has demonstrated strong capability in predicting building energy performance, most existing studies focus primarily on improving predictive accuracy rather than supporting architectural decision-making [18][6]. Furthermore, relatively few studies have specifically investigated courtyard geometry despite evidence that building form, orientation, surface exposure, and window-to-wall ratio significantly influence thermal performance [15] [10] [2].

More importantly, predictive modelling alone does not provide actionable design recommendations. The integration of prescriptive analytics with machine learning to optimise courtyard building configurations remains largely absent, particularly within the Malaysian tropical context [22] [13]. Consequently, architects are able to estimate building energy performance but still lack practical guidance on selecting the most energy-efficient design configuration during the early stages of design. This study addresses this gap by integrating Gradient Boosting with a prescriptive optimisation framework and scenario-based what-if analysis to transform energy prediction into practical design recommendations for L-shaped courtyard buildings.

https://doi.org/10.31436/ijpcc.v12i2.673

III. METHODOLOGY

This study proposes a prescriptive analytics framework that extends machine learning beyond prediction by transforming predicted building energy performance into practical architectural recommendations. The framework follows the Cross-Industry Standard Process for Data Mining (CRISP-DM), consisting of business understanding, data understanding, data preparation, modelling, evaluation, and prescriptive analytics. While the predictive modelling stage identifies the most accurate machine learning model, the prescriptive stage applies optimisation and scenario analysis to recommend energy-efficient courtyard building configurations.

A. Dataset and Predictive Model

The study utilised the L-shaped courtyard building dataset developed by [4], which contains validated thermal performance and energy simulation data representing buildings located in Kuala Lumpur, Malaysia. The dataset includes seven architectural design variables, namely WIDTH, LENGTH, HEIGHT, ORIENTATION, WINDOWS_RATIO, FORM_FACTOR, and S_V_RATIO, while the response variables consist of COOLING_LOAD, HEATING_LOAD, and TOTAL_LOAD. These architectural parameters represent key geometric and envelope characteristics that influence building energy performance under Malaysia's tropical climate.

Three machine learning algorithms, namely Random Forest (RF), Gradient Boosting (GB), and Artificial Neural Network (ANN), were evaluated for predicting cooling load, heating load, and total energy load. Model performance was assessed using statistical error measures, and Gradient Boosting consistently demonstrated the most stable and accurate predictive performance across both single-output and multi-output prediction tasks. Consequently, the trained Gradient Boosting model was selected as the surrogate model for the subsequent prescriptive analytics stage.

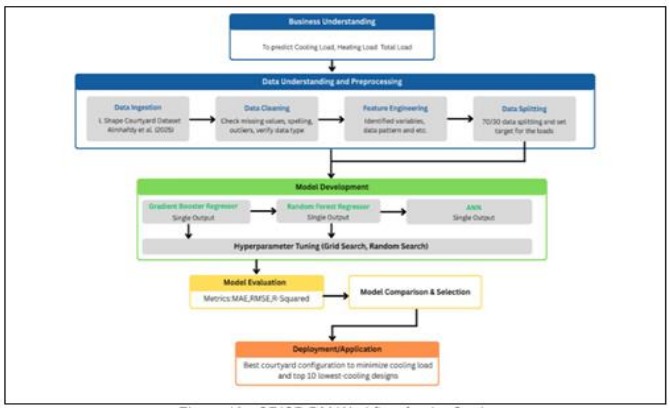


Fig 1. Machine Learning Model Development Workflow for L-Shaped Courtyard Energy Prediction.

B. Prescriptive Optimisation Framework

Unlike conventional machine learning studies that terminate after prediction, this research integrates the Gradient Boosting model into a prescriptive optimisation framework to identify energy-efficient architectural configurations. The predictive model serves as a surrogate model capable of rapidly estimating building energy performance without repeatedly performing computationally intensive building simulations.

The optimisation process began by generating thousands of possible combinations of the architectural design variables within the ranges defined by the dataset. Each generated design alternative was evaluated using the trained Gradient Boosting model to predict its corresponding cooling load, heating load, and total energy load. The predicted results were subsequently ranked according to three optimisation objectives:

- minimisation of cooling load,
- minimisation of heating load, and
- minimisation of total energy load.

For each objective, the design configuration with the lowest predicted energy demand was selected as the recommended solution. This optimisation process transforms predictive outputs into actionable architectural recommendations by identifying the most suitable design parameters for different performance objectives rather than merely estimating energy consumption.

C. Scenario-Based What-If Analysis

To improve the practical applicability of the proposed framework, a scenario-based what-if analysis was conducted following the optimisation stage. Rather than evaluating individual design variables independently, three representative architectural design packages were developed to simulate realistic design strategies commonly encountered during the early stages of building design. These packages enable designers to compare the expected energy performance of different architectural approaches before finalising a building configuration.

The three design scenarios consist of:

- Package A: Cooling-optimised design characterised by a compact building form with a low window-to-wall ratio.
- Package B: Conventional design representing moderate values across the principal architectural variables.
- Package C: A design with relatively higher glazing ratio and a vertically compact configuration.

Each package was evaluated using the trained Gradient Boosting model, allowing the predicted cooling load, heating load, and total energy load to be compared across alternative design strategies. This scenario-based evaluation

https://doi.org/10.31436/ijpcc.v12i2.673

provides architects with interpretable design guidance by illustrating the consequences of selecting different architectural configurations instead of relying solely on individual optimisation results.

IV. RESULTS AND DISCUSSION

A. *Gradient Boosting Performance*

Among the three machine learning models evaluated, Gradient Boosting achieved the highest predictive accuracy across all target variables and was therefore selected as the surrogate model for prescriptive optimisation. In the multi-output framework, Gradient Boosting achieved an RMSE of 0.01967 and R^2 = 0.999684 for Cooling Load, RMSE = 0.01012 and R^2 = 0.998808 for Heating Load, and RMSE = 0.01434 and R^2 = 0.999695 for Total Load. These values consistently outperformed Random Forest and Artificial Neural Network models, demonstrating the robustness of Gradient Boosting for subsequent optimisation tasks.

Since the objective of this study is to provide design recommendations rather than improve prediction accuracy alone, the trained Gradient Boosting model was subsequently employed as the optimisation engine for prescriptive analytics.

B. *Optimal Design Configuration*

The optimisation process evaluated multiple feasible L-shaped courtyard configurations to identify the architectural design that minimised each energy objective independently. The results indicate that different optimisation objectives produce different architectural solutions, confirming that no single design simultaneously minimises cooling, heating, and total energy demand.

TABLE 1 –
OPTIMAL DESIGN CONFIGURATIONS IDENTIFIED BY THE PRESCRIPTIVE OPTIMISATION FRAMEWORK

Parameter	Cooling Load	Heating Load	Total Load
Width (m)	60	60	60
Length (m)	60	30	60
Height (m)	8	16	8
Orientation	45°	0°	45°
Window-to-Wall Ratio	0.10	0.40	0.10
Form Factor	1.8007	1.3007	1.8007
S/V Ratio	0.4502	0.3476	0.4502
Predicted Load (kWh/m²)	43.81	0.68	45.55

For cooling load minimisation, the optimisation converged towards a 60 m × 60 m building footprint with a building height of 8 m, an orientation of 45°, and a window-to-wall ratio (WWR) of 0.10, producing a predicted cooling load of 43.81 kWh/m². Interestingly, the optimisation for total energy load converged to an almost identical configuration, yielding a predicted total load of 45.55 kWh/m². This finding indicates that, under Malaysia's hot-humid climate, cooling demand dominates total building energy consumption. Consequently, architectural strategies that minimise cooling load also minimise total energy demand.

Conversely, the optimal configuration for heating load differed substantially. The optimisation recommended a 60 m × 30 m footprint, 16 m building height, 0° orientation, and a WWR of 0.40, producing the lowest predicted heating load of 0.68 kWh/m². Compared with the cooling-optimised configuration, this design exhibits a lower form factor and lower surface-to-volume ratio, reflecting a more vertically compact geometry with greater façade transparency.

These results clearly demonstrate that optimisation objectives are conflicting. While cooling and total energy optimisation favour compact horizontal buildings with minimal glazing, heating optimisation benefits from taller buildings with larger glazing areas.

C. *Scenario-Based What-If Analysis*

To improve practical applicability, the optimisation results were translated into three representative architectural design packages.

TABLE 2
SCENARIO-BASED DESIGN PACKAGES

Package	Cooling (kWh/m²)	Heating (kWh/m²)	Total (kWh/m²)
Package A	43.847	1.727	45.553
Package B	45.651	1.194	46.845
Package C	47.279	0.681	47.964

Package A, representing the cooling-oriented design, achieved the lowest cooling load (43.847 kWh/m²) and total energy load (45.553 kWh/m²). This package adopted a low-rise configuration (8 m), a 60 m × 60 m footprint, a diagonal orientation (45°), and the minimum WWR of 0.10. Despite exhibiting the highest form factor (1.80) and surface-to-volume ratio (0.45), it consistently outperformed the remaining scenarios.

Package B represents conventional Malaysian design practice. Compared with Package A, cooling load increased from 43.847 to 45.651 kWh/m², while total load increased to 46.845 kWh/m², indicating that moderate glazing and increased building height reduce overall energy efficiency.

Package C produced the highest cooling (47.279 kWh/m²) and total energy loads (47.964 kWh/m²), despite achieving the most compact geometry. Although heating demand decreased to 0.681 kWh/m², the increase in cooling demand

outweighed this benefit. Relative to Package A, cooling load increased by 3.432 kWh/m² (7.8%), confirming that increasing glazing ratio substantially increases operational energy demand under tropical climatic conditions.

V. CONCLUSIONS

This study proposed a prescriptive analytics framework that extends machine learning beyond prediction to support architectural decision-making for L-shaped courtyard buildings in Malaysia's tropical climate. While previous studies have primarily focused on predicting building energy performance, the proposed framework demonstrates how predictive models can be transformed into practical decision-support tools through optimization and scenario-based what-if analysis. Gradient Boosting was identified as the best-performing predictive model and subsequently employed as the optimization engine to recommend energy-efficient courtyard building configurations. By integrating prediction, optimization, and scenario analysis within a single framework, this study addresses the gap between predictive modelling and practical architectural decision-making [22][13].

The optimization results demonstrate that compact building forms with a low window-to-wall ratio (WWR = 0.10) consistently minimize cooling load (43.81 kWh/m²) and total energy load (45.55 kWh/m²), whereas heating load is minimized by a taller configuration with a higher window-to-wall ratio (WWR = 0.40), producing a predicted heating load of 0.68 kWh/m². Furthermore, the scenario-based what-if analysis confirmed that the cooling-oriented design package achieved the lowest overall energy demand, while the high-glazing package produced the highest cooling and total energy consumption. These findings demonstrate that cooling and heating objectives cannot be optimized simultaneously, highlighting the importance of objective-specific design recommendations for architects during the conceptual design stage.

The study further confirms that the window-to-wall ratio is the dominant architectural parameter governing cooling and total energy demand in tropical courtyard buildings. By transforming predictive machine learning into a prescriptive optimization framework, the proposed approach provides architects with practical guidance for selecting energy-efficient courtyard configurations instead of relying solely on predicted energy values. This contribution demonstrates how machine learning can evolve from a predictive technique into an evidence-based decision-support tool for sustainable building design.

Despite these contributions, several limitations should be acknowledged. The proposed framework was developed using a parametric simulation dataset for L-shaped courtyard buildings and therefore does not capture real-world operational variability, occupant behavior, shading devices, natural ventilation strategies, material thermal properties, or active building system performance. In addition, the optimization recommendations were generated using surrogate model predictions rather than direct field validation. These limitations may affect the generalizability of the findings to other building typologies and climatic conditions.

Future research should validate the recommended design configurations using real-world building case studies and measured operational data to further verify the effectiveness of the proposed framework. The framework may also be extended to other courtyard typologies and different climatic regions to improve its broader applicability. In addition, incorporating additional architectural variables such as shading devices, natural ventilation strategies, material thermal properties, and active building system efficiencies would enable a more comprehensive evaluation of building energy performance. Future studies may also investigate advanced neural network architectures and enhanced optimization techniques to further improve predictive performance and strengthen prescriptive decision support for sustainable building design.

ACKNOWLEDGMENT

The author would like to express sincere appreciation to Asst. Prof. Dr. Raini Hassan, Kulliyyah of Information and Communication Technology (KICT), International Islamic University Malaysia (IIUM), for her invaluable guidance, constructive feedback, and continuous encouragement throughout this research. The author also gratefully acknowledges International Islamic University Malaysia (IIUM) for the recognition and support through the IIUM Sejahtera Award, which has provided motivation and encouragement in completing this study. The author extends heartfelt gratitude to her family for their unwavering love, patience, prayers, and continuous support throughout this research journey. Sincere appreciation is also extended to friends and colleagues for their encouragement, thoughtful discussions, and moral support, which have contributed significantly to the successful completion of this work.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHOR(S) CONTRIBUTION STATEMENT

All authors contributed equally to this work.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ETHICS STATEMENT

This study did not require ethical approval.

REFERENCES

[1] S. R. S. Aldhshan, K. N. A. Maulud, W. S. W. M. Jaafar, O. A. Karim, and B. Pradhan, "Energy consumption and spatial assessment of renewable energy penetration and building energy efficiency in Malaysia: A review," Sustainability, vol. 13, no. 16, p. 9244, 2021, doi: 10.3390/su13169244.

[2] A. Al-Fayyad, S. Al-Mamoori, and A. Al-Shimmary, "The impact of courtyard geometry on energy efficiency in hot-dry climates: A parametric study," Journal of Engineering, vol. 2023, p. 883421, 2023.

[3] A. Ali, X. Zhou, and T. Hong, "Machine learning as a surrogate to building performance simulation: A review," Energy and Buildings, vol. 285, p. 112848, 2023.

[4] A. Almhafdy, N. Ibrahim, S. S. Ahmad, E. Almahmoud, and A. Alashwal, "Optimization of U-enclosure courtyard (UEC) design factors in tropical climate: Energy assessment," Civil Engineering and Architecture, 2025.

[5] B. Al Shawa, "Assessing the validity of simplified heating and cooling demand calculation methods: The case of passive house planning package (PHPP) and radiant time series method (RTSM)," Frontiers in Built Environment, vol. 10, p. 1426774, 2024, doi: 10.3389/fbuil.2024.1426774.

[6] G. Bekdaş, A. Dutta, and A. Baki, "Prediction of cooling load of tropical buildings with machine learning," Sustainability, vol. 15, no. 6, p. 4583, 2023, doi: 10.3390/su15064583.

[7] V. R. Chaganti, K. Roy, and A. Kannan, "Building heating and cooling load prediction using ensemble machine learning model," Sensors, vol. 22, no. 19, p. 7692, 2022, doi: 10.3390/s22197692.

[8] Z. Chen, F. Xioa, F. Guo, and J. Yan, "Interpretable machine learning for building energy management: A state-of-the-art review," 2023.

[9] P. De Wilde, "Building performance simulation in the brave new world of artificial intelligence and digital twins: A systematic review," 2023.

[10] E. Diz-Mellado, C. Barreneche, and L. F. Cabeza, "Unravelling the impact of courtyard geometry on cooling energy consumption in buildings," Building and Environment, vol. 240, p. 110442, 2023, doi: 10.1016/j.buildenv.2023.110442.

[11] K. Gunasagaran, N. Ibrahim, and A. Bahauddin, "The significance of courtyards in Melaka terrace houses: A study of thermal comfort and cultural heritage," Journal of the Malaysian Institute of Planners, vol. 19, no. 2, pp. 123–136, 2021, doi: 10.21837/pm.v19i22.904.

[12] V. Gupta and C. Deb, "Envelope design for low-energy buildings in the tropics: A review," Renewable and Sustainable Energy Reviews, vol. 186, p. 113650, 2023.

[13] Z. He, J. Wang, and S. Liu, "Prescriptive analytics for energy-efficient building envelope optimization using tree-based genetic programming," 2022.

[14] IRENA, "Malaysia energy transition outlook," International Renewable Energy Agency, Abu Dhabi, 2023.

[15] A. M. Nugroho, A. Citraningrum, W. Iyati, and M. H. Ahmad, "Courtyard as tropical hot humid passive design strategy: Case study of Indonesian contemporary houses in Surabaya, Indonesia," 2020.

[16] M. Ouf, H. Gohary, and V. Ugursal, "Energy efficiency in sustainable buildings: A systematic review of current research and future directions," Sustainable Cities and Society, vol. 61, p. 102335, 2020.

[17] N. A. Rahim and A. H. M. Ali, "Comparative analysis of building energy performance in Malaysia: A multiple linear regression study on green vs. conventional buildings," Journal of Energy and Safety Technology (JEST), vol. 7, no. 2, pp. 54–82, 2024, doi: 10.11113/jest.v7.177.

[18] M. K. M. Shapi, N. A. Ramli, and L. J. Awalin, "Energy consumption prediction by using machine learning for smart building: Case study in Malaysia," 2021.

[19] S. Studer, T. B. Bui, C. Drescher, L. Winkler, S. Peters, K. Müller, and M. Matzner, "Towards CRISP-ML(Q): A machine learning process model with quality assurance methodology," Software and Systems Modeling, vol. 20, no. 4, pp. 1135–1155, 2021, doi: 10.1007/s10270-021-00879-3.

[20] X. Wang, J. Harrison, R. Teigland, and A. Hollberg, "Machine learning (ML) as a surrogate model for early-stage heating demand optimization," in Proc. Building Performance Analysis Conf. (SimBuild 2024), 2024.

[21] P. Westermann and R. Evins, "besos: Building and energy simulation, optimization and surrogate-modelling (version 3.0)," Journal of Open Source Software, vol. 6, no. 59, p. 2677, 2021.

[22] C. Yu, X. Li, and J. Zhou, "Predictive and prescriptive modeling for optimizing courtyard design under different climatic zones using LightGBM and genetic algorithms," Mathematics, vol. 11, no. 1142, pp. 1–20, 2023, doi: 10.3390/math111142.

[23] A. Zwain and A. Bahauddin, "The significance of the traditional courtyard components of shophouses in George Town, Penang Malaysia," Advances in Environmental Biology, vol. 9, no. 5, pp. 32–34, 2015.

International Journal on Perceptive and Cognitive Computing (IJPCC) Vol 12, Issue 2 (2026)
https://doi.org/10.31436/ijpcc.v12i2.713

Predicting Energy Load of L-Shaped Courtyard Buildings in Malaysia Using Machine Learning

Alea Syaffa Binti Mohd Sabri, Raini Binti Hassan*
Kulliyyah of Information and Communication Technology, International Islamic University Malaysia, Gombak, Malaysia

*Corresponding author: hrai@iium.edu.my
(Received: 17th March 2026; Accepted: 23rd June, 2026; Published on-line: 30th July, 2026)

Abstract— Building's account for a substantial share of global energy consumption, with cooling demand dominating in hot and humid tropical climates such as Malaysia. Although dynamic simulation tools such as EnergyPlus and TRNSYS provide accurate performance evaluation, their high computational cost and reliance on specialised expertise limit their application during early-stage design. This study develops a machine learning-based surrogate modelling framework to predict the cooling load, heating load, and total energy load of L-shaped courtyard buildings in Kuala Lumpur. A validated parametric simulation dataset was generated using key architectural and geometric variables, including WIDTH, LENGTH, HEIGHT, ORIENTATION, WINDOWS_RATIO, FORM_FACTOR, and S_V_RATIO. Three supervised regression models, Random Forest (RF), Gradient Boosting (GB), and Artificial Neural Network (ANN-MLP), were evaluated using Root Mean Square Error (RMSE) and Coefficient of Determination (R^2). The results show that ensemble learning models consistently outperform ANN, with Gradient Boosting achieving the highest predictive accuracy across all energy outputs. For cooling load prediction, GB achieved an RMSE of 0.0188 and an R^2 of 0.9997, while also producing the best performance for heating load and total energy load. The findings further indicate that envelope-related variables, particularly the window-to-wall ratio, have a significant influence on cooling and overall energy demand in tropical courtyard buildings. By integrating climate-specific parametric simulation with ensemble machine learning, the proposed framework accurately reproduces simulation-derived energy loads while reducing the need for repeated simulation during early-stage design. The developed surrogate model provides a reliable decision-support tool for rapid performance assessment during early-stage architectural design and contributes a validated predictive framework for energy-efficient tropical courtyard buildings in Malaysia.

Keywords— Building energy prediction, courtyard buildings, tropical climate, machine learning, Gradient Boosting

I. Introduction

The building sector remains one of the largest consumers of global energy, accounting for nearly 40% of total final energy demand and a significant share of greenhouse gas emissions [1], [10]. In tropical climates, such as Malaysia, building energy consumption is predominantly driven by cooling demand rather than heating, due to persistently high temperatures and humidity throughout the year [11], [18]. As urbanisation accelerates and indoor comfort expectations increase, electricity consumption in residential and commercial buildings continues to rise, placing additional pressure on national energy systems and carbon reduction targets [10], [26]. Improving building energy efficiency, particularly during the early design stage, has therefore become a critical strategy for sustainable development.

Architectural design decisions made during conceptual planning have long-term consequences for operational energy performance. Parameters such as building geometry, façade transparency, orientation, and compactness directly influence solar heat gain, surface heat transfer, and internal thermal behaviour [5], [18], [25]. In tropical contexts, envelope-related factors such as window-to-wall ratio (WWR) play a particularly significant role in determining cooling load [5]. Among passive design typologies courtyard architecture has historically demonstrated climatic responsiveness through enhanced natural ventilation, shading, and daylight distribution [8], [15], [16]. In Malaysia, L-shaped courtyard configurations are commonly adopted in educational institutions, commercial developments, and residential projects due to their spatial flexibility and environmental performance potential [9].

Despite the recognised climatic benefits of courtyard typologies, quantitative frameworks linking courtyard geometry to whole-building energy performance remain limited. Traditional approaches for evaluating building energy demand rely heavily on dynamic simulation tools such as EnergyPlus and TRNSYS [2], [4], [12]. While these tools provide detailed transient heat balance modelling, they require extensive input parameters, calibration, and computational time. Comparative studies have shown that

https://doi.org/10.31436/ijpcc.v12i2.713

simulation outputs may vary depending on modelling assumptions and software configurations [4], [13]. Consequently, architects often face practical constraints in testing multiple design alternatives during early-stage development, where rapid iteration is essential.

Simplified analytical methods such as the Cooling Load Temperature Difference (CLTD) method and Radiant Time Series Method (RTSM) offer faster approximations but rely on steady-state assumptions and tabulated correction factors, limiting their ability to capture nonlinear interactions between geometry and climatic variables [2], [14]. These deterministic approaches are not well suited for complex forms such as L-shaped courtyard buildings, where surface exposure, orientation, and enclosure degree interact dynamically.

Recent advances in machine learning (ML) provide an alternative paradigm for building energy prediction. Rather than explicitly solving thermodynamic equations, ML models learn nonlinear relationships directly from structured datasets [3], [6]. Ensemble learning techniques such as Random Forest and Gradient Boosting have demonstrated strong predictive performance in heating and cooling load estimation tasks [7], [20], [21]. Surrogate modelling frameworks, where ML models are trained on simulation-generated datasets, enable near-instantaneous prediction of new design configurations while maintaining high accuracy [6], [22]. This capability is particularly valuable for early-stage architectural design, where numerous parametric variations must be evaluated efficiently.

Although machine learning applications in building energy research have expanded globally, relatively few studies focus specifically on tropical courtyard geometries under Malaysian climatic conditions [9], [17]. Most existing ML-based energy studies utilise generic building archetypes or temperate climate datasets, limiting their contextual relevance to cooling-dominated environments [18], [23]. Furthermore, the integration of ensemble machine learning techniques with courtyard-specific parametric variables remains underexplored.

To address this gap, this study develops and evaluates a machine learning-based predictive framework for estimating cooling load, heating load, and total load of L-shaped courtyard buildings in Kuala Lumpur. Using a validated parametric simulation dataset, the study compares three supervised regression algorithms: Random Forest, Gradient Boosting, and Artificial Neural Network (MLP). Model performance is assessed using Root Mean Square Error (RMSE) and Coefficient of Determination (R^2), enabling rigorous comparison of predictive accuracy.

By integrating climate-specific architectural parameters with advanced ensemble learning techniques, this research contributes a high-accuracy surrogate modelling approach tailored to Malaysian tropical conditions. The findings aim to support performance-driven architectural decision-making, reduce reliance on repeated dynamic simulations, and promote energy-efficient courtyard design aligned with national sustainability objectives [1], [10], [26].

II. LITERATURE REVIEW

The building sector remains one of the largest contributors to global energy consumption, accounting for nearly forty percent of total final energy demand worldwide [1]. In tropical regions such as Malaysia, energy consumption is predominantly cooling-driven due to high temperature and humidity throughout the year. Consequently, reducing cooling load has become central to achieving sustainability and carbon reduction targets [1].

Architectural design parameters, particularly geometry and façade configuration, significantly influence cooling demand in hot–humid climates. However, the methodologies used to evaluate building energy performance vary in complexity, accuracy, and computational demand. The literature can broadly be categorised into three methodological streams: traditional analytical methods, dynamic simulation approaches, and machine learning-based predictive modelling.

Table 1 shows the traditional cooling and heating load estimation techniques were developed to simplify thermal load calculations during early engineering design. Among the most commonly applied analytical methods are the Cooling Load Temperature Difference (CLTD) method, Total Equivalent Temperature Difference (TETD), and the Radiant Time Series Method (RTSM) [2], [14]. These approaches rely on predefined correction factors and simplified heat balance assumptions derived from standard climatic conditions.

The CLTD method estimates peak cooling loads by adjusting temperature differences based on construction type and solar exposure. While straightforward and computationally efficient, it assumes steady-state conditions and limited dynamic interaction between building components [2]. TETD expands this by incorporating equivalent temperature adjustments to better represent solar gains, yet it remains manual and sensitive to tabulated values. RTSM improves temporal representation by separating radiant and convective heat components; however, it assumes periodic load patterns and simplified occupancy behaviour [14].

Although these techniques remain useful for quick approximations, their deterministic nature restricts their applicability to complex geometries such as courtyard buildings. They cannot effectively capture nonlinear interactions between geometric compactness, façade transparency, and climatic variability.

https://doi.org/10.31436/ijpcc.v12i2.713

TABLE I
TRADITIONAL METHODS USED FOR COOLING LOAD AND HEATING LOAD PREDICTION

Method	Core Principle	Strengths	Limitations	Key Source
CLTD	Tabulated temperature differences to estimate peak cooling	Simple and fast	Assumes steady-state; limited climate adaptability	[2]
TETD	Equivalent temperature difference incorporating solar gains	Improved solar correction	Manual and time-intensive	[2]
RTSM	Time-series heat balance simplification	Accounts for radiant and convective split	Assumes periodic conditions; limited annual accuracy	[2]

Dynamic simulation tools such as EnergyPlus, TRNSYS, IDA ICE, and Ecotect as shown in Table 2 represent the industry standard for detailed building energy modelling [4], [12], [13]. These tools solve transient heat balance equations using hourly weather data, material properties, occupancy schedules, and HVAC system parameters.

Dynamic simulation allows accurate modelling of solar radiation, shading devices, thermal mass effects, and ventilation behaviour, making it suitable for complex building forms such as courtyard typologies. However, modelling accuracy depends heavily on input assumptions and calibration processes. Comparative analyses have reported discrepancies in predicted energy loads between software platforms even when modelling identical geometries [4].

Moreover, simulation-based parametric exploration requires significant computational time. Testing multiple architectural variations during conceptual design can be impractical, particularly in professional practice where time constraints are strict.

TABLEII
THE INDUSTRY STANDARD FOR DETAILED BUILDING ENERGY MODELLING

Tool	Capability	Advantages	Limitations	Key Source
EnergyPlus	Detailed transient simulation	High accuracy; industry standard	Time-intensive; complex input setup	[4]
TRNSYS	Modular system simulation	Flexible HVAC modelling	High computational cost	[4]
IDA ICE	Whole-building simulation	Integrated airflow modelling	Software variability in results	[4]
Ecotect	Concept-stage environmental analysis	Faster modelling	Less detailed physics modelling	[4]

A. *Machine Learning in Building Energy Prediction*

Machine learning (ML) has gained attention as an alternative to physics-based modelling due to its ability to learn nonlinear relationships directly from data [3], [6]. Supervised regression algorithms are commonly used to predict heating and cooling loads based on geometric and envelope features.

Based on Table 3, Artificial Neural Networks (ANN) were among the earliest ML models applied in energy research. While capable of modelling complex nonlinear behaviour, ANNs require careful hyperparameter tuning and may overfit small datasets [3].

Ensemble learning methods such as Random Forest (RF) and Gradient Boosting (GB) have demonstrated superior predictive performance in structured tabular datasets typical of building energy modelling [7], [20], [21]. RF reduces variance through bagging, while GB improves accuracy by sequentially correcting residual errors. These methods are particularly effective in capturing interactions among façade transparency, surface exposure, and orientation.

TABLE III
MACHINE LEARNING ALGORITHM AND THE MODEL STRENGTH AND WEAKNESS IN ENERGY LOAD PREDICTION.

Algorithm	Model Type	Strengths	Weaknesses	Key Sources
ANN (MLP)	Neural Network	Captures nonlinear relationships	Sensitive to hyperparameters; risk of overfitting	[3]
Random Forest	Ensemble (Bagging)	Robust; reduces variance	Slightly less interpretable	[7]
Gradient Boosting	Ensemble (Boosting)	High accuracy; strong generalisation	Computationally heavier than RF	[7]

B. *Surrogate Modelling for Early-Stage Design*

As shown in Table 4, Surrogate modelling integrates machine learning with dynamic simulation outputs to enable rapid prediction of building performance [6], [22]. In this framework, ML models are trained on simulation-generated datasets and subsequently used to provide near-instantaneous predictions for new design configurations.

This approach significantly reduces computational burden and supports rapid parametric exploration during conceptual design. Rather than replacing simulation tools entirely, surrogate models serve as decision-support systems. Promising configurations identified through ML prediction can later be validated using high-fidelity simulation. In cooling-dominated climates, early design decisions related to envelope transparency and compactness strongly influence operational performance

https://doi.org/10.31436/ijpcc.v12i2.713

[5], [20]. Surrogate modelling therefore enhances the feasibility of performance-driven architectural optimisation.

TABLE IV
THE ASPECT OF MACHINE LEARNING SURROGATE MODEL.

Aspect	Dynamic Simulation	ML Surrogate Model
Computational Time	High	Very low after training
Accuracy	High (physics-based)	High (data-driven approximation)
Early-Stage Suitability	Limited	Highly suitable
Parametric Exploration	Time-intensive	Rapid iteration possible
Design Decision Support	Detailed validation	Fast screening tool

C. *Courtyard Architecture and Energy Performance*

Courtyard architecture is widely recognised as a climate-responsive typology in tropical regions. Traditional Malaysian shophouses use air-well courtyards to enhance stack ventilation and daylight penetration [8], [15]. Contemporary buildings continue to adopt courtyard forms due to their passive cooling potential referring to Table 5.

Courtyard geometry influences solar exposure, airflow patterns, and surface heat transfer. Studies show that compact forms and reduced façade transparency significantly reduce cooling demand in hot–humid climates [5], [18], [25]. Envelope-related parameters, particularly window-to-wall ratio, are consistently identified as dominant predictors of cooling load.

Despite its environmental relevance, quantitative energy modelling of L-shaped courtyard buildings under Malaysian climate conditions remains limited [9], [17]. Most courtyard studies focus on microclimate or thermal comfort rather than multi-output energy load prediction. Integrating courtyard-specific geometric parameters with machine learning techniques therefore provides a novel contribution to tropical building performance modelling.

The literature reveals three critical gaps:

- Limited machine learning studies focused specifically on tropical courtyard geometry
- Insufficient integration of courtyard morphology with surrogate modelling frameworks
- Lack of climate-specific predictive tools tailored to Malaysian design practice

The validated L-shaped courtyard dataset used in this study enables systematic evaluation of geometric and envelope variables under Kuala Lumpur weather conditions. By integrating ensemble machine learning techniques with courtyard design parameters, this research addresses the identified methodological and contextual gaps.

TABLE V
FOCUS OF THE STUDY, KEY FINDING AND GAP IDENTIFIED BASED FROM PREVIOUS STUDY.

Study Focus	Climate Context	Key Finding	Gap Identified	Key Source
Traditional Malaysian courtyards	Hot–humid	Improved natural ventilation	No quantitative cooling load modelling	[8]
Parametric courtyard simulations	Hot–dry	Courtyard geometry reduces cooling load	Limited tropical validation	[4]
Envelope optimisation	Tropical	WWR strongly influences cooling	Not courtyard-specific	[5]

III. METHODOLOGY

This study adopts a structured machine learning framework to predict cooling load, heating load, and total load of L-shaped courtyard buildings under Malaysia's hot–humid tropical climate. The methodological approach follows the CRISP-DM (Cross Industry Standard Process for Data Mining) framework The workflow progresses systematically from problem definition to deployment of the best-performing predictive model.

The first stage is business understanding. The primary objective of the study is to develop predictive models capable of estimating three energy performance indicators: cooling load, heating load, and total load. These outputs are derived from a validated parametric simulation dataset of L-shaped courtyard buildings located in Kuala Lumpur. The aim is to create a surrogate modelling tool that reduces dependence on repeated dynamic simulations while maintaining high predictive accuracy for early-stage architectural design decisions.

The second stage is data understanding and pre-processing. The dataset used in this study was derived from a validated parametric simulation dataset of L-shaped courtyard buildings developed in [9]. The dataset represents building energy performance under Kuala Lumpur weather conditions and consists of architectural variables describing building geometry, façade transparency, orientation, and compactness. Compared with operational datasets collected from existing buildings, the simulation-based dataset provides a controlled environment for systematically evaluating the influence of individual design parameters on building energy performance while eliminating variations caused by occupant behaviour and operational uncertainties (see Table 6).

https://doi.org/10.31436/ijpcc.v12i2.713

TABLE VI
DATASET UNDERSTANDING

Variable	Description	Unit
WIDTH	Building width	m
LENGTH	Building length	m
HEIGHT	Building height	m
ORIENTATION	Building orientation relative to north	Degree
WINDOWS_RATIO	Window-to-wall ratio	Ratio
FORM_FACTOR	Building compactness indicator	Dimensionless
S_V_RATIO	Surface-to-volume ratio	Dimensionless
COOLING_LOAD	Annual cooling energy demand	kWh/m²
HEATING_LOAD	Annual heating energy demand	kWh/m²
TOTAL_LOAD	Sum of annual cooling and heating energy demand	kWh/m²

Seven architectural input variables were considered, namely WIDTH, LENGTH, HEIGHT, ORIENTATION, WINDOWS_RATIO, FORM_FACTOR, and S_V_RATIO. These variables represent building dimensions, façade characteristics, and geometric compactness that directly influence solar heat gain, heat transfer, and thermal behaviour. The corresponding output variables consist of COOLING_LOAD, HEATING_LOAD, and TOTAL_LOAD, representing the annual building energy demand expressed in kWh/m².

Prior to model development, the dataset underwent a structured data quality assessment. The original dataset was inspected for missing values, inconsistent variable names, incorrect data types, and statistical outliers. A typographical inconsistency in the original dataset was corrected by renaming the variable "HIGHT" to "HEIGHT" to ensure semantic consistency throughout the analytical workflow. Missing value analysis confirmed that the dataset contained no null observations, eliminating the need for data imputation. Outlier detection using the Interquartile Range (IQR) method further confirmed that all predictor and target variables were within acceptable statistical limits, indicating a high-quality simulation dataset suitable for machine learning modelling.

Following data quality assessment, the dataset was divided into training (70%) and testing (30%) subsets to evaluate model generalisation performance. Continuous variables were standardised where appropriate to improve optimisation stability, particularly for the Artificial Neural Network model. Exploratory Data Analysis (EDA) was subsequently conducted to examine statistical distributions, identify relationships among variables, and evaluate correlation patterns prior to model training. The analysis consistently indicated that envelope-related variables, particularly WINDOWS_RATIO, exhibited the strongest relationship with cooling load and total energy demand.

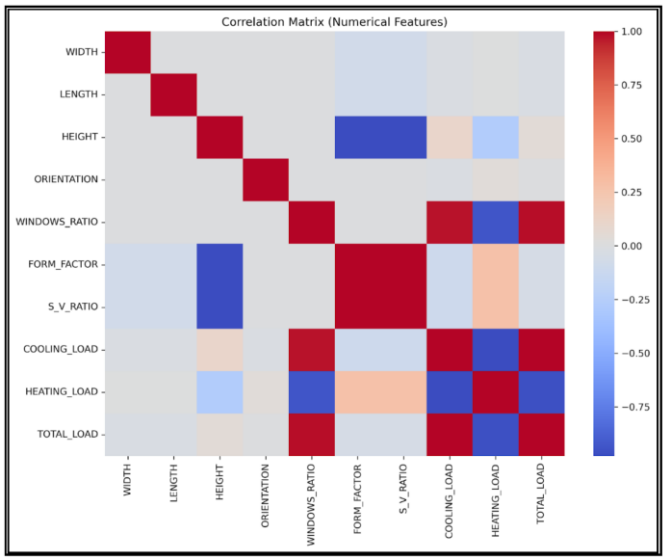


Fig 1 – Correlation Matrix between architectural design variables and target energy load during EDA.

Figure 1 illustrates the correlation matrix between the architectural design variables and the target energy loads. Among all input variables, Window-to-Wall Ratio (WINDOWS_RATIO) exhibits the strongest positive correlation with both Cooling Load and Total Load, indicating that increasing façade glazing substantially increases energy demand in Malaysia's hot-humid climate. Conversely, WINDOWS_RATIO shows a strong negative correlation with Heating Load, reflecting the inverse relationship between glazing and heating energy requirements.

https://doi.org/10.31436/ijpcc.v12i2.713

FORM_FACTOR and S_V_RATIO demonstrate moderate relationships with the energy loads, suggesting that building compactness also contributes to thermal performance. In contrast, WIDTH, LENGTH, HEIGHT, and ORIENTATION exhibit relatively weak correlations, indicating that their independent influence on energy demand is less significant than façade transparency. These observations support the subsequent machine learning results, where Gradient Boosting successfully captured the dominant influence of envelope-related variables in predicting building energy performance.

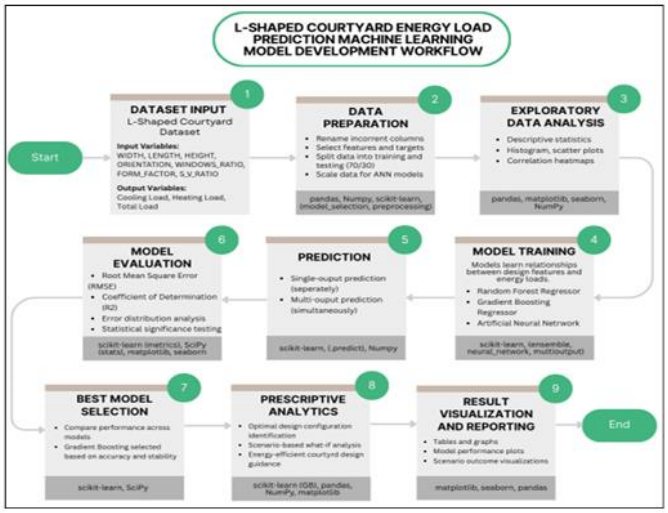


Fig 2 - Detailed Machine Learning Development Process for L-Shaped Courtyard Energy Load Prediction.

Figure 2, supports the explanation of dataset input, data preparation, exploratory analysis, model training, evaluation, best model selection, prescriptive analytics, and result visualisation.

The third stage is model development. Three supervised regression algorithms were implemented: Random Forest (RF), Gradient Boosting (GB), and Artificial Neural Network (ANN – Multilayer Perceptron). Random Forest constructs multiple decision trees using bootstrapped sampling and averages their predictions to reduce variance and improve robustness [7]. Gradient Boosting builds models sequentially, where each new tree corrects the residual errors of the previous one, resulting in high predictive accuracy and strong generalisation performance [7]. The ANN model captures nonlinear relationships through hidden layers and backpropagation optimisation [3].

The machine learning models were configured using predefined hyperparameter settings as shown in Table 7 selected to provide stable predictive performance. For the Random Forest model, the number of decision trees (n_estimators) and random state were specified prior to training. The Gradient Boosting model was configured using predefined values for the number of estimators, learning rate, and random state. For the Artificial Neural Network (ANN), the network architecture, activation function, optimisation algorithm, maximum number of iterations, and early stopping parameters were specified before model training. Using fixed hyperparameter settings ensured a consistent and reproducible modelling framework for comparing the predictive performance of the three machine learning algorithms.

TABLE VII
HYPERPARAMETER THAT IS USED BASED ON THE MACHINE LEARNING MODEL.

Model	Hyperparameter	Value
Random Forest	Number of estimators (n_estimators)	200
	Random state	42
Gradient Boosting	Number of estimators (n_estimators)	200
	Learning rate	0.1
	Random state	42
Artificial Neural Network	Hidden layer sizes	(64, 32)
	Activation function	ReLU
	Solver	Adam
	Maximum iterations	1000
	Early stopping	TRUE
	Validation fraction	0.1
	n_iter_no_change	20
	Random state	42

The fourth stage is model evaluation. Model performance was assessed using Root Mean Square Error (RMSE) and Coefficient of Determination (R^2). RMSE measures the magnitude of prediction error, while R^2 indicates how well the model explains variance in the target variable. Lower RMSE and higher R^2 values indicate better predictive performance. Residual analysis and comparative evaluation across the three models were conducted separately for cooling load, heating load, and total load. The model demonstrating the lowest prediction error and highest stability across outputs was selected as the best-performing algorithm.

The final stage involves deployment and application. The selected model functions as a surrogate predictive engine capable of providing near-instantaneous energy load predictions for new L-shaped courtyard configurations. This enables rapid parametric exploration during early-stage

https://doi.org/10.31436/ijpcc.v12i2.713

architectural design without repeatedly running computationally intensive dynamic simulations [6]. Rather than replacing physics-based simulation entirely, the model serves as a decision-support layer that accelerates conceptual evaluation while preserving the option for final simulation-based validation.

Overall, the methodology integrates climate-specific parametric simulation data with ensemble machine learning techniques to produce a reliable, efficient, and scalable predictive framework tailored to tropical courtyard buildings in Malaysia

IV. RESULTS

This section presents the predictive performance of the three machine learning models, namely Gradient Boosting (GB), Random Forest (RF), and Artificial Neural Network (ANN), for estimating cooling load, heating load, and total load of L-shaped courtyard buildings. Model evaluation was conducted using Root Mean Square Error (RMSE) and Coefficient of Determination (R^2). Lower RMSE and higher R^2 indicate superior predictive accuracy.

A. *Cooling Load Prediction Performance*

The results shown Table 8 that Gradient Boosting achieved the highest predictive performance for cooling load estimation. The model recorded an RMSE of 0.0188 and an R^2 value of 0.9997, indicating near-perfect prediction accuracy. Random Forest also performed strongly, with an RMSE of 0.0374 and R^2 of 0.9989. In contrast, the ANN model demonstrated lower predictive stability, with a higher RMSE of 0.2685 and R^2 of 0.9412.

TABLE VIII
COOLING LOAD PREDICTION PERFORMANCE

Model	RMSE	R^2	Performance Rank
Gradient Boosting	**0.0188**	**0.9997**	1 (Best)
Random Forest	0.0374	0.9989	2
ANN (MLP)	0.2685	0.9412	3

The results demonstrate that ensemble learning algorithms outperform the Artificial Neural Network (ANN) in predicting cooling load for L-shaped courtyard buildings. Among the evaluated models, Gradient Boosting achieved the lowest prediction error (RMSE = 0.0188) and the highest coefficient of determination (R^2 = 0.9997), indicating excellent agreement between predicted and simulated cooling loads. Random Forest also produced highly accurate predictions but exhibited slightly higher prediction errors than Gradient Boosting, whereas ANN showed substantially larger error dispersion and lower predictive stability.

The superior performance of Gradient Boosting can be attributed to its sequential residual-learning mechanism. Unlike Random Forest, which constructs multiple independent decision trees and averages their predictions, Gradient Boosting iteratively builds new trees that specifically correct the prediction errors generated by previous trees. This sequential optimisation enables the model to capture subtle nonlinear relationships between architectural parameters, façade characteristics, and energy performance more effectively. The comparatively lower performance of ANN may be explained by the characteristics of the simulation dataset. Although multilayer perceptron are capable of modelling highly complex nonlinear relationships, they generally require careful hyperparameter optimisation and large training datasets to achieve stable generalisation. For structured tabular datasets with a limited number of design variables, tree-based ensemble methods frequently demonstrate superior prediction accuracy and robustness.

These findings are consistent with previous building energy prediction studies. Divina et al. (2022) reported that Gradient Boosting outperformed deep learning approaches for structured building energy datasets, while Zhang et al. (2021) similarly identified tree-based ensemble models as highly effective surrogate models for building energy prediction. The present findings therefore reinforce previous evidence that Gradient Boosting provides an appropriate balance between predictive accuracy, model stability, and computational efficiency for simulation-based building energy datasets.

B. *Heating Load Prediction Performance*

Although heating demand is minimal in Malaysia's tropical climate, predictive modelling was conducted to maintain methodological completeness. Gradient Boosting again achieved the best performance referring to Table 9, with RMSE of 0.0095 and R^2 of 0.9989. Random Forest followed with RMSE of 0.0202 and R^2 of 0.9952, while ANN produced RMSE of 0.0527 and R^2 of 0.9673.

TABLE IX
HEATING LOAD PREDICTION PERFORMANCE

Model	RMSE	R^2	Performance Rank
Gradient Boosting	**0.0095**	**0.9989**	1 (Best)
Random Forest	0.0202	0.9952	2
ANN (MLP)	0.0527	0.9673	3

The consistently strong performance of Gradient Boosting across outputs suggests high model generalisation and robustness. The low prediction error for heating load

https://doi.org/10.31436/ijpcc.v12i2.713

reflects the model's ability to learn small-scale variations within a cooling-dominated climate.

Although annual heating demand is relatively insignificant in Malaysia's hot-humid climate, including heating load prediction provides a comprehensive assessment of model performance across multiple energy outputs. Gradient Boosting consistently achieved the lowest prediction error, demonstrating that the model can accurately capture even small variations within the simulation dataset. The consistent ranking of Gradient Boosting, Random Forest, and ANN across cooling and heating loads further indicates that the observed performance differences are attributable to the modelling capability of each algorithm rather than the characteristics of a specific target variable.

C. *Total Load Prediction Performance*

For total load prediction can be seen in Table 10, Gradient Boosting remained the best-performing model, achieving RMSE of 0.0143 and R^2 of 0.9997. Random Forest produced RMSE of 0.0223 and R^2 of 0.9993, while ANN recorded RMSE of 0.2339 and R^2 of 0.9192.

The results demonstrate that ensemble learning techniques consistently outperform ANN in this dataset. The superior accuracy of Gradient Boosting confirms its suitability as the final surrogate model for courtyard energy load prediction.

The consistently strong predictive performance observed for total energy load indicates that Gradient Boosting effectively captures the combined influence of multiple architectural parameters on building energy performance in Table 10. Since total load integrates both cooling and heating demand, accurate prediction requires the model to simultaneously learn complex interactions among geometric configuration, façade transparency, and compactness indicators. The near-perfect R^2 achieved by Gradient Boosting demonstrates that the surrogate model successfully approximates the dynamic simulation outputs while maintaining excellent predictive stability.

TABLE X
TOTAL LOAD PREDICTION PERFORMANCE

Model	RMSE	R^2	Performance Rank
Gradient Boosting	**0.0143**	**0.997**	1 (Best)
Random Forest	0.0223	0.9993	2
ANN (MLP)	0.2339	0.9192	3

D. *Overall Comparative Summary Across Energy Outputs*

Residual distribution analysis indicated that Gradient Boosting and Random Forest produced tightly clustered residuals around zero, demonstrating minimal bias and strong stability. ANN exhibited wider error dispersion, particularly for extreme values of cooling and total loads (see Table 11).

TABLE XI
OVERALL COMPARATIVE SUMMARY

Energy Output	Best Model	Lowest RMSE	Highest R^2	Conclusion
Cooling Load	Gradient Boosting	0.0188	0.9997	Excellent predictive accuracy
Heating Load	Gradient Boosting	0.0095	0.9989	Very stable and precise
Total Load	Gradient Boosting	0.0143	0.9997	Strong generalization

Figure 3,4 and 5 illustrates the residual distributions of the three machine learning models for cooling load, heating load, and total load prediction. Across all target variables, Gradient Boosting exhibits the narrowest residual distribution around the zero-error line, indicating consistently small prediction errors and minimal systematic bias. Random Forest demonstrates similarly stable behaviour but with slightly larger residual variation, whereas ANN produces substantially wider error dispersion and several extreme residuals, particularly for cooling load and total load prediction.

These observations are consistent with the RMSE and R^2 values reported in Tables I–III, confirming that Gradient Boosting provides the most accurate and reliable predictions among the evaluated models. The residual analysis further demonstrates the strong generalisation capability of the ensemble learning approach for simulation-based building energy prediction.

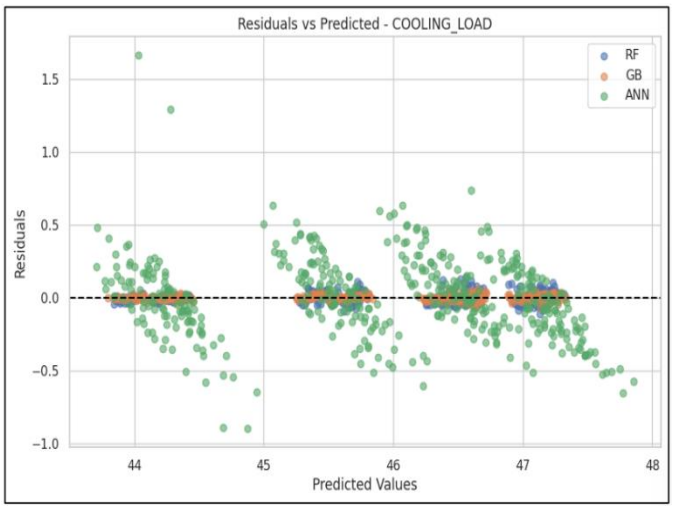


Fig 3 – Residual distribution and predicted for Cooling load

https://doi.org/10.31436/ijpcc.v12i2.713

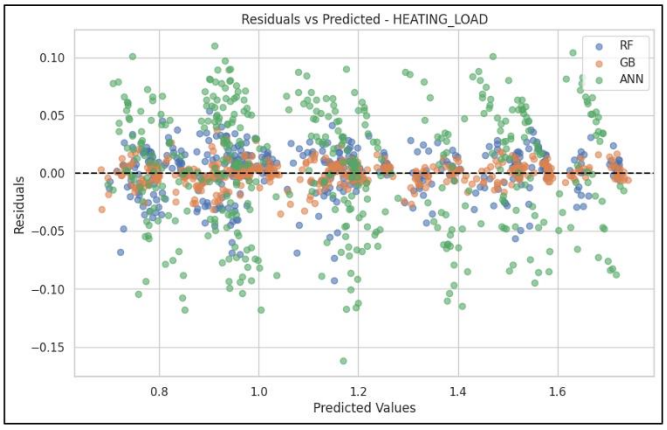


Fig 4. Residual distribution and predicted for Heating Load

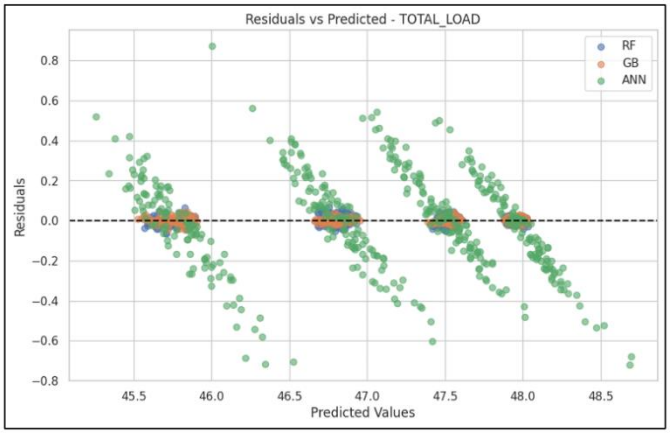


Fig. 5 - Residual distribution and predicted for Total Load

The statistical consistency across all three target outputs supports the hypothesis that ensemble-based machine learning algorithms provide superior predictive accuracy compared to standalone neural network models [3], [7].

Based on RMSE, R^2, and residual stability, Gradient Boosting was selected as the best-performing model. The model demonstrates:

- Highest predictive accuracy across all outputs
- Lowest prediction error
- Strong generalisation capability
- Stable residual distribution

The selected model serves as the final surrogate predictive engine for L-shaped courtyard energy load estimation. The high R^2 values (greater than 0.998 for ensemble models) indicate that machine learning can replicate simulation-derived energy loads with very high fidelity. This confirms that the trained model can replace repeated simulation runs during early-stage architectural design exploration.

The results validate the feasibility of integrating machine learning into tropical courtyard building design workflows, supporting rapid optimisation and energy-efficient configuration selection.

Methodologically, this research demonstrates that machine learning can function effectively as a surrogate modelling tool. The trained model enables near-instantaneous prediction of energy loads for new L-shaped courtyard configurations, significantly reducing reliance on computationally intensive dynamic simulations. Rather than replacing simulation entirely, the developed framework complements existing tools by accelerating early-stage design exploration and enabling rapid parametric assessment.

From a practical perspective, the proposed predictive framework supports architects and engineers in making data-informed design decisions during conceptual stages. By identifying high-performing geometric configurations and minimizing cooling demand, the approach contributes to energy-efficient tropical building design and aligns with Malaysia's broader sustainability and low-carbon objectives.

Overall, the results demonstrate that Gradient Boosting provides the performed best among the evaluated models for predicting building energy performance of L-shaped courtyard buildings in Malaysia. Its consistently superior performance across cooling load, heating load, and total load suggests that boosting-based ensemble learning is particularly well suited for structured simulation datasets characterised by nonlinear interactions among architectural variables.

The results also indicate that window-to-wall ratio is the dominant architectural parameter influencing cooling and total energy demand, followed by geometric compactness indicators such as FORM_FACTOR and S_V_RATIO. These findings are consistent with recent studies on tropical building envelopes, which identify façade transparency as one of the primary drivers of cooling energy consumption in hot-humid climates. Consequently, reducing excessive glazing while maintaining compact building geometry represents an effective passive design strategy for improving energy performance during the early stages of architectural design.

V. Conclusions

This study developed and evaluated a machine learning-based predictive framework for estimating cooling load, heating load, and total load of L-shaped courtyard buildings under Malaysia's hot–humid tropical climate. Using a validated parametric simulation dataset, three supervised regression models Gradient Boosting (GB), Random Forest (RF), and Artificial Neural Network (ANN) were compared to determine the most accurate and stable predictive approach.

The results demonstrate that ensemble-based models consistently outperform neural network models in this structured architectural dataset.

Among the three algorithms tested, Gradient Boosting achieved the highest predictive accuracy across all energy outputs, with R^2 values exceeding 0.998 and the lowest RMSE values for cooling, heating, and total loads. These findings confirm that Gradient Boosting provides superior generalization performance and stability for tropical building energy prediction tasks.

Although the proposed models achieved high predictive accuracy, the study has several limitations. The models were developed using a simulation-generated dataset representing L-shaped courtyard buildings under Kuala Lumpur climatic conditions. Consequently, the findings may not be directly generalisable to other building typologies, climatic regions, or operational building datasets without further validation. Future work should evaluate the proposed framework using measured building data and additional cross-validation strategies to improve robustness and generalisability.

The study also suggests that envelope-related parameters, particularly window-to-wall ratio, exert strong influence on cooling and total energy demand in courtyard buildings. Geometric compactness indicators such as form factor and surface-to-volume ratio further contribute to performance variation. These findings reinforce the importance of façade design and massing strategies in reducing cooling loads within tropical climates.

ACKNOWLEDGMENT

The author gratefully acknowledges the International Islamic University Malaysia (IIUM) and Asst. Prof. Dr Raini Hassan for providing the academic environment and institutional support that made this research possible. Special appreciation is extended for the recognition received through the Sejahtera Award, which served as motivation to pursue this work with dedication and excellence. Heartfelt thanks are also extended to the author's family for their unwavering support, encouragement, and prayers, which provided strength and perseverance in completing this study.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHOR(S) CONTRIBUTION STATEMENT

All authors contributed equally to this work.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ETHICS STATEMENT

This study did not require ethical approval.

REFERENCES

[1] A. Aldhshan, K. Maulud, O. Jaafar, H. Karim, and B. Pradhan, "Building energy consumption trends in tropical climates," Sustainable Cities and Society, vol. 72, pp. 103–118, 2021.

[2] R. Al Shawa, "Assessing the validity of simplified heating and cooling demand calculation methods: The case of Passive House Planning Package (PHPP) and Radiant Time Series Method (RTSM)," Frontiers in Built Environment, vol. 10, pp. 1–15, 2024.

[3] X. Chen, Y. Xiao, Y. Guo, and D. Yan, "Machine learning approaches for building energy prediction: A review," Energy and Buildings, vol. 279, pp. 112–130, 2023.

[4] A. Garcia-Ballesta, J. Perez-Fadon, and J. Almendros-Ibanez, "Assessment of building energy simulation tools to predict heating and cooling energy consumption at early design stages," Sustainability, vol. 15, no. 4, pp. 3450–3465, 2023.

[5] R. Gupta and C. Deb, "Envelope optimisation for cooling load reduction in tropical buildings," Building and Environment, vol. 228, pp. 109–124, 2023.

[6] P. Westermann and R. Evins, "Surrogate modelling for sustainable building design: A review," Renewable and Sustainable Energy Reviews, vol. 134, pp. 110–130, 2021.

[7] Y. Zhang, J. Wen, H. Chen, Y. Ye, and W. Livingood, "Tree-based ensemble learning for building energy prediction," Applied Energy, vol. 285, pp. 116–134, 2021.

[8] I. Zwain and A. Bahauddin, "Thermal performance of courtyard shophouses in Malaysian hot-humid climate," Journal of Asian Architecture and Building Engineering, vol. 14, no. 3, pp. 541–548, 2015.

[9] A. Almhafdy, N. Ibrahim, M. Ahmad, M. Almahmoud, and M. Alashwal, "Dataset on energy consumption in buildings within tropical climate based on design aspects of courtyards," Data in Brief, vol. 61, Art. no. 111834, Aug. 2025, doi: 10.1016/j.dib.2025.111834.

[10] International Energy Agency, Energy Efficiency 2023. Paris, France: International Energy Agency, 2023.

[11] ASEAN Centre for Energy, Southeast Asia Energy Outlook 2024. Jakarta, Indonesia: ASEAN Centre for Energy, 2024.

[12] D. Zhou, Y. Zhang, X. Li, X. Huai, and M. Xu, "Energy, environmental, and economic feasibility assessment of solar adsorption cooling system under different climate conditions in China," International Journal of Energy Research, vol. 2025, Art. no. 5377062, 2025, doi: 10.1155/ER/5377062.gap between predicted and measured energy performance of buildings: A review," Building Research & Information, vol. 52, no. 2, pp. 180–196, 2023.

[14] Y. Ren, Z. Chen, H. Zhang, and L. Liu, "Optimal classification of radiant time series for cooling load calculation," Energy and Buildings, vol. 290, pp. 113–129, 2024.

[15] G. Gunasagaran, N. Ibrahim, and A. Bahauddin, "Environmental performance of traditional Malaysian courtyard houses," Journal of Design and Built Environment, vol. 22, pp. 75–89, 2021.

[16] M. Diz-Mellado, S. Barreneche, and L. Cabeza, "Impact of courtyard geometry on building energy performance," Energy Reports, vol. 9, pp. 520–532, 2023.

[17] A. Al-Fayyad, M. Al-Mamoori, and S. Al-Shimmary, "The impact of courtyard geometry on energy efficiency in hot-dry climates," Journal of Engineering, vol. 29, no. 3, pp. 140–154, 2023.

[18] L. Liu, H. Zhang, C. Hu, Y. Tang, X. Wu, and J. Wang, "Influence of building form and envelope on cooling energy demand in hot-humid climates," Energy and Buildings, vol. 250, pp. 111–126, 2021.

[19] Z. Zhu, H. Feng, and Y. Li, "Courtyard morphology and microclimate performance in warm regions," Building and Environment, vol. 216, pp. 108–122, 2022.
[20] Y. Ali, X. Zhou, and T. Hong, "Artificial intelligence applications in building energy modelling: A systematic review," Energy and AI, vol. 7, pp. 100–118, 2023.
[21] T. Kaghembega, X. Chen, and R. Tchewafei, "Optimisation of building envelope parameters using gradient boosting models," Applied Energy, vol. 332, pp. 120–138, 2024.
[22] D. Balmer, M. Kuhn, B. Bischof, D. Salamanca, J. Kaufmann, F. Perez-Cruz, and M. Kraus, "Machine learning surrogates for parametric building performance exploration," Building Simulation, vol. 17, pp. 345–360, 2024.
[23] J. Bekdaş, B. Duta, and A. Baki, "Comparison of ANN and ensemble learning models for energy load prediction," Journal of Building Engineering, vol. 66, pp. 105–118, 2023.
[24] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001.
[25] R. Ouf, M. Gohary, and V. Ugursal, "Energy efficiency strategies in hot-humid climates: A review," Renewable and Sustainable Energy Reviews, vol. 132, pp. 110–125, 2020.
[26] Government of Malaysia, National Energy Transition Roadmap (NETR), Putrajaya, Malaysia, 2023.

International Journal on Perceptive and Cognitive Computing (IJPCC) Vol 12, Issue 2 (2026)
https://doi.org/10.31436/ijpcc.v12i2.719

An Intelligent NLP-Based System for Resume Screening and Applicant Tracking Systems (ATS)-Optimised Resume Generation

Amir 'Aatieff Amir Hussin*, Sherina Aleeyah Suparman, Hannah Sufiyyah Rushdan
Department of Computer Science, Kulliyyah of Information and Communication Technology
International Islamic University Malaysia, Kuala Lumpur, Malaysia.

*Corresponding author: amiraatieff@iium.edu.my
(Received: 7th May 2026; Accepted: 3rd June, 2026; Published on-line: 30th July, 2026)

Abstract— This paper presents the design, development, and evaluation of a web-based intelligent Progressive Web Application (PWA), that solves key issues arising from inefficiencies of current recruiting and resume writing processes. Traditionally, Applicant Tracking Systems (ATS) depend on keyword-based searching and matching, often miss out on relevant candidates because of semantic mismatch, and do not give constructive feedback to the job applicants. In this regard, a novel system that features a double-module architecture, comprising a Recruiter Module and a Job Seeker Module, has been designed to fill this void. The system makes use of named entity recognition (using spaCy and regex) for skills extraction and Sentence-BERT (all-MiniLM-L6-v2) embedding, cosine similarity for deep semantic resume-to-job matching. As a contingency plan, TF-IDF is used to ensure the system's reliability. Evaluation results show a 100% user acceptance testing (UAT) functional pass rate in 20 tests, 100% semantic matching accuracy for high, partial, and low match cases, and 97.5% Precision, 87.7% Recall, and 92.2% F1-score for NLP pipeline skills extraction.

Keywords— Natural Language Processing, Sentence-BERT, Resume Screening, Applicant Tracking Systems, Skill Extraction, Explainable AI.

I. Introduction

With competition at its peak in the current global job market environment, HR departments receive an unusually high number of resumes for each available job. To cope with such large numbers of incoming resumes, the use of Applicant Tracking Systems (ATS) tools has become common practice within organizations to screen resumes automatically at the very first stage [1]. First generation Applicant Tracking Systems, however, depend on superficial and inflexible matching based on key words. This technology does not involve any contextual understanding and therefore filters out qualified applicants whose resumes contain skills or experience expressed in synonyms [2].

Existing online recruitment platforms are silos, either designed specifically to aid recruiters in the screening process or designed to enable jobseekers to better format resumes. There is almost no system that bridges these two worlds [3]. Online resume builder tools tend to be focused solely on helping with visual design elements instead of job specific optimizations. Moreover, there is little ability to compare multiple candidates or rank them against each other. There is also an apparent lack of transparency into the workings of the automatic score-calculating "black box" algorithms. This makes it impossible to validate whether a recruiter did everything right when setting up their criteria and completely impossible for a job seeker to know why an application was rejected or how well one did at each stage of the selection process [4].

To resolve these dual limitations, this paper proposes an integrated, intelligent, web-based Progressive Web Application (PWA) designed to support both recruiters and job seekers. The proposed system introduces a dual-module architecture: a Recruiter Module that automates candidate ranking and generates AI-driven semantic summaries, and a Job Seeker Module that provides granular skill gap analysis (matched and missing skills) alongside a live, ATS-friendly resume builder. By integrating advanced Natural Language Processing (NLP) techniques including regular expressions, spaCy-based Named Entity Recognition (NER), and contextual Sentence-BERT embeddings, the proposed system replaces traditional black-box scoring with highly explainable decision support.

The main contribution of this paper is embodied in the realization of an integrated system for the bidirectional recruitment support. Specifically, the proposed tool is intended to achieve five main objectives: (1) automate the extraction of unstructured information from resumes using powerful parsing libraries; (2) perform context-aware semantic matching of resumes and job descriptions by means of transformer-based embeddings; (3) compare many candidates and rank them objectively; (4) audit the

https://doi.org/10.31436/ijpcc.v12i2.719

format and keywords of the ATS; and (5) provide structured and real-time feedback along with dynamic document generation.

II. Review of Previous Work

The recent surge in academic and commercial work indicates a growing trend towards the implementation of artificial intelligence-based technologies such as machine learning (ML) and natural language processing (NLP) into the processes of recruiting and staffing [5]. Several techniques have been developed to automate the processes of parsing, classifying and recommending resumes [6], [8]. Examples of current implementations include, ResuMate AI is a heuristic screening software that utilizes a number of rules based on ML algorithms to classify jobs and determine the experience of job applicants. While the tool employs keyword dictionaries, it lacks the ability to process deeper levels of meaning and logic or to use any type of non-standard language constructs since it is based on a set of fixed keywords. Commercial companies that develop tools for creating resumes, such as Enhancv and Zety, develop user-friendly interfaces and visual customization options, although they cannot match jobs and do not allow recruiting companies to screen resumes in bulk or rank candidates.

TABLE I
COMPARATIVE ANALYSIS OF RELATED SYSTEMS

System / Work	Methodology	Primary Strengths	Identified Gaps
Enhancv (2024)	Rule-based analysis	Strong UI; diverse layout formatting	Lacks semantic AI matching; no ranking
ResuMate AI (2025)	Heuristic NLP; MySQL	Predicts professional domain and experience	Limited contextual accuracy; rule-bound
JobAnalista (2024)	Keyword matching; NLP	Simple analysis and job recommendations	Relies on exact keywords; no ranking
ResumeCraft (2024)	Machine learning	Guided builder with automated suggestions	Focuses only on building; no screening
Prasad et al. (2023)	GNN; GloVe; Cosine	Advanced semantic and multi-graph matching	Highly complex; computationally heavy
Proposed System	spaCy NER; BERT; Cosine	Semantic ranking; explainable skill-gap feedback	Evaluated on academic scale prototype

However, as can be seen from Table I, while some of the approaches center on machine learning and sophisticated semantic matching, such as Graph Neural Network (GNN) and GloVe embedding model suggested by [9], their excessive computational cost and complexity makes them non-applicable for prototype implementation in a university/medium-size enterprise setting. Additionally, other NLP approaches, like that of [10], implement spaCy-based information extraction without any ranking of candidates or user feedback. This is where our proposed approach combines recruiter's bulk screening and job seeker optimization into one lightweight, scalable web-based system.

III. Methodology

This section provides an overview of the architecture, logical design, hybrid text extraction process, and contextual semantic matching engine that together constitute the intelligent resume screening and ranking system.

A. System Analysis and Logical Design

The system has been designed using the Agile Prototyping Model where the process can be iterated continuously depending upon the feedback of the supervisor and the users. The logical design of the system is created by means of UML diagrams in the conventional manner. There is segregation in the processes of recruiters and candidates in the system.

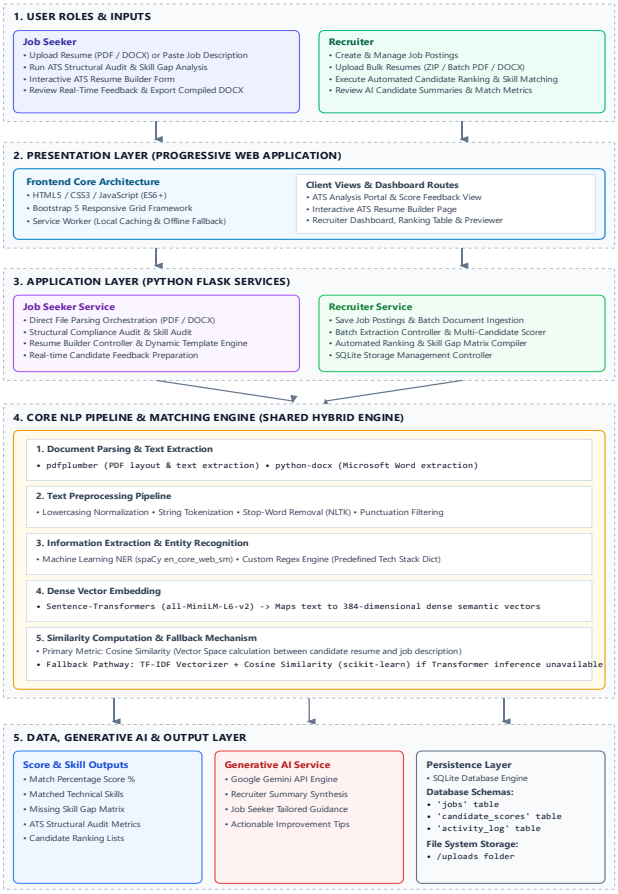


Fig. 1 System architecture and data flow diagram showing presentation, application, and data layers.

https://doi.org/10.31436/ijpcc.v12i2.719

Fig. 1 illustrates the overall system architecture, which is partitioned into three key layers: (1) the Presentation Layer comprising a cross-platform Progressive Web Application (PWA) styled with Bootstrap 5, which supports local installation and offline fallback using a service worker; (2) the Application Layer powered by a Python Flask backend and the double-core NLP engine; and (3) the Data Layer which employs a local, serverless SQLite database. The use of SQLite is selected because it requires zero setup, making it ideal for rapid prototyping and deployment on cloud hosting. The database manages three tables: 'jobs', 'candidate_scores', and 'activity'.

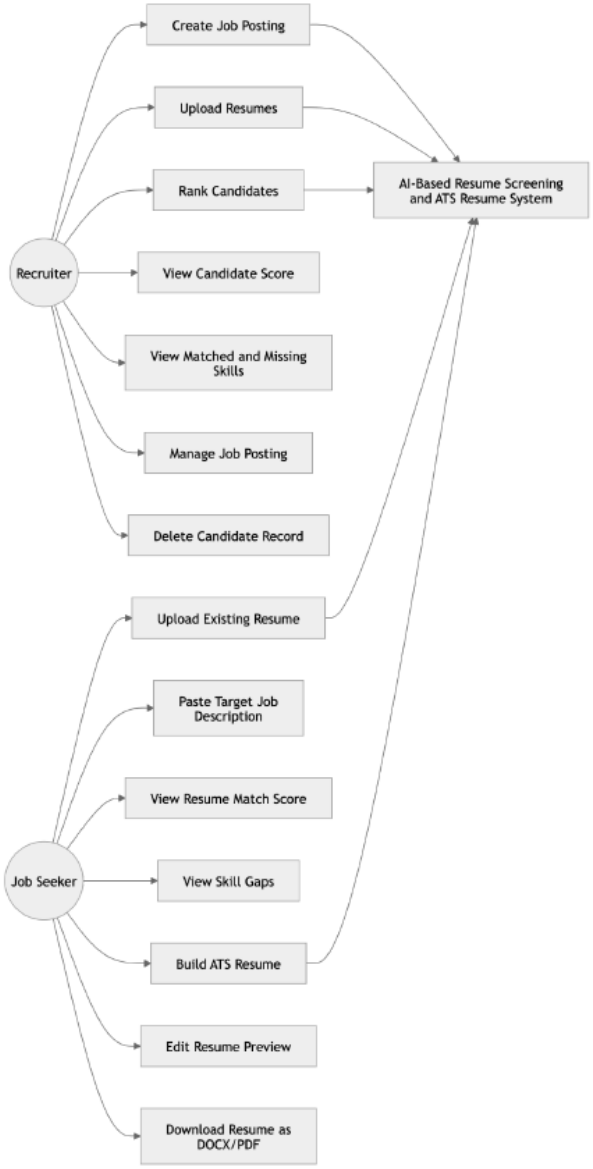


Fig. 2 Use case diagram for the dual-module intelligent resume screening and generation system

The functional boundaries of the system are described using the Use Case Diagram in Fig. 2. Recruiter actor is allowed to post jobs, perform bulk resume uploads, perform candidate ranking, and view AI summary. On the other hand, the Job Seeker actor is allowed to upload an existing resume for ATS auditing purpose or fill up the structured ATS Resume Builder form. Database ERD in Fig. 3 describes the structural relationship in the recruiter data as 1-to-many relationship of jobs and candidate score and jobs and system log records.

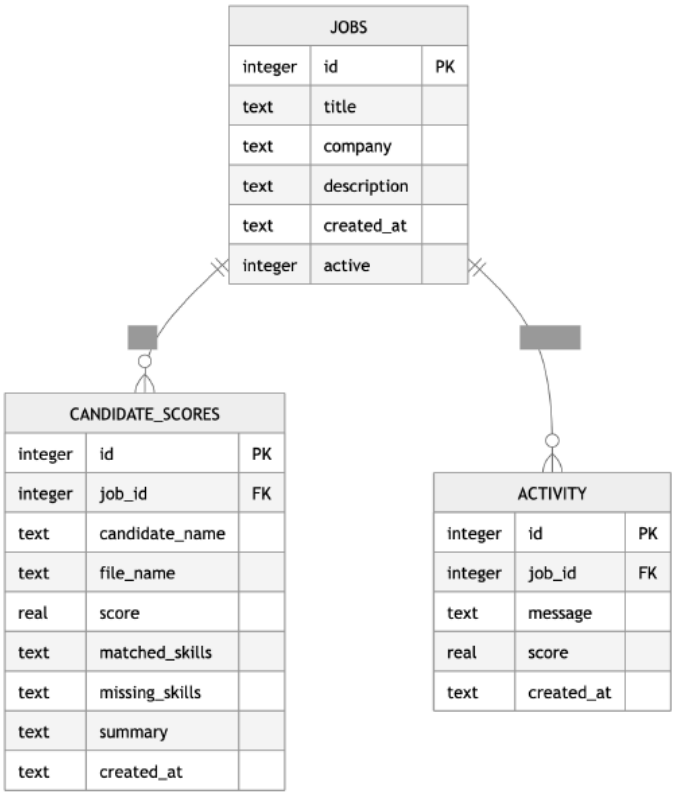


Fig. 3 Entity relationship diagram (ERD) of the recruiter-side SQLite database schema

B. NLP Parsing and Information Extraction

Once a PDF/DOCX document is uploaded, the system uses the 'pdfplumber' and 'python-docx' packages to retrieve the raw unstructured text. Preprocessing includes converting the text into lower case, tokenization, stop words removal, and punctuation removal. Extraction of information is done using the combination NLP pipeline of: (1) Regex rule-based, which matches specific programming languages, databases and tools from the pre-defined dictionary; (2) machine learning based NER using the English language model of spaCy package for extracting unstructured experience and soft skills information. This combined approach helps in maximizing information extraction while keeping the false

https://doi.org/10.31436/ijpcc.v12i2.719

positive rate almost zero. Choice of optimal representation of the features and elimination of non-informative tokens is very important in data extraction through automation, as seen in previous research on feature selection [18].

C. *Sentence-BERT Semantic Matching*

In order to evaluate the deep semantic similarity measure, both the processed text from the resume and the job description are transformed into dense vectors in high dimensional space by means of 'all-MiniLM-L6-v2' model of Sentence-Transformer model. Using this model, it is possible to express sentences in the space of 384-dimensional vectors and identify the synonyms of words based on the context and semantics.

The percentage of matching is calculated using Cosine Similarity measure. The fallback channel involves the application of content-based TF-IDF vectorizer in conjunction with cosine similarity method. The choice of the light-weight transformers is based on empirical research that proves the rationale of the usage of the low-dimensional sentence embeddings in providing the required balance between contextual accuracy and inference efficiency in matching specific domains [16], [17].

IV. Results and Discussion

This section evaluates the proposed system through user requirement validation surveys, functional User Acceptance Testing, and directional BERT scoring accuracy across baseline candidate profiles. Furthermore, it presents comprehensive performance benchmarking for the hybrid natural language processing skill extraction algorithmic pipeline.

A. *User Requirement Validation Survey*

Before development, a Google Form survey was conducted among 36 participants split evenly between recruiters (n=18) and job seekers (n=18). Survey results confirmed the importance of the problem under analysis: 94.4% of recruiters strongly agreed on manual resume screening being very time-consuming, and 88.9% of recruiters strongly agreed on the challenge of finding the right candidate among many applicants.

Moreover, 77.8% of recruiters agreed that traditional exact-match keyword searching tools often miss the most qualified candidates, and 83.3% viewed explainable scoring (e.g., lists of skills that match and those that do not match the description) as crucial. In addition, 88.9% of job seekers did not know whether their resumes matched the target job descriptions, and 72.2% had no information about ATS formatting rules.

B. *Functional User Acceptance Testing*

For evaluation of system performance, structured User Acceptance Testing (UAT) was performed with real participants. A total of 20 detailed functional tests were conducted which included 10 tests for the Recruiter Module (UAT-01 to UAT-10) involving tasks such as the posting of jobs, upload of resumes, ranking of candidates, and showing AI summarization.

There were also 10 UAT tests conducted for the Job Seeker Module (UAT-11 to UAT-20), which included analyses of resumes, ATS feedback, and resume generation/downloading. As summarized in Table II, the proposed system achieved a 100% functional pass rate across all 20 test cases, demonstrating complete functional correctness and operational reliability from end to end

TABLE II
UAT FUNCTIONAL TEST RESULTS

Module	Cases	Pass	Fail	Rate (%)
Recruiter	10	10	0	100%
Job Seeker	10	10	0	100%
Overall	**20**	**20**	**0**	**100%**

C. *BERT Scoring Directional Validation*

Since the semantic match scores provide continuous similarity percentages and not binary classifications, conventional metrics of confusion matrices are not applicable. Thus, assessing the scoring accuracy was done consistently on three pairs of known baseline resumes-job pairs including (T1) a Data Scientist profile with high match, (T2) a System Analyst profile with partial match and (T3) a Junior Quantity Surveyor profile without matching work, against the standard Data Scientist job description.

TABLE III
BERT SCORING DIRECTIONAL VALIDATION

Test	Match Level	Expected	Actual	Result
T1	High	>70%	72.0%	Pass
T2	Partial	40%-70%	45.0%	Pass
T3	No Match	<30%	8.8%	Pass

As is evident from Table III, the BERT system matched the human recruiters' judgments quite well. In T1, BERT attained 72.0%, indicating a good semantic match. In T3, BERT generated 8.8% similarity as there was no relation between the texts at all. What is most important is that in T2, BERT obtained 45.0% as it was able to understand contextual relations such as matching 'built web applications' to 'software development experience' where standard keyword indexes fail.

D. *NLP Skill Extraction Performance*

In order to evaluate the efficiency of the hybrid skill parser using the combination of NER (spaCy NER) and Regex approaches, the outputs were checked with the reference

counts for confirmed skills collected from five separate resumes. The utilization of multi-stage NLP heuristics and scoring with transformers is an example of machine learning techniques, which show that hybrid/stacked architecture provides better generalization and classification performance than isolated architectures [19].

TABLE IV
NLP SKILL EXTRACTION PERFORMANCE

Resume	Actual	Correct	Missed	Precision	Recall	F1
A	10	10	0	100.0%	100.0%	100%
B	18	14	4	87.5%	77.78%	82.5%
C	12	11	1	100.0%	91.7%	96.67%
D	9	8	1	100.0%	91.7%	96.67%
E	10	8	2	100.0%	80.0%	88.9%
Avg	11.8	10.2	1.6	97.5%	87.7%	92.2%

Table IV shows the results of the skill extraction pipeline. The system showed an average Precision of 97.5%, Recall of 87.7% and F1-Score of 92.2%. Examination of the errors (for example, skills missing from Resumes B and E) revealed that the main reason behind the false negatives was the very unusual phrasing of the skills (e.g. "experience with version control" as opposed to the word "Git"). However, the exceptional F1-score clearly shows how resilient the hybrid pipeline is.

E. Threats to Validity

Potential threats to the validity of these empirical findings were evaluated. The skill extraction benchmark relied on a small sample of ground-truth candidate resumes that while representative of standard technical layouts, other more extreme multi-column structural variations may introduce edge-case extraction variance. On external validity, the BERT semantic evaluation prioritized computer science and software engineering domains; extending these performance bounds to non-technical or creative industries may require domain-specific dictionary adjustments.

V. Conclusions and Future Work

This paper discussed the creation of a web application that combines intelligent resume screening and ATS optimization for both candidate assessment and resume building. This tool automates candidate assessment and provides understandable feedback through the integration of spaCy NER [11], Sentence-BERT semantic embeddings [12], and Generative AI [13]. Empirical testing showed successful system performance, as it passed all 20 user acceptance testing scenarios with a 100% functional pass rate, scored 100% directionally accurate, and achieved an F1-score of 92.2% for skills extraction.

However, despite such impressive performance, there are some technical constraints with the current prototype. For example, document parsing is problematic in the case of complicated multi-column structures and scanned image files [14]. Moreover, semantic matching depends largely on the job description, whereas the serverless SQLite database allows only single-thread concurrency. Finally, job-seeker analysis data is held only temporarily in memory.

The next stage of development will concentrate on scalability and parsing performance improvement. The database will be updated to PostgreSQL to ensure multi-tenant persistence and concurrent users [15]. Moreover, the use of Optical Character Recognition [14] will provide for the analysis of image-based resumes, and flexible scoring will allow recruiters to configure evaluation priorities.

The underlying web application prototype, anonymized survey datasets, and User Acceptance Testing task sheets are available from the corresponding author upon reasonable request.

Acknowledgment

The authors express their sincere gratitude to the Department of Computer Science (DCS) and the Kulliyyah of Information and Communication Technology (KICT) at the International Islamic University Malaysia (IIUM) for providing the necessary facilities, computing infrastructure, and supportive environment to conduct this project.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHOR(S) CONTRIBUTION STATEMENT

All authors contributed equally to this work.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ETHICS STATEMENT

This study did not require ethical approval.

References

[1] P. R. Chavan et al., "Enhancing recruitment efficiency: An advanced Applicant Tracking System (ATS)," 2024.
[2] A. Ashutosh and V. Kumar, "AI-powered Applicant Tracking System: An intelligent approach to modern recruitment," Int. J. Res. Appl. Sci. Eng. Technol., vol. 13, no. 11, pp. 1913–1918, 2025, doi: 10.22214/ijraset.2025.75489.
[3] C. G. Onukwugha, C. I. Ofoegbu, O. B. Aliche, and C. U. Betrand, "Resume optimization model using machine learning techniques," Int. J. Intell. Inf. Syst., vol. 13, no. 5, pp. 109–116, 2024, doi: 10.11648/j.ijiis.20241305.12.
[4] A. Beretta et al., "Requirements of eXplainable AI in algorithmic hiring," in CEUR Workshop Proc., 2024.

[5] R. Chayka, "Artificial intelligence in personnel selection: A literature review of organizational psychology," Kyiv J. Mod. Psychol. Psychother., vol. 9, pp. 122–138, 2025, doi: 10.48020/mppj.2025.01.09.
[6] S. Mohanty, A. Behera, S. Mishra, A. Alkhayyat, D. Gupta, and V. Sharma, "Resumate: A prototype to enhance recruitment process with NLP based resume parsing," in Proc. 2023 4th Int. Conf. Intell. Eng. Manage. (ICIEM), 2023, pp. 1–6, doi: 10.1109/iciem59379.2023.10166169.
[7] K. Thangaramya, G. Logeswari, S. Gajendran, J. Deepika Roselind, and N. Ahirwar, "Automated resume parsing and ranking using natural language processing," in Proc. 2024 3rd Int. Conf. Artif. Intell. Internet Things (AIIoT), 2024, pp. 1–6, doi: 10.1109/aiiot58432.2024.10574696.
[8] N. Akram, A. Majeed, S. Khan, Z. A. A. Salam, A. Sohail, and S. Shahbaz, "Automation of resume classification using machine learning algorithm," in Proc. 2023 IEEE 21st Stud. Conf. Res. Dev. (SCOReD), 2023, pp. 403–407, doi: 10.1109/scored60679.2023.10563408.
[9] R. Pradeepa, A. C. Charumathi, R. Siva, S. I. Unais Sulthan, and B. Vishnukumar, "Intelligent resume evaluation tool based on machine learning for analysis and career advancement," in Proc. 2024 Int. Conf. Emerg. Res. Comput. Sci. (ICERCS), Coimbatore, India, 2024, pp. 1–8, doi: 10.1109/ICERCS63125.2024.10895464.
[10] G. Jaiswal, A. Uttam, D. D. Dubey, and P. K. Mall, "Resume analyser and job recommendation system based on NLP," in Proc. 2nd Int. Conf. Disruptive Technol. (ICDT), 2024, pp. 1584–1587, doi: 10.1109/ICDT61202.2024.10489058.
[11] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing," Explosion AI, 2017.
[12] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in Proc. 2019 Conf. Empirical Methods Natural Lang. Process. (EMNLP-IJCNLP), 2019, pp. 3982–3992.
[13] L. Ouyang et al., "Training language models to follow instructions with human feedback," Adv. Neural Inf. Process. Syst., vol. 35, pp. 27730–27744, 2022.
[14] R. Smith, "An overview of the Tesseract OCR engine," in Proc. 9th Int. Conf. Doc. Anal. Recognit. (ICDAR), vol. 2, 2007, pp. 629–633.
[15] R. Elmasri and S. B. Navathe, Fundamentals of Database Systems, 7th ed. Boston, MA, USA: Pearson, 2015.
[16] H. Ahmad and M. N. Sulaiman, "Cognitive and Perceptive Models for Unstructured Text Matching in Human Resource Systems," International Journal on Perceptive and Cognitive Computing (IJPCC), vol. 9, no. 1, pp. 12–19, 2023.
[17] S. Z. Mohammad and K. Wan, "An Evaluation of Lightweight Sentence Embeddings for Low-Resource Domain Adaptation," International Journal on Perceptive and Cognitive Computing (IJPCC), vol. 10, no. 2, pp. 45–52, 2024.
[18] N. H. M. Khalid, A. R. Ismail, N. A. Aziz, and A. A. A. Hussin, "Performance comparison of feature selection methods for prediction in medical data," in Proc. Int. Conf. Soft Comput. Data Sci., 2023, pp. 92–106, doi: 10.1007/978-981-99-0405-1_7.
[19] M. Z. Islam, M. K. A. Hassan, A. A. A. Hussin, and M. S. Sha, "Enhancing diabetes prediction accuracy using stacked machine learning and deep learning models: A public health approach," Indonesian J. Comput. Sci., vol. 14, no. 4, 2025, doi: 10.33022/ijcs.v14i4.4947.

International Journal on Perceptive and Cognitive Computing (IJPCC) Vol 12, Issue 2 (2026)
https://doi.org/10.31436/ijpcc.v12i2.720

Enhancing Student Performance through EduMate: An AI-Powered Smart Tutoring System

Zobayer Md Ahsanul Mahbub, Mahamat Ali Adam, Norsaremah Salleh

Department of Computer Science, Kulliyyah of ICT, IIUM

*Corresponding author: norsaremah@iium.edu.my
(Received: 7th June 2026; Accepted: 3rd July, 2026; Published on-line: 30th July, 2026)

Abstract— In the modern educational landscape, students possess unique learning styles, yet many institutions continue to rely on a one-size-fits-all teaching approach. This traditional methodology often results in struggling students falling behind while advanced learners remain unchallenged. This paper presents EduMate, an AI-powered personalized tutoring system designed to adapt to individual learning styles, proficiency levels, and academic progress. Utilizing a combination of Machine Learning (ML) and Large Language Models (LLM), EduMate functions as a web-based platform that offers tailored lessons, dynamically generated quizzes, and real-time AI conversational feedback. The system features dedicated interfaces for students, teachers, and administrators, effectively bridging learning gaps while equipping educators with actionable insights into student performance.

Keywords— Personalized learning, smart tutoring system, educational technology.

I. Introduction

Education is the cornerstone of human development, yet traditional educational systems struggle to keep pace with rapid technological advancements. Traditional teaching methods predominantly treat all students as identical learners, despite the reality that each student has distinct learning styles, strengths, and pacing requirements (Luckin et al., 2016) [1]

Recently, Artificial Intelligence (AI) has emerged as a transformative tool in education. According to Holmes, Bialik, and Fadel (2019), AI enables personalized instruction, allowing students to learn at their own pace and preferred style [2]. Implementations of AI in education offer innovative ways to support students, making the learning process highly effective and engaging [3][4].

The quality of education in large classrooms is frequently compromised because teachers lack the time and resources to provide individualized attention [5]. Consequently, struggling students often go unnoticed and quietly fall behind. Furthermore, every learner possesses unique requirements; visual learners may need diagrams, while others may require interactive exercises or more time to grasp concepts [6]. When these diverse needs are unmet, students experience frustration and a loss of academic interest. Additionally, many students hesitate to ask questions due to shyness, creating cumulative learning gaps [7]. While various online learning platforms exist, most fail to offer genuine personalization, delivering static content without adjusting to the learner's knowledge level [8]. An intelligent, affordable, and flexible AI-powered tutoring system is urgently needed to provide personalized learning experiences that function akin to a private human tutor [9].

The primary objective of this project is to develop a smart tutoring system that utilizes artificial intelligence to support personalized learning by adjusting content and teaching styles based on student behavior and performance. To achieve this, the specific objectives are:

i) To study and understand different learning styles and their effects on student performance.

ii) To design and develop an AI-based system that collects data on student learning behavior and uses it to generate customized learning paths;

iii)To integrate dynamic features such as adaptive quizzes, real-time feedback, and progress tracking;

iv)To test the system with real users to analyze its effectiveness in improving learning outcomes and;

v) To ensure the resulting platform is highly accessible and easy to use for students from all backgrounds.

The proposed system is designed to facilitate personalized learning for a distinct, multi-level user base. The primary target audience includes secondary school and university students who require tailored educational content that adapts to their unique learning speed and preferences. Additionally, the system is intended for teachers and educators who seek smart technological assistance to monitor student analytics, reduce manual workloads, and offer targeted interventions. Finally, the platform serves educational institutions looking for affordable, scalable, and AI-driven tutoring solutions to elevate overall academic quality.

https://doi.org/10.31436/ijpcc.v12i2.720

This paper is organized as follows: Section II describes the Literature review. Section III describes the methodology whereas Section IV presented the analysis of results. Section V presented discussion of the findings. Finally, Section VI concludes this study.

II. Literature Review

The integration of Artificial Intelligence in education has opened new possibilities for personalized learning [9]. Several notable systems have been developed to utilize AI for tutoring and academic support. One notable platform is known as Khanmigo (by Khan Academy) [10]. It is an AI-driven virtual tutor developed in collaboration with OpenAI. It uses a conversational AI model to engage with students naturally, guiding them through problem-solving steps rather than directly providing answers. Squirrel AI Learning is another platform that pioneered adaptive learning platform from China [11]. It employs machine learning algorithms to build individualized learning paths based on detailed data gathered from student interactions.

MATHia (by Carnegie Learning) has been developed as an intelligent tutoring system [12]. It focused on mathematics that uses cognitive modeling and AI to simulate the behavior of a one-on-one human tutor. Another application is known as Syntea, which has been developed by IU International University of Applied Sciences [13]. It is an AI-powered tutoring assistant designed for university students, offering on-demand academic assistance using Natural Language Processing (NLP).

Each of these platforms brings valuable benefits to learners and educators, which demonstrate the potential of AI in education. Khanmigo guide students with problem-solving steps rather than giving direct answers. It supports critical thinking and student independence [10]. Its conversational style is less intimidating, making it highly suitable for shy students and helping them build problem-solving skills that are useful beyond specific subjects. Meanwhile, Squirrel AI Learning is a platform that excels in fine-grained personalization by detecting minor patterns in student behavior and adapting in real-time [11]. This precision allows for highly targeted interventions, reducing time wasted on already-mastered content and building student confidence. MATHia, on the hands, provides sharp, research-based mathematical instruction supported by detailed analytics. This allows teachers to identify and address learning challenges efficiently, effectively bridging the gap between traditional teaching and one-on-one tutoring. Finally, Syntea is designed for higher education and it offers flexible and immediate academic support that is especially useful for distance and adult learners. It empowers students to study independently by providing instant explanations across a wide range of subjects.

Despite their strengths, these systems exhibit distinct limitations that create barriers to widespread, general-purpose use. For instance, Khanmigo operates as a fully cloud-based service requiring constant internet access, which limits its usefulness in low-resource environments and poses a risk of students becoming overly reliant on the AI without fully engaging in deeper learning [10]. Similarly, Squirrel AI Learning relies heavily on extensive data collection from students, raising ethical and privacy concerns regarding data usage. Because it was originally designed for the Chinese education system, its global adaptations also still face challenges with diverse user interfaces and curriculum alignment. Furthermore, platforms like MATHia are exclusively focused on mathematics, restricting their usefulness for students seeking support in other academic subjects. Moreover, MATHia's interface can be overwhelming for younger or less tech-savvy users, and it lacks interactive or fun features, which can make learning feel tedious [12].

Finally, the effectiveness of systems like Syntea is strictly limited to the internal curriculum of its host institution, IU International University. This lack of broader scope restricts its usefulness for external students and limits its ability to support diverse academic needs outside of that specific environment. EduMate addresses these gaps by offering an affordable (open-source foundations), multi-subject (Science, English, Computer Science, Math), real-time adaptive platform with a highly simplified, accessible user interface.

III. Methodology

To build a flexible and responsive AI system, Extreme Programming (XP) was adopted as the core development methodology [14]. XP relies on rapid creation in smaller phases, emphasizing fast feedback, frequent testing, and user-centric design. The system was broken down into manageable modules (e.g., personalized quizzes, LLM chatbot, dashboard), allowing the team to code in short cycles, test immediately, and refine the interface continuously based on user input. Fig. 1 shows the development step employed in this study.

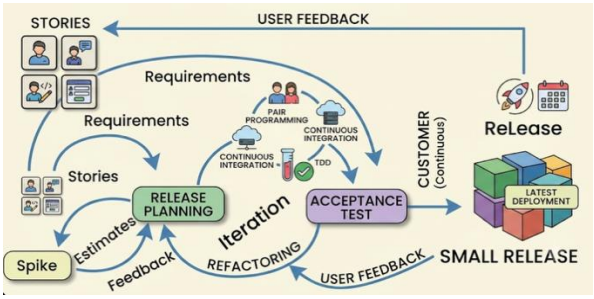


Fig. 1 Extreme Programming process

https://doi.org/10.31436/ijpcc.v12i2.720

In this study, XP was used to ensure development teams working in short, iterative cycles, typically lasting one to three weeks. Each iteration begins with the user describing desired features through "user stories," which we can then estimate for effort. The developers and users jointly negotiate which stories will be completed in the upcoming iteration, ensuring the most valuable features are prioritized and delivered in small, functional increments.

During development, XP focuses on a collaborative and technical practices to maintain quality and adaptability [14]. These core practices include pair programming, where two developers work together at a single workstation to continuously review code and solve problems; test-driven development (TDD), where automated unit tests are written before any production code; and continuous integration, requiring developers to integrate code multiple times into a universal test suite to catch errors immediately. Furthermore, the user remains actively involved as a part of the development team, providing feedback and conducting acceptance testing to verify that each small release meets their needs. Through this iterative process of listening, testing, coding, and designing, we are able to produce high-quality software that remains aligned with the customer's evolving requirements.

A. *Questionnaire Design*

To better understand the needs, preferences, and challenges faced by students in their learning journey, we conducted a survey questionnaire as part of the initial research for EduMate. The goal of the survey was to assess how students perceive current educational tools, their struggles with traditional learning methods, and their openness to AI-based personalized education. The survey was distributed online among IIUM undergraduate students in Gombak Campus. The survey questions comprise of the following questions:

- What is your preferred learning style?
- How often do you use online learning platforms? (e.g. Coursera, Udemy, Khan Academy)
- Have you ever used an AI-based learning tool before? (e.g. ChatGPT, Deepseek)
- Would you be interested in using a system that suggests lesson, quizzes, and exercises based on your learning style and performance?
- What features would you find helpful in an AI-powered tutoring system?

B. *Requirements Specification*

System requirements were driven directly by user needs. EduMate consists of the following functional requirements.

TABLE I
FUNCTIONAL REQUIREMENTS

Req ID	Description
FR-AUTH-01	The system shall allow users to register an account.
FR-AUTH-02	The system shall allow users to log in and log out securely.
FR-ROLE-01	The system shall support role-based access for Student, Teacher, and Admin.
FR-COURSE-01	The system shall allow teachers to create, update, and delete courses.
FR-ENR-01	The system shall allow teachers to enroll students into a course.
FR-ENR-02	The system shall restrict course access so only enrolled students can view course content.
FR-LESSON-01	The system shall allow teachers to create and manage lessons under a course.
FR-LESSON-02	The system shall allow teachers to upload learning materials (files/links) for lessons.
FR-QUIZ-01	The system shall allow teachers to create and manage quizzes and questions.
FR-QUIZ-02	The system shall allow students to attempt quizzes and submit answers
FR-FB-01	The system shall display quiz results and feedback after submission
FR-PROG-01	The system shall store quiz attempts and update student progress records
FR-TDASH-01 `	The system shall provide a teacher dashboard showing student performance and progress analytics.
FR-CHAT-01	The system shall provide an AI tutor chat feature for student questions and explanations.

C. *Technologies Used*

The technology stack was chosen to maximize AI integration and system scalability. Python was selected as the primary programming language due to its robust ML libraries [15]. Hugging Face Transformers (DistilBERT) and the Gemini API were utilized for lightweight, real-time natural language processing and chatbot capabilities [16]. The frontend was developed using React and Bootstrap for a responsive interface, paired with FastAPI for asynchronous backend API management [17][18][19]. Firebase and MySQL were integrated for real-time data syncing and structured relational data storage [20][21].

D. *Logical Design*

The logical design of EduMate describes how the system works before full implementation. It identifies the main users (student, teacher, and admin), key processes such as lessons, quizzes, feedback, and progress updates, and how these processes interact through the frontend, backend, and database. This design helps keep the system structured,

https://doi.org/10.31436/ijpcc.v12i2.720

reduces development errors, and ensures all features match the project requirements.

The class diagram for EduMate shows how the main parts of the system are structured and connected. At the centre is the User class, which is the base for three types of users: Student, Teacher, and Admin. Each of them has specific roles. For instance, students take quizzes and receive feedback, teachers create lessons and track progress, and admins manage users and content. The system includes Lesson, Quiz, and Question classes that represent the learning materials, and a Feedback class that gives students their results. To make the system smart and personalized, an AI Engine class generates lessons, quizzes, and feedback based on each student's performance. There is also a Chatbot that answer students' questions and a progress tracker to monitor learning progress. This structure helps organize the system clearly so that every part works together to create a smooth and personalized learning experience.

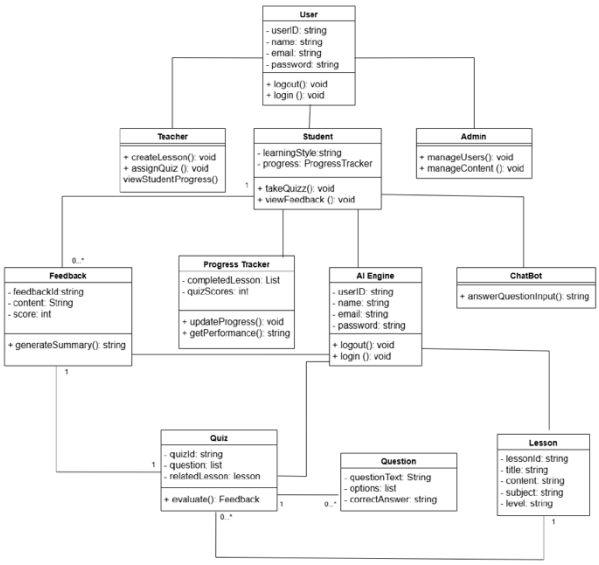


Fig. 2 Class Diagram

E. Frontend and Backend Integration

The frontend was developed using React and Bootstrap. React was used to build the pages as components. Bootstrap was used for layout and styling. It keeps the design consistent. It also supports responsive display. EduMate includes public pages such as Home, Login, and Registration. It also provides separate views for Student, Teacher, and Admin. After a user logs in, the system shows the correct dashboard based on the selected role. The frontend stores the login token and sends it with future requests. This allows the user to access protected modules. Examples include Courses, Lessons, Quizzes, AI Help, Notifications, and Profile. When the user clicks a button like "View Lesson" or "Start Quiz", the frontend sends a request to the backend and waits for the response.

The backend was developed using FastAPI. It handles the main business logic of the system. It provides API endpoints for login, user role checking, course and lesson retrieval, quiz operations, and progress updates. The backend validates each request before returning 55 data. It also ensures that students cannot access teacher or admin functions. The same rule applies to other roles.

IV. Results

A total of 54 students has responded to the survey. The findings from the survey revealed that 72.2% of respondents prefer visual, auditory, or reading/writing learning styles, with visual learning being the most favored at 33.3%. Consequently, 72.2% of students reported that they sometimes or often struggle in class because traditional teaching methods do not align with how they learn best. The findings also highlighted a strong readiness for digital and AI solutions, showing that nearly 89% of students regularly use online learning platforms and 72.2% have previously interacted with AI tools. Notably, 100% of the surveyed students expressed interest in using an AI-based educational system that personalizes lessons and quizzes to their performance. When asked about specific system features, the majority of students requested personalized learning paths (61.1%), progress tracking dashboards (61.1%), and instant feedback on quizzes (50%).

The implemented EduMate prototype successfully delivered specialized modules for distinct user roles:

- Student Dashboard: Features visual mastery tracking, enrolled course summaries, and a "Weak Topics" identifier to guide focus. Students access Personalized Quizzes, which adjust to their knowledge level to reduce stress, and receive instant feedback upon submission
- AI Help Chatbot: An integrated LLM-powered chat interface where students can ask complex academic questions and receive conversational, natural-language explanations without waiting for human intervention
- Teacher Dashboard: Provides educators with tools to upload PDF/text lessons, assign dynamic quizzes generated automatically by the AI based on lesson text, and monitor specific student metrics
- Admin Dashboard: Centralizes system monitoring, displaying active API/Database connections, and facilitating the management of user roles and course catalogs

The Landing page is the first screen displayed to users (see Fig. 3). It introduces EduMate and explains the main purpose of the platform. It also provides a clear "Get Started" button and navigation links to sections such as About,

https://doi.org/10.31436/ijpcc.v12i2.720

Features, Workflow, and Impact. From this page, users can move to the login process.

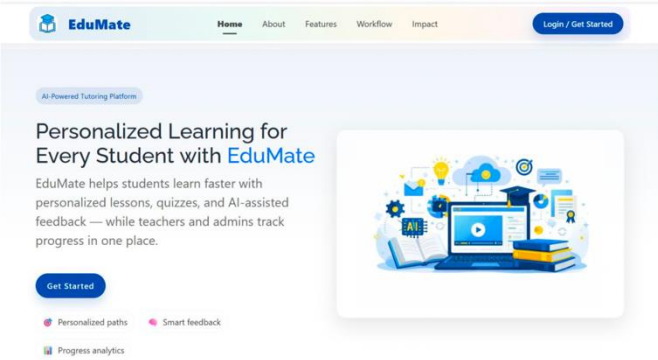


Fig. 3 Landing Page

The student interface is designed to support learning and practice. The student dashboard gives a quick summary. It shows mastery, enrolled courses, lesson count, and weak topics. Students can open Courses and Lessons to access materials (see Fig 4-5). They can also go to Quizzes and pick a lesson to attempt a quiz. The AI Help page lets students ask questions during study. It returns short explanations to support learning. These features help students follow their progress over time.

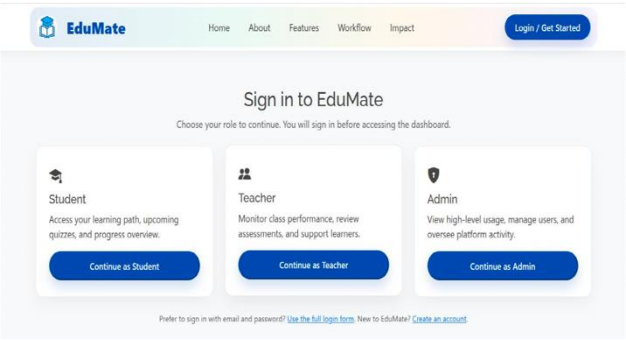


Fig. 4 User Sign-in Page

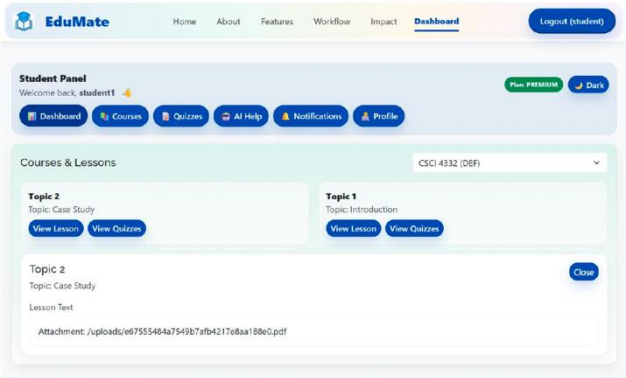


Fig. 5 Course and Lessons Page

System testing was done after the core modules of EduMate were implemented. The aim was to confirm that the system works as expected. It also checks that the system is stable during normal use. Testing was carried out from the user point of view. This means the testers followed real tasks that a student, teacher, or admin would do. The testing covered the full flow of the system. It starts from the Home page and sign-in role selection. It continues with registration and login. After login, it checks if the correct dashboard is shown. It also verifies that each role only sees the modules allowed for them. For example, students should not see admin functions. Teachers should not access admin management pages. This helps confirm role-based access control. Functional testing was applied to the key modules. For students, the tests include viewing courses, opening lesson materials, attempting quizzes, and using AI Help. For teachers, the tests include uploading lesson materials and generating quizzes. For admins, the tests include viewing system statistics and managing users and courses. During testing, the expected results were compared with the actual results. Any mismatch was recorded. After that, fixes were applied and the test was repeated. This approach helps ensure EduMate meets the main project requirements and provides a smooth experience for users.

The test plan uses task-based functional testing. Each tester followed a set of common tasks in EduMate. The tasks were based on the main modules shown in the interface. Testing involved different roles such as Student, Teacher, and Admin. All the test case includes pre-conditions, dependencies, and step-by-step actions. The expected output and actual output were recorded. The test status was marked as Pass or Fail. This method helps verify that EduMate meets the key project requirements and works smoothly for users.

V. Discussion

The deployment of EduMate demonstrates that integrating lightweight AI frameworks (DistilBERT) and external LLM APIs (Gemini) into a web-based educational platform can effectively bridge the personalization gap prevalent in traditional schooling. As identified in our survey, students are actively seeking adaptive tools. We found that from this study, EduMate meets this demand by dynamically parsing academic strengths and weaknesses to construct targeted learning paths. In contrast to existing monolithic systems that are often restricted by high subscription costs or single-subject focuses [22].

A critical factor in EduMate's success was the strategic selection of its technology stack. Python was chosen as the primary programming language due to its rich machine learning libraries, clean syntax, and superior flexibility

compared to languages like Java or C++. To power the system's natural language processing and real-time intelligent responses, the Hugging Face DistilBERT framework was integrated. DistilBERT, a lighter and faster version of the original BERT model, proved ideal for balancing high performance with efficiency, especially in an educational environment with limited computational resources [19]. Furthermore, integrating the Gemini API enabled a highly interactive conversational AI chatbot, which is capable of breaking down complex academic topics into simple, natural language explanations for students. To ensure a responsive and scalable platform, the architecture was built using React and Bootstrap for the frontend, FastAPI for asynchronous backend operations, and a hybrid database approach utilizing both Firebase (for real-time syncing) and MySQL (for structured relational data).

Despite the successful implementation, the project faced several developmental and operational constraints. Because the project was developed under a fixed semester schedule and with limited funding, the scope was restricted strictly to core functional modules, relying heavily on free or affordable open-source tools. Operationally, the system's heavy reliance on cloud-based services and external APIs (like Gemini) means it requires a highly stable internet connection. Users in low-resource environments with slow internet may experience noticeable delays when opening lesson files or waiting for AI-generated feedback. Additionally, system testing was conducted with a relatively small group of participants; while this provided valuable baseline feedback, it may not fully represent the diverse technical and educational needs of a broader global user base.

Future enhancements will focus on deepening AI personalization, improving teacher controls, and boosting student motivation. The current system's mastery logic will be refined to analyze a student's repeated mistakes and weak topics over longer periods, enabling the AI to automatically recommend highly specific practice exercises. To maintain high academic standards, future iterations will provide teachers with advanced oversight features, allowing them to preview and manually edit AI-generated quizzes before they are published to the student dashboard. Finally, to foster long-term student engagement and consistent study habits, subsequent versions of EduMate will introduce gamification elements, such as motivation badges, learning streaks, and milestone achievements.

VI. Conclusions

The EduMate project successfully conceptualized, designed, and implemented an AI-powered personalized tutoring system capable of addressing the limitations of one-size-fits-all education. By leveraging Extreme Programming and a modern Python/React technology stack, the platform provides secure, role-based environments for students, teachers, and administrators. Testing and survey results conclusively indicate that the system's dynamic quizzes, real-time AI conversational support, and automated progress tracking successfully facilitate a more engaging, tailored, and supportive learning experience. EduMate establishes a robust, accessible foundation for the future of AI-assisted, personalized education.

Acknowledgment

We would like to thank the participants who have responded to the survey and involved in giving feedback on EduMate.

Conflict of Interest

The authors declare that there is no conflict of interest.

Author(s) Contribution Statement

All authors contributed equally to this work.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Ethics Statement

This study did not require ethical approval

References

[1] R. Luckin, W. Holmes, M. Griffiths, and L. B. Forcier, *Intelligence Unleashed: An Argument for AI in Education*, London, U.K.: Pearson Education, 2016.

[2] W. Holmes, M. Bialik, and C. Fadel, Artificial Intelligence in Education: Promises and Implications for Teaching and Learning. Center for Curriculum Redesign, 2021.

[3] Y. D. Assmar, R. Contreras-Masse, M. V. H. Dozal, I. O. H. Andrade, and L. F. A. González, "Application of Artificial Intelligence in Personalized Education," in *Advances in Computational Intelligence. MICAI 2025 International Workshops* (Lecture Notes in Computer Science, vol. 16264), L. Martínez-Villaseñor *et al.*, Eds., Cham, Switzerland: Springer, 2026, pp. 110–122, doi: 10.1007/978-3-032-17930-2_10.

[4] X. Liu and B. Zhong, "Integrating generative Artificial Intelligence into student learning: A systematic review from a TPACK perspective," *Educational Research Review*, vol. 49, Art. no. 100741, 2025.

[5] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education," International Journal of Educational Technology in Higher Education, vol. 16, art. 39, 2019.

[6] B. J. Zimmerman, "Becoming a self-regulated learner: An overview," Theory Into Practice, vol. 41, no. 2, pp. 64–70, 2002.

[7] O. Henkel, H. Horne-Robinson, N. Kozhakhmetova, and A. Lee, "Effective and Scalable Math Support: Experimental Evidence on the Impact of an AI-Math Tutor in Ghana," in *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and*

Blue Sky, Cham, Switzerland: Springer Nature Switzerland, 2024, pp. 373–381.

[8] G. Kwid, N. Sarty, and D. Yang, 'A Review of AI Tools: Definitions, Functions, and Applications for K-12 Education', ACRT, vol. 3, Nov. 2024, doi: 10.5772/acrt.20240048.

[9] A. Fortuna, F. Prasetya, A. D. Samala, S. Rawas, S. Criollo-C, D. Kaya, M. Raihan, W. Andriani, D. Safitri, and R. A. Nabawi, "Artificial intelligence in personalized learning: A global systematic review of current advancements and shaping future opportunities," *Social Sciences & Humanities Open*, vol. 12, p. 102114, 2025, doi: 10.1016/j.ssaho.2025.102114.

[10] Khan Academy, "Introducing Khanmigo," 2025. [Online]. Available: https://www.khanacademy.org/khan-labs

[11] Squirrel AI Learning, "Squirrel AI," 2025. [Online]. Available: https://squirrelai.com/

[12] Carnegie Learning, "MATHia: Intelligent math software," 2025. [Online]. Available: https://www.carnegielearning.com/solutions/math/mathia

[13] IU International University of Applied Sciences, "Syntea: Your digital study assistant," 2023. [Online]. Available: https://www.iu.org

[14] A. Salem, "Extreme Programming (XP) methodology," *Scrum-Master.org*, Mar. 13, 2024. [Online]. Available: https://scrum-master.org/en/extreme-programming-xp-a-beginners-guide-to-the-agile-method/. [Accessed: May 30, 2026].

[15] Python Software Foundation, "About Python," 2025. [Online]. Available: https://www.python.org/doc/essays/blurb/

[16] Google, "Gemini API documentation," n.d. [Online]. Available: https://ai.google.dev/

[17] React, "React documentation," n.d. [Online]. Available: https://react.dev/learn

[18] Bootstrap, "Bootstrap documentation," n.d. [Online]. Available: https://getbootstrap.com/docs/

[19] FastAPI, "FastAPI documentation," n.d. [Online]. Available: https://fastapi.tiangolo.com/

[20] Firebase, "Firebase documentation," Google Developers, 2025. [Online]. Available: https://firebase.google.com/docs

[21] SQLite, "SQLite documentation," n.d. [Online]. Available: https://www.sqlite.org/docs.html

[22] S. Ritter, J. R. Anderson, K. R. Koedinger, and A. Corbett, "Cognitive Tutor: Applied research in mathematics education," Psychonomic Bulletin & Review, vol. 14, no. 2, pp. 249–255, 2016.

International Journal on Perceptive and Cognitive Computing (IJPCC) Vol 12, Issue 2 (2026)
https://doi.org/10.31436/ijpcc.v12i2.710

Sack Key: An Offline Mobile Password Manager Application

Muhammad Afiq Haikal Ahmad Tarmizi, Muhammad Anwar Mohd Asri, Norlia Md Yusof*

Department of Computer Science, Kulliyyah of Information & Communication Technology,
International Islamic University Malaysia, Gombak, Malaysia

*Corresponding author: norliayusof@iium.edu.my
(Received: 7th June 2026; Accepted: 3rd July, 2026; Published on-line: 30th July, 2026)

Abstract— The rapid growth of online services has significantly increased the number of credentials users must manage, leading to widespread password reuse and weak authentication practices. This issue is particularly severe on mobile devices, where usability constraints often conflict with secure password management. Existing password managers commonly rely on cloud-based storage and subscription models, which introduce privacy concerns and increase exposure to centralized attacks. This paper presents Sack Key, a privacy-centric mobile password manager designed with an offline-first architecture to enhance security and user control. Sack Key securely stores credentials locally using AES-256 encryption and eliminates dependency on cloud infrastructure. The system incorporates multi-layer security mechanisms, including master password, Two-Factor Authentication (2FA), automatic session locking and privacy-preserving features such as manual encryption and decryption. Additionally, Sack Key supports secure peer-to-peer password sharing over local networks without internet connectivity. The results of functional, security and user testing indicate that Sack Key provides reliable password protection, strong user trust in offline security and intuitive user experience. The findings demonstrate that an offline-first password manager can effectively balance usability, security and privacy in mobile environments.

Keywords— Password Manager, Offline-First, AES Encryption, Two-Factor Authentication, Mobile Security, Privacy.

I. Introduction

Mobile devices have become essential tools for communication, online transactions and access to digital services. As a result, users increasingly rely on smartphones to manage sensitive information such as banking credentials, personal accounts and private communications. Despite advancements in mobile security technologies, passwords remain the dominant authentication mechanism across most applications. However, managing strong and unique passwords on mobile devices is challenging due to small interfaces, frequent app switching and user preference for convenience.

Studies indicate that a significant proportion of data breaches are caused by human-related factors, including weak passwords, password reuse and phishing attacks [1]. These risks are amplified in mobile environments, where users often prioritize speed over security. Password managers have emerged as practical solutions to mitigate these issues by securely storing and generating credentials. However, many existing password managers rely heavily on cloud-based storage and continuous internet connectivity. This dependency introduces privacy risks, centralized attack surfaces and trust concerns among users who prefer full control over their sensitive data.

In addition, usability remains a critical challenge. Complex interfaces and advanced security configurations often discourage non-technical users from adopting password managers [2], [3], [4] reducing their effectiveness. Furthermore, some password managers have been shown to expose credentials through insecure autofill mechanisms [5], improper memory handling [6], or insufficient authentication protection.

To address these limitations, this paper proposes Sack Key, an offline-first [7] mobile password manager designed with a strong emphasis on privacy, security and usability. Sack Key stores all credentials locally in encrypted form using AES-256 and does not require cloud synchronization. The system integrates Two-Factor Authentication (2FA), automatic session locking and additional privacy features such as manual encrypt and decrypt to protect users in coercive or high-risk situations. Secure local network password sharing is also supported, enabling controlled credential exchange without internet exposure.

The main contributions of this work are as follows:

1) The design and implementation of an offline-first mobile password manager that eliminates cloud dependency.
2) Integration of multi-layer security mechanisms, including AES-256 encryption and 2FA.
3) Support for secure peer-to-peer password sharing over local networks.
4) Evaluation of usability and security effectiveness through user acceptance testing.

https://doi.org/10.31436/ijpcc.v12i2.710

II. RELATED WORK

Password managers have been widely examined in academic research as an effective countermeasure against weak passwords, password reuse and credential overload. Prior studies consistently demonstrate that password managers enable users to adopt unique and high-entropy passwords, thereby improving overall authentication security [1], [8], [9]. However, recent literature emphasizes that the security of password managers cannot be evaluated solely based on encryption strength, but must also consider architectural design, runtime behaviour and user interaction models [6].

Recent reviews and empirical studies published from 2024 to early 2026 highlight that most modern password managers rely on cloud-based synchronization to support multi-device access [5]. Although these systems typically employ strong encryption algorithms such as AES-256 or modern authenticated encryption schemes, researchers argue that cloud-centric designs introduce centralized trust assumptions and additional attack surfaces [10]. As a result, password manager security must be analysed holistically rather than focusing exclusively on data-at-rest protection.

A. *Security Vulnerabilities and Threat Models*

Recent academic work has expanded the threat model of password managers beyond traditional database breaches. Empirical studies show that decrypted credentials may remain in system memory during runtime, exposing them to memory dump and forensic attacks when a device is compromised [6], [11]. These findings challenge the assumption that encrypted storage alone is sufficient to protect sensitive credentials.

Browser-integrated autofill mechanisms have also been identified as a significant source of credential leakage. Research demonstrates that automated credential injections can expose passwords to malicious or hidden input fields, particularly in phishing and script-based attacks [5], [12]. Consequently, scholars increasingly recommend restricting automatic disclosure and introducing explicit user confirmation or secondary authentication before revealing sensitive information.

On mobile platforms, recent studies further reveal weaknesses in poorly implemented multi-factor authentication mechanisms. Insecure local verification and flawed session handling may allow attackers to bypass One-Time Password (OTP) or biometric checks, undermining the intended security benefits of such mechanisms [10], [13]. These findings highlight the importance of secure local enforcement rather than relying solely on authentication at login.

B. *Usability, Trust and Accessibility*

Despite the security benefits demonstrated, usability remains a major barrier to password manager adoption. Longitudinal studies involving first-time users indicate that perceived complexity, fear of credential loss and limited understanding of security mechanisms significantly reduce long-term usage [2], [3]. Importantly, these challenges are largely behavioural rather than technical.

Human-centered security research further shows that poor usability directly undermines security outcomes, as users adopt insecure coping strategies such as reusing master passwords, disabling security features, or avoiding password managers altogether [3]. Recent accessibility-focused studies reveal that users with visual impairments or low technical proficiency face additional challenges when interacting with password manager interfaces, further limiting the effectiveness of existing solutions [11]. These findings emphasize that secure systems must be designed with usability and accessibility as core requirements rather than secondary considerations.

C. *Offline and Client-Centric Password Managers*

In response to privacy and trust concerns associated with cloud-based synchronization, academic researchers have proposed offline-first or fully local password manager architectures [9], [14]. These designs aim to minimize centralized attack surfaces by ensuring that credentials and cryptographic keys remain solely on the user's device. Studies suggest that offline architecture can significantly reduce exposure to server-side attacks and third-party data handling risks.

However, existing offline solutions discussed in the literature often lack modern usability features, secure authentication layers, or mobile-centric optimization [13]. As a result, their real-world adoption remains limited despite strong theoretical security advantages. This gap indicates the need for practical implementations that balance offline security with usability and accessibility, particularly in mobile environments.

D. *Research Gap and Motivation*

Based on recent academic literature, three key gaps are identified. First, many studies emphasize encryption strength while under addressing runtime exposure risks such as memory-based attacks. Second, usability and accessibility challenges continue to hinder adoption, especially among non-technical users. Third, although offline password managers are theoretically appealing, few mobile-oriented implementations successfully integrate strong security guarantees with user-centered design.

To address these gaps, this research proposes Sack Key, a mobile password manager designed with a fully offline

https://doi.org/10.31436/ijpcc.v12i2.710

architecture, strong local encryption, enhanced authentication during credential access and usability-driven interaction. By grounding its design in recent academic findings (see Table 1), this work aims to contribute toward more secure and accessible password management solutions for everyday users.

TABLE I
COMPARISON OF FEATURES OF PASSWORD MANAGER SYSTEM

Ref	Platform	Encryption algorithms	Storage Architecture	Cloud Dependency	Authentication Support
[14]	Offline	AES-256	Local encrypted vault	None	Master password
[5]	Web	AES-based	Browser-stored credentials	Required	Browser authentication
[10]	Mobile	OS-managed	OS-level credential handling	Not required	2FA / Password-less
[6]	Mobile	AES-based	Local encrypted storage	Optional	Master password
[15]	Mobile	AES-based	Local + runtime memory	Optional	Master password
Sack Key	Mobile	AES-256	Fully local encrypted vault	None	Master password + 2FA

III. METHODOLOGY

This section describes the architectural design and security-oriented development approach adopted for Sack Key. The system was designed to operate entirely offline while maintaining strong protection of sensitive credentials and ensuring usability for non-technical users.

A. Development Methodology

Sack Key was developed following the Object-Oriented System Development (OOSD) methodology [16]. It consists of planning, analysis, design, development and testing phase (see Fig. 1). Security risks such as unauthorized access, insecure local storage and mobile-specific attacks were identified early and addressed through architectural and implementation decisions. This approach ensured that both functional and non-functional security requirements were systematically enforced.

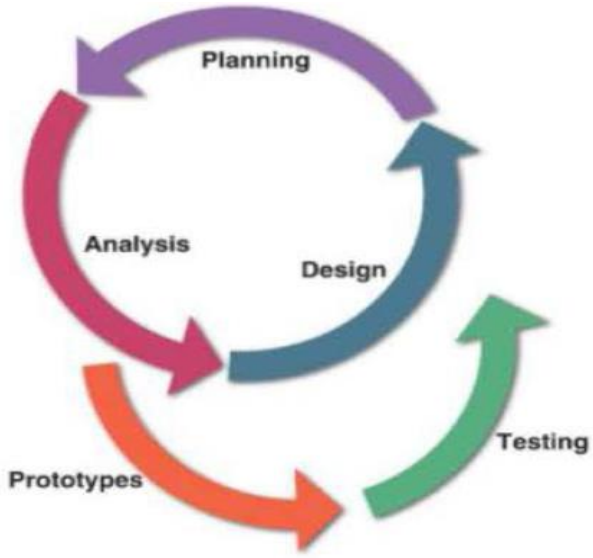


Fig. 1. OOSD adopted in Sack Key

B. System Architecture Design

Sack Key adopts a standalone mobile application architecture with an offline-first design. The architecture is composed of four main layers as illustrated in Fig. 2.

1) Presentation Layer: Provides the mobile user interface developed using Flutter. It manages user interactions, input validation and secure display of credential data by masking sensitive fields by default.

2) Application Logic Layer: Handles core functionalities including authentication, password vault operations, session management and access control. All user requests are validated before any data operation is executed.

3) Data Storage Layer: Utilizes SQLite for local data persistence. All stored credentials are encrypted prior to storage, ensuring that plaintext data is never written to disk.

4) Security Layer: Implements AES-256 encryption, 2FA, automatic session locking and privacy-preserving mechanisms such as manual encrypt and decrypt. This layer operates across all system components to enforce consistent protection of sensitive information.

https://doi.org/10.31436/ijpcc.v12i2.710

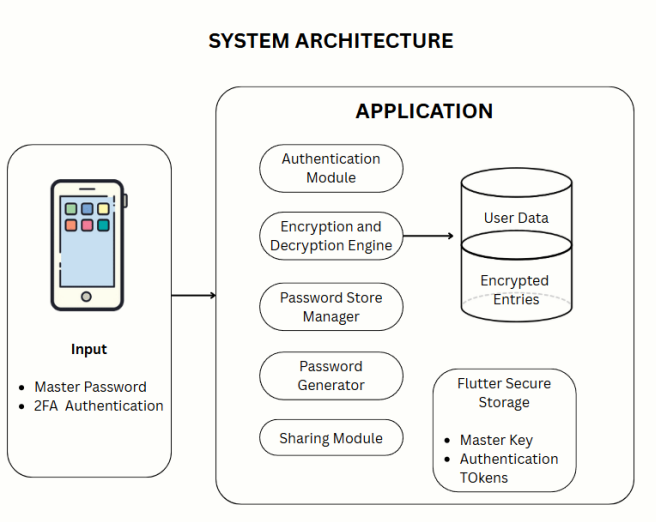


Fig. 2. Overall system architecture of Sack Key

C. Threat Model

To clarify the security scope of Sack Key, this subsection defines the threat model considered during system design. The primary threat explicitly addressed in the current implementation is SQL injection targeting the local SQLite database; all database queries are implemented using parameterized statements combined with input validation at the application logic layer, preventing malicious query fragments from being injected through credential input fields.

Runtime exposure and device-compromise threats are only partially addressed in the current design and warrant more explicit discussion, given their prominence in the literature reviewed in Section II-A [6], [11]. Decrypted vault data exists in memory only while a session is active and is dereferenced once the session ends or the application is locked (Section III-D). However, because Sack Key is implemented in Dart on the Flutter runtime, memory reclamation is handled by a managed garbage collector rather than by explicit, immediate zeroing of memory; consequently, residual data may persist in memory for a short period after dereferencing, and Sack Key does not currently perform explicit memory wiping. This is consistent with the runtime-exposure risk documented for password managers generally in prior work [6], [11], rather than a limitation unique to this system, but it means the memory-dump attack surface identified in the literature is not yet fully closed here. Similarly, the master key and authentication tokens are stored using Flutter Secure Storage (Fig. 3), which delegates to platform-level secure storage (Android Keystore or iOS Keychain) and provides hardware-backed protection against extraction on a non-compromised device. On a rooted or jailbroken device with kernel-level access, however, an attacker may still be able to circumvent these platform protections or perform live memory scraping; Sack Key does not currently implement root/jailbreak detection or anti-tampering measures to detect or respond to such compromise. Cold-boot memory extraction and backup-file extraction were likewise considered during design discussions but are not yet explicitly mitigated or empirically evaluated. Addressing these runtime and device-compromise threats, for example through explicit memory wiping where the Flutter platform permits it, root/jailbreak detection, and anti-tampering safeguards is identified as a priority direction for future development (see Section V).

D. Implementation Overview

The system is divided into four tightly integrated functional modules to support secure password management. The authentication, vault management, encryption and local sharing modules interact through the application logic layer to ensure controlled access and secure data flow.

1) Authentication Module: Manages user registration and login using a master password and optional 2FA via OTP. Credentials and security settings are stored using secure local storage to prevent exposure through local files or memory dumps. Fig. 3 and Fig. 4 depicts the user authentication and OTP verification interface respectively.

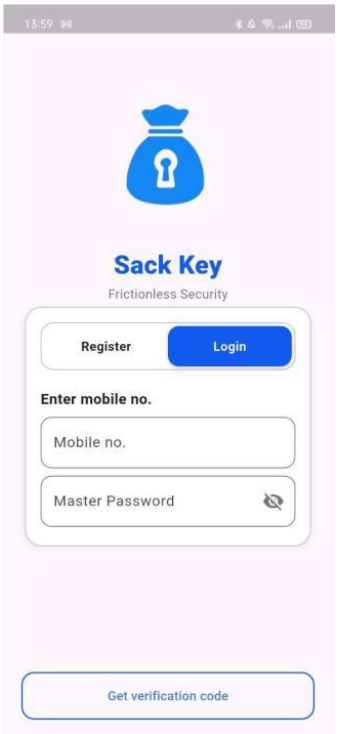


Fig. 3. User authentication and OTP verification interface

https://doi.org/10.31436/ijpcc.v12i2.710

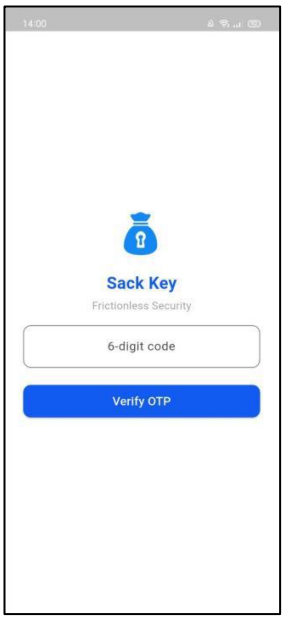


Fig. 4. User authentication and OTP verification interface

2) Password Vault Module: This module is combination of password store manager and password generator in the system architecture. It allows users to generate, view (see Fig. 5), update (see Fig. 6) and delete (see Fig. 7) stored credentials. Vault data is decrypted only in memory after successful authentication and is immediately cleared when the session ends or the application is locked.

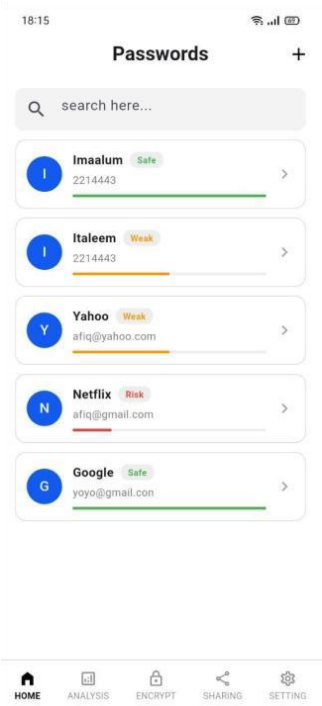


Fig. 5. Encrypted password vault interface

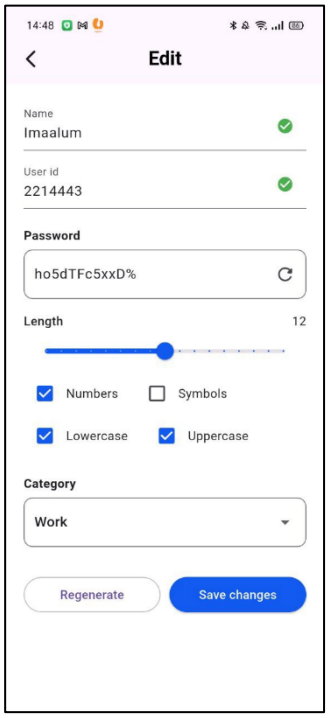


Fig. 6. Edit password interface

Fig. 7. Delete password interface

https://doi.org/10.31436/ijpcc.v12i2.710

3) Encryption Module: Provides AES-256 encryption for all sensitive data (see Figure 8). Encryption keys are derived from the user's master password, ensuring that stored data remains protected even if device storage is compromised.

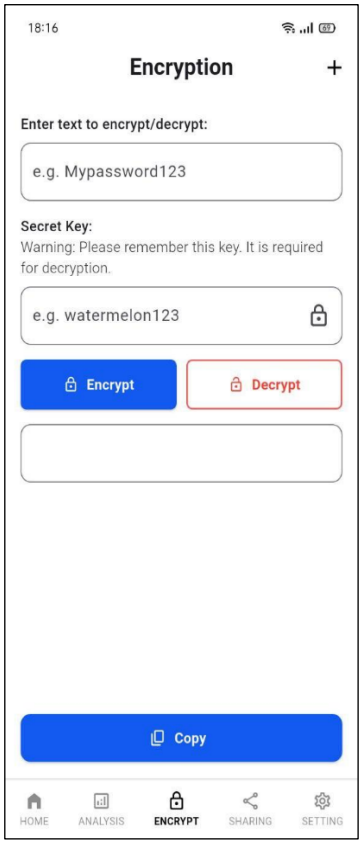


Fig. 8. Encryption and decryption feature output

4) Local Network Sharing Module: Enables secure peer-to-peer password sharing over a trusted local network using Wi-Fi Direct (see Figure 9). Shared credentials are encrypted prior to transmission and are decrypted only after successful authentication on the receiving device.

The architectural design of Sack Key integrates offline-first principles with layered security controls to protect sensitive credentials. By combining secure local storage, strong encryption and modular design, the system achieves a balance between robust security and practical usability.

IV. Results and Discussions

This section presents the results obtained from the development and evaluation of Sack Key. The evaluation focuses on functional correctness, security behavior and usability based on controlled testing and user acceptance testing.

A. *Functional Testing Results*

1) All core functionalities of Sack Key operated as intended during testing. User registration and authentication using a master password and 2FA were completed successfully without critical failures. After authentication, users were consistently granted access to the encrypted password vault.

Fig. 9. Local network password sharing interface

2) Password vault operations, including adding, editing, viewing, searching and deleting credentials, were executed correctly. Stored credentials remained encrypted at rest and decrypted values were displayed only temporarily in memory. The password generator produced strong and random passwords, supporting improved password practices among users.

B. *Security Testing Results*

Security mechanisms were validated through repeated access and usage scenarios. The system successfully enforced encryptions for all stored credentials using AES-256, preventing plaintext exposure in local storage. Automatic session locking was triggered correctly after periods of inactivity, requiring re-authentication before vault access.

https://doi.org/10.31436/ijpcc.v12i2.710

Beyond functional verification, the SQL-injection threat identified in Section III-C was assessed through two complementary checks. First, a code-level review confirmed that all database queries interacting with user-supplied input are implemented using parameterized statements, with no raw string concatenation used to construct SQL queries. Second, manual testing was conducted by submitting common SQL-injection payloads (e.g., ' OR '1'='1, '; DROP TABLE users; --) through the credential input fields. In all cases, the payloads were treated as literal input by the parameterized queries rather than executable SQL fragments; each attempt simply failed as an invalid login rather than altering the query structure or exposing any stored data, and no unauthorized data access or query manipulation was observed.

2FA effectively added an additional security layer during login and sensitive operations. The password vault feature functioned as intended by displaying non-sensitive data when activated, providing protection in coercive scenarios. Manual encryption and decryption feature also performed reliably, allowing users to securely encrypt custom text using user-defined keys.

C. *Local Network Sharing Testing Result*

The feature establishes a direct connection between two devices using Wi-Fi Direct, without relying on an intermediary access point or internet connectivity. The local peer-to-peer password sharing feature was tested within a trusted local network environment. The system successfully established secure connections between devices and transmitted encrypted credential data without internet connectivity. Shared credentials were only accessible after successful authentication on the receiving device, confirming controlled and secure data exchange.

Connection status messages and error notifications were displayed clearly when sharing was successful or interrupted, ensuring transparency during the sharing process.

Because Wi-Fi Direct relies on device-level pairing rather than a centralized trust authority, mitigating man-in-the-middle risk requires an explicit device-verification step during pairing. In Sack Key, the receiving device verifies the sharing device by scanning a QR code displayed on the sharing device before the connection is established, over the same local network, providing mutual device verification and preventing an unverified device from joining the pairing process. A formal evaluation of this QR-based pairing mechanism against relay or proximity-based attacks is noted as an area for future work.

D. *User Testing Results*

User Acceptance Testing (UAT) was conducted with users who had limited prior experience with password manager applications [2]. Table II shows that most users reported that the application was easy to understand and navigate. Authentication flows and password creation processes were rated as intuitive, while the password generator was highlighted as particularly useful.

TABLE II
OVERALL USER ACCEPTANCE TESTING

Test Area	Key Criteria	Average Rating (1–5)
Registration and Login	Ease, clarity, smoothness	4.6
Add Password	Ease, form clarity, experience	4.3
Manage Passwords	Navigation, edit/delete, clarity	3.9
Search Function	Speed, accuracy, ease	4.8
Security Features	Trust, safety, unlocking	4.2
Overall User Experience	Usability, security, acceptance	4.2

It should be noted that the "Security Features" rating in Table II reflects participants' subjective sense of trust and perceived safety while using the application, rather than a formal measure of technical security. Perceived trust and validated security are assessed separately in this study: user trust is captured through the acceptance testing reported here, while technical security is evaluated through the encryption enforcement and SQL-injection validation described in Section IV-B.

Some usability limitations were identified, including limited interaction options within the password list view and the absence of clearly visible delete actions. Despite these minor issues, users expressed strong confidence in the system's offline security model and overall reliability.

The results demonstrate that Sack Key effectively addresses key limitations of existing mobile password managers, particularly in terms of privacy, offline functionality and usability for non-technical users. The offline-first architecture successfully eliminates cloud dependency, reducing exposure to centralized data breaches and third-party trust concerns.

From a security perspective, the integration of AES-256 encryption, 2FA and automatic session locking provides layered protection against unauthorized access [4], [6], [9]. The manual encryption features further strengthen user privacy in coercive or high-risk scenarios. These design choices align with prior findings that emphasize minimizing attack surfaces and securing sensitive data at rest and in memory.

The local network sharing feature demonstrates a practical alternative to cloud-based synchronization by enabling secure credential exchange within trusted environments. While effective, this approach inherently limits scalability and cross-network access, reflecting a deliberate design trade-off favoring privacy over convenience.

UAT results indicate that strong security mechanisms can be implemented without significantly increasing user complexity. Most users were able to perform core tasks such as authentication, password storage and retrieval with minimal guidance. This suggests that security-focused mobile applications can remain accessible when interfaces are carefully designed. However, minor usability limitations observed during testing highlight the trade-off between simplicity and feature visibility.

Overall, the discussion confirms that Sack Key achieves a balanced design that prioritizes user control and data privacy while maintaining acceptable usability. The findings support the feasibility of offline-first password management systems as viable solutions for users who prefer strong security without reliance on external infrastructure.

V. Conclusion

This paper presented Sack Key, a privacy-centric mobile password manager designed to address security and usability challenges associated with cloud-dependent credential management solutions. By adopting an offline-first architecture, Sack Key eliminates reliance on centralized storage while maintaining strong protection of sensitive data through AES-256 encryption and secure local storage mechanisms.

The system integrates multiple security layers, including a master password, 2FA, automatic session locking and privacy-preserving features such as manual encrypt and decrypt. In addition, secure peer-to-peer password sharing over local networks enables controlled credential exchange without internet exposure. The application was implemented using Flutter, Dart and SQLite, ensuring portability and efficient local performance.

Evaluation results demonstrate that Sack Key operates reliably under normal usage conditions and is accessible to non-technical users. Functional testing confirmed correct behavior of core features, while user acceptance testing indicated positive perceptions of usability and trust in the offline security model. Although minor interface improvements were identified, the system successfully achieves a balance between security, privacy and ease of use.

Sack Key demonstrates the feasibility of an offline-first password management approach for mobile environments. Future work may extend this solution through biometric authentication, encrypted backup mechanisms and enhanced sharing controls, further strengthening security while preserving user autonomy.

As identified in Section III-C, future work should also prioritize runtime memory protection (e.g., explicit memory wiping where the platform allows it) and device-compromise safeguards such as root/jailbreak detection and anti-tampering measures, to close the residual attack surface highlighted by prior research on memory-based credential exposure [6], [11].

ACKNOWLEDGMENT

This study received no external funding.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

AUTHORS CONTRIBUTION STATEMENT

All authors contributed equally to this work.

DATA AVAILABILITY STATEMENT

There is no external or third-party data that support the findings of this study.

ETHICS STATEMENT

This study did not require ethical approval

References

[1] Verizon, *2025 Data Breach Investigations Report*, Verizon Enterprise, 2025.

[2] P. A. Cabarcos, M. G. López and J. L. Vázquez, "The more accounts I use, the less I have to think: A longitudinal study on the usability of password managers for novice users," in *Proc. USENIX Symp. Usable Privacy Secur. (SOUPS)*, 2025.

[3] A. Ponticello, R. Kuber and S. Branham, "How blind and low-vision users manage their passwords," in *Proc. ACM Conf. Comput. Commun. Secur. (CCS)*, Nov. 2025, doi: 10.1145/3719027.3765081, 2025.

[4] G. W. W. Mukti and R. Roestam, "Enhancing password manager application security by root detection with usability and security evaluation," in *Proc. 8th Int. Conf. Informatics Comput. (ICIC)*, pp. 1–7, doi: 10.1109/ICIC60109.2023.10382115, Dec. 01, 2023.

[5] Y. Fu and D. Wang, "Leaky autofill: An empirical study on the privacy threat of password managers' autofill functionality," in *Proc. Annu. Comput. Secur. Appl. Conf. (ACSAC)*, pp. 288–303, doi: 10.1109/ACSAC63791.2024.00037, Nov. 2024.

[6] E. Chatzoglou, V. Kampourakis, Z. Tsiatsikas, G. Karopoulos and G. Kambourakis, "Keep your memory dump shut: Unveiling data leaks in password managers," in *IFIP Adv. Inf. Commun. Technol.*, vol. 710. Cham, Switzerland: Springer, pp. 63–78, doi: 10.1007/978-3-031-65175-5_5, 2024.

[7] S. H. Pothineni, "Offline-first mobile architecture: enhancing usability and resilience in mobile systems," *Journal of Artificial Intelligence General Science (JAIGS)*, vol. 7, issue 1, Dec. 2024. [Online]. Available: https://www.researchgate.net/publication/393910615_Offline-First_Mobile_Architecture_Enhancing_Usability_and_Resilience_in_Mobile_Systems

[8] D. Beh, N. Kaur, S. Verma and A. Joshi, "Cryptographic encryption techniques in a password manager," in *IET Conf. Proc.*, vol. 2023, no. 11, pp. 346–353, doi: 10.1049/icp.2023.1803, Jan. 2023.

[9] H. Alsharaya, "Password managers: A critical review of security, usability and innovative designs," *J. Comput. Sci. Inst.*, vol. 36, pp. 328-335, June 2025.

[10] A. T. Mahdad and N. Saxena, "Mobile login bridge: Subverting 2FA and passwordless authentication via Android Debug Bridge," in *Proc. 21st Int. Conf. Privacy, Security Trust (PST)*, pp. 1–12, doi: 10.1109/PST62714.2024.10788081, Aug. 2024.

[11] E. Chatzoglou, V. Kampourakis, Z. Tsiatsikas, G. Karopoulos and G. Kambourakis, "Unmasking the hidden credential leaks in password managers," *Comput. Secur.*, vol. 140, Art. no. 103620, doi: 10.1016/j.cose.2024.103620, 2025.

[12] C. Anliker, M. Staub and D. Basin, "Phishing attacks against password manager browser extensions," in *Proc. USENIX Secur. Symp.*, 2025.

[13] M. Jubur, A. Alsaadi and K. Almarhabi, "An in-depth analysis of password managers and two-factor authentication (2FA) schemes," *ACM Comput. Surv.*, early access, doi: 10.1145/3711117, Jan. 2025.

[14] M. Abdulkadir, S. Alketbi, H. Lamaazi, R. Altamimi, S. Alblooshi and A. Lakas, "Vault-PMS: A vault-based password management system for secure offline data storage," in *Proc. 20th Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, pp. 1510–1515, doi: 10.1109/IWCMC61514.2024.10592442, May 2024.

[15] A. Gangwal, S. Singh and A. Srivastava, "AutoSpill: Credential Leakage from Mobile Password Managers," *Proceedings of the Thirteenth ACM Conference on Data and Application Security and Privacy*, pp. 39–47, doi: https://doi.org/10.1145/3577923.3583658. Apr. 2023.

[16] S. Tilley, *System Analysis and Design*, 12th ed., Boston, USA: Cengage, 2020.

International Journal on Perceptive and Cognitive Computing (IJPCC) Vol 12, Issue 2 (2026)
https://doi.org/10.31436/ijpcc.v12i2.721

GrocerEase: A Mobile Application for Recipe-based Grocery List Generation and Quantity Optimization

Alin Farhain Abdul Rajat@Abdul Razak[1], Nur Eilyana Eirdyna Ismail[2], Shuhaili Talib[3*]
1,2,3 Department of Information Systems, Kulliyyah of Information and Communication Technology, International Islamic University Malaysia (IIUM), 53100, Kuala Lumpur, Malaysia

*Corresponding author: shuhaili@iium.edu.my
(Received: 17th May 2026; Accepted: 13rd June, 2026; Published on-line: 30th July, 2026)

Abstract— GrocerEase is a mobile application created to address common challenges faced by individuals when managing meals, grocery shopping, pantry inventory, and food-related expenses. These challenges have become increasingly complex for individuals with busy lifestyles, particularly students and young adults. Existing mobile applications often address these tasks separately, resulting in fragmented experiences, repeated manual input, and limited integration between meal planning, grocery lists, and pantry tracking. This paper presents GrocerEase, a smart mobile application designed to integrate meal planning, grocery list management, pantry monitoring, and budget tracking within a single platform. A review of related applications, including Mealime, AnyList, Eat This Much, and MealBoard reveals that while they provide useful individual features, they lack comprehensive integration, intelligent scanning capabilities, and strong budget awareness. These gaps were further supported by findings from distributing a questionnaire survey, which highlighted the demand for automated grocery list generation, pantry expiry alerts, customizable recipes, and reduced manual data entry. Guided by these findings, the project adopts a structured system development methodology involving requirement analysis, system modelling, and prototype design. Unified Modelling Language (UML) diagrams, such as use case, activity, sequence, and data flow diagrams, were developed to represent system functionality, workflows, and data movement. A high-fidelity prototype was then created to demonstrate key features, including weekly and monthly meal planning, smart scanning for food and recipes, automatic grocery list generation, pantry tracking, and budget monitoring. GrocerEase apps aims to improve user efficiency, reduce food waste, and support informed decision-making through an integrated and user-centred mobile application design.

Keywords— Meal Planning, Grocery Management, Pantry Tracking, Smart Scanning, Budget Monitoring, Mobile Application.

I. INTRODUCTION

Meal planning and grocery shopping are routine activities that can become inefficient and stressful when handled without proper organisation [1]. Many individuals rely on handwritten lists or basic note-taking applications to manage their grocery needs. However, these methods do not automatically combine ingredients from multiple recipes, adjust quantities based on servings or consider items already available at home. As a result, users often forget important ingredients, purchase duplicate items or overestimate quantities, leading to food waste and unnecessary spending.

Despite the availability of various grocery and meal-planning applications, many existing systems focus only on isolated features such as recipe browsing or simple list creation. There remains a lack of integrated tools that intelligently process recipes, optimise ingredient quantities and support pantry awareness within a single mobile application [2]. These limitations highlight the need for a smarter grocery planning solution.

Current grocery planning practices suffer from several limitations. Users face difficulty managing ingredient lists when preparing multiple meals, especially when recipes require similar items in different quantities. Most existing tools require manual calculations and do not provide automated quantity scaling or ingredient consolidation [2]. In addition, the absence of pantry tracking causes users to buy items they already own, contributing to overspending and food wastage [3].

Without intelligent automation, grocery planning remains inefficient, error-prone and time-consuming [1]. These issues demonstrate the need for a system that can automatically generate optimised grocery lists by analysing recipes, adjusting quantities accordingly and filtering out items already available at home.

https://doi.org/10.31436/ijpcc.v12i2.721

The main objective of this study is to develop a mobile application that supports efficient grocery planning through automation. The specific objectives of the study are to:

1. automatically generate grocery lists based on selected recipes
2. aggregate and optimise ingredients quantities according to the number of servings
3. reduce duplicate and unnecessary purchases through pantry awareness
4. provide a structured and organized grocery planning experience for users

The scope of the study focusses on the design and development of GrocerEase, a mobile-based system/apps intended to assist users in managing meal planning, grocery lists, pantry inventory, and food-related budgeting in an integrated manner. The apps primarily target individual users, particularly students and young adults, who require an efficient and convenient solution for organizing daily food planning activities

The main entities involved in the apps include users, meal plans, recipes, grocery lists, pantry and fridge items, budget records, and scanned inputs. Users can register and log in to the system to access personalized features such as creating and managing meal plans, saving and customizing recipes, generating grocery lists automatically, tracking pantry inventory with expiry dates, and monitoring monthly food budgets. The apps also incorporate a smart scanning feature that allows users to scan food items, written recipes, and pantry contents to reduce manual data entry.

II. Literature Review

Existing apps related to grocery and meal planning were reviewed to identify their features, strengths, and limitations. An analysis highlighting similarities and differences among the systems is presented to provide a clear comparison and to identify gaps that motivate the proposed solution. The comparison is summarised in Table I.

A. Review of Existing Systems

Several existing grocery and meal-planning applications were analysed, including Mealime, AnyList, Eat This Much, and MealBoard. These applications were selected due to their relevance to meal planning, grocery list generation, and recipe management.

Mealime provides personalised meal planning and automatic grocery list generation but offers limited pantry tracking and requires subscription access for advanced features [4], [5].

AnyList focuses on shared grocery lists and recipe imports, however it lacks automated meal planning and quantity optimisation [6].

Eat This Much emphasises calorie-based meal planning and nutrition goals but is less suitable for general grocery organisation and pantry awareness [7], [8].

MealBoard supports pantry tracking and recipe scaling, but its availability is limited to IOS platforms, and its interface is less modern compared to newer applications [9].

B. Comparison Between Existing Systems

Table I presents a comparison of key features across existing grocery and meal-planning applications. While most apps support basic meal planning and grocery list creation, pantry tracking and automated quantity scaling are often missing or only partially supported [6]. Some advanced features are also restricted behind paid subscriptions, reducing accessibility for users [4].

TABLE I.

COMPARISON BETWEEN EXISTING SYSTEMS

System	Features				
	Meal Planning	**Grocery List Creation**	**Pantry/ Inventory Tracking**	**Recipe Import**	**Budget Tracking**
Mealime	✓	✓	×	✓	×
AnyList	✓	✓	×	✓	×
Eat This Much	✓	✓	×	×	✓
MealBoard	✓	✓	✓	✓	×
GrocerEase	✓	✓	✓	✓	✓

C. System Adaptation Summary

Based on the review of existing apps, GrocerEase was designed to address the identified gaps by integrating the most effective features from each application while introducing improvements. From Mealime, GrocerEase adopts structured meal planning and recipe-based grocery list generation. From AnyList, it incorporates flexible grocery list management and recipe organization concepts. From Eat This Much, it draws inspiration from budget-aware planning and nutritional considerations. These features are further enhanced through the introduction of a smart scanning function, pantry expiry alerts, and a centralized dashboard. By combining and improving upon existing

https://doi.org/10.31436/ijpcc.v12i2.721

solutions, GrocerEase aims to provide a more comprehensive, user-centred system that supports efficient food planning, reduces manual effort, and promotes better decision-making in daily meal and grocery management.

III. Methodology

A. Development Approach

Fig. 1 shows the agile development methodology adopted for this study due to its iterative and flexible nature [10], [11]. Agile supports continuous feedback, allowing requirements and system features to be refined throughout the development process [12]. This approach is suitable for GrocerEase, as user needs are derived from survey findings and may evolve based on usability evaluation and supervisor feedback [13].

The Agile lifecycle used in this project consists of requirement analysis, design, development, testing, deployment and review phases. Each iteration focuses on delivering incremental improvements, ensuring steady progress and early identification of issues [14].

Fig. 1 Agile Development Lifecycle for GrocerEase Apps

B. Requirement Gathering

User requirements were collected through an online questionnaire distributed to 50 respondents [15]. The survey aimed to understand users' grocery shopping habits, challenges and interest in potential features such as meal planning, pantry tracking, and budget monitoring [16]. The collected data was analysed to prioritise core features for the GrocerEase prototype. Ethical considerations such as voluntary participation and response anonymity were maintained throughout the data collection process [17].

The survey results indicated a strong preference for an integrated system that combines meal planning, grocery listing, and pantry tracking within a single platform [18]. Many respondents expressed interest in features that reduce manual input, particularly smart scanning for food items and recipes. Budget tracking and expiry date alerts were also highlighted as important functions for improving cost control and reducing unnecessary purchases [19]. These findings were analysed and used to prioritise the core features incorporated into the GrocerEase prototype, ensuring that the system design aligns closely with real user needs and expectations.

C. System Design Overview

Fig 2. presents the use case diagram that shows the functional scope of the GrocerEase apps and the interactions between the user and the apps [20]. The user is the primary actor who accesses the system after completing the registration or login process. Once authenticated, the user can interact with five main modules: Home (Dashboard), Plan (Meal Plan and Recipes), Grocery List, Pantry, and Profile. Each module represents a set of system functionalities designed to support meal planning, grocery organization, pantry tracking, and budget management.

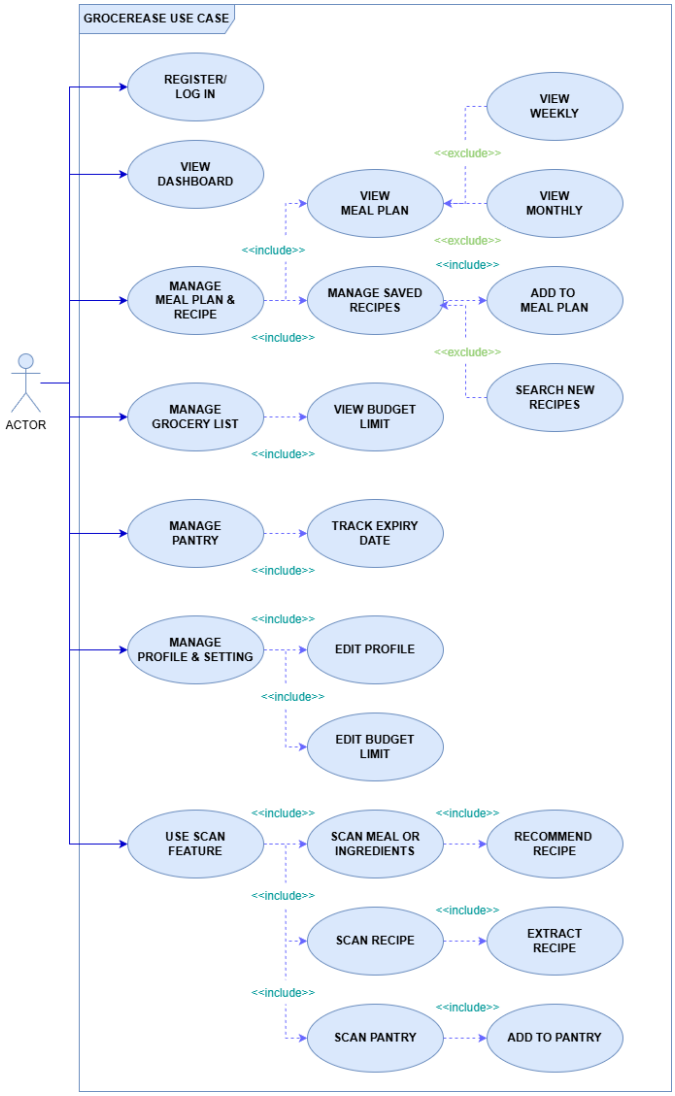


Fig. 2. Use Case Diagram for GrocerEase Application

https://doi.org/10.31436/ijpcc.v12i2.721

The diagram also includes a smart scanning feature, which is shared across multiple modules and modelled using an <<include>> relationship to represent its reuse within the system. Optional actions, such as sharing a recipe or generating a grocery list after adding a meal plan, are represented using <<extend>> relationships [21]

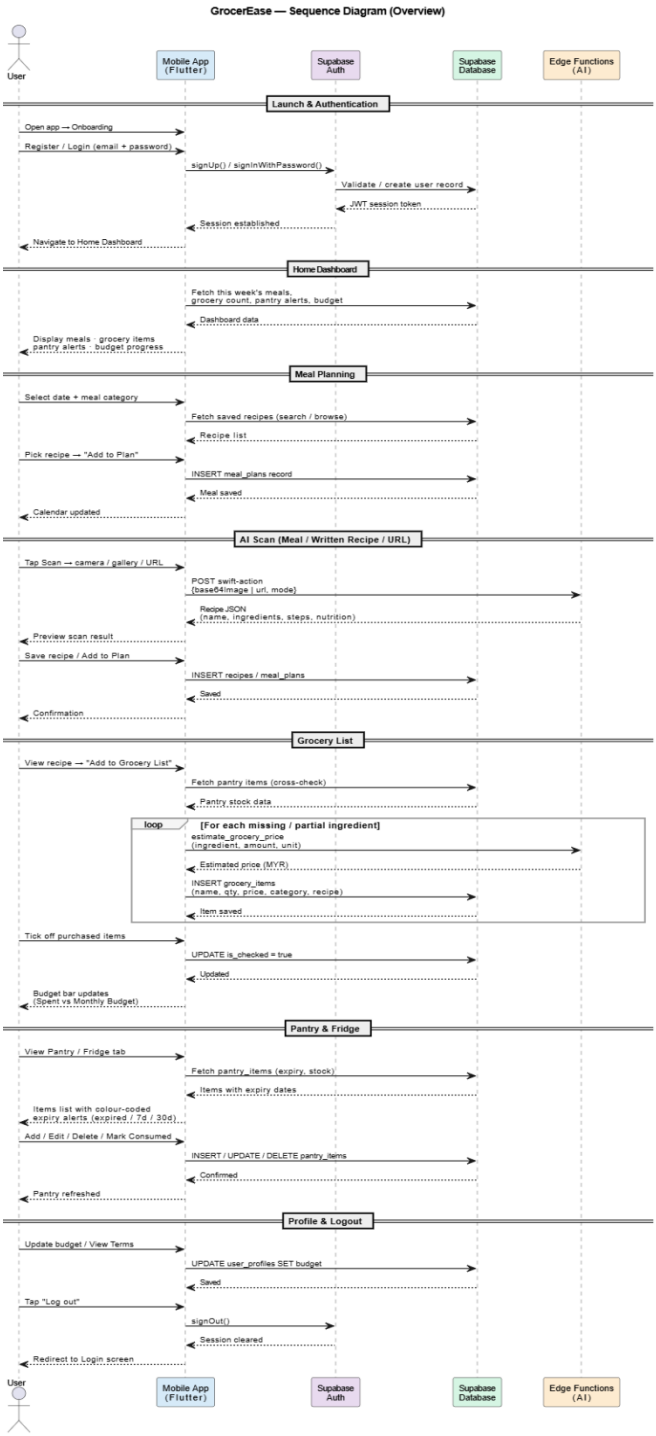


Fig. 3. Sequence Diagram for GrocerEase Application

Fig. 3 presents the end-to-end interaction flow of the GrocerEase mobile application, depicting how the user, the Flutter-based mobile interface, Supabase Authentication, Supabase Database, and the AI Edge Functions communicate with one another across the system's core features. The flow begins with the user launching the application, where they are guided through an onboarding experience before being directed to either register a new account or log in, both of which are handled securely through Supabase Authentication with email verification support via deep linking. Once authenticated, the home dashboard is loaded by simultaneously fetching the user's weekly meal plan, grocery list summary, pantry alerts, and budget information from the database, giving the user an at-a-glance overview of their current status. From there, the user is able to plan meals by selecting recipes from their saved collection or through search, with each selection being recorded into the meal plans table. The application also supports an AI-powered scanning feature, where the user can either capture or upload a photo of a meal, a set of ingredients, or a written recipe, all of which are sent to the Supabase Edge Functions for AI analysis and returned as structured recipe data that can be saved or added to the meal plan [22]. When adding ingredients to the grocery list from a recipe, the system first cross-checks the user's existing pantry and fridge items to identify what is already available, then calls the AI pricing service to estimate the cost of remaining items in the Malaysian retail market before inserting them into the grocery list. The grocery list itself supports checking off purchased items, which simultaneously updates the budget spending tracker, while the pantry module allows the user to manage food items with expiry date colour-coding to reduce food waste [23]. The sequence concludes with the user's ability to log out, which clears the active session through Supabase and redirects the application back to the login screen.

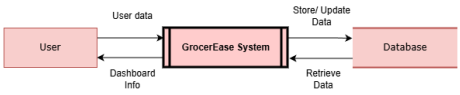


Fig. 4. Data Flow Diagram - Level 0 for GrocerEase

Fig. 4 illustrates the Level 0 Data Flow Diagram, which is also known as the context diagram, represents GrocerEase as a single process interacting with external entities [24]. In this diagram, the user is the main external entity who provides inputs such as meal plans, grocery items, pantry updates, and budget information. The system processes these inputs and returns outputs including dashboard summaries, alerts, and recommendations. A central data store represents the system database, which stores and retrieves all persistent data required by the system.

https://doi.org/10.31436/ijpcc.v12i2.721

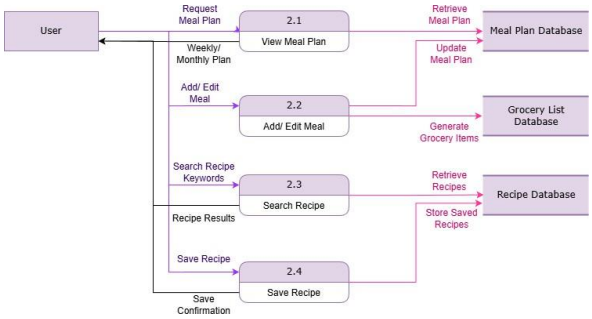


Fig. 5. Data Flow Diagram - Level 1 for GrocerEase

Fig. 5 depicts the Level 1 Data Flow Diagram that expands the single process shown in Level 0 into several major subprocesses [25]. These include managing user accounts, handling meal plans and recipes, generating grocery lists, tracking pantry inventory, and processing scanned inputs. Each process interacts with its respective data store, such as user data, meal and recipe data, grocery data, and pantry data. Data flows between the user, processes, and data stores are clearly defined to show how information is processed and stored within the system.

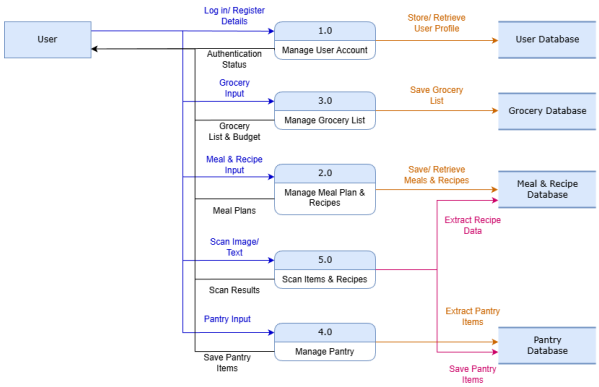


Fig. 6. Data Flow Diagram - Level 2 for GrocerEase

Fig. 6 shows the Level 2 Data Flow Diagram that further decomposes one of the main processes, specifically the meal planning and recipe management process [25]. This diagram shows detailed subprocesses such as viewing meal plans, adding or editing meals, searching for recipes, and saving selected recipes. When a meal is added, the system updates the meal plan data store and generates related grocery items, which are then sent to the grocery list data store. This detailed view illustrates how data flows between closely related features within the system. Flows between the user, processes, and data stores are clearly defined to show how information is processed and stored within the system.

D. Prototype UI/UIX Design

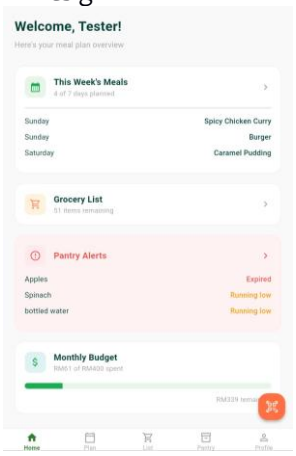


Fig. 7. Home Page

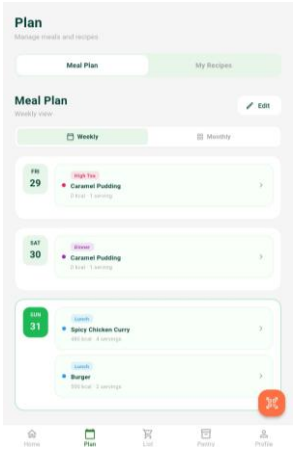


Fig. 8. Plan Page

Fig. 7 displays the home page functions as the main dashboard of the GrocerEase application. It provides a summarized view of key information, including the user's meal plan status, grocery list progress, pantry reminders, and monthly budget usage. Each section includes a navigation button that directs the user to the relevant detailed page. This design allows users to quickly understand their current food planning status and access important features with minimal effort [26].

Fig. 8 portrays the plan page that supports meal organization and recipe management. It is divided into two sections, which allow users to view meal plans and manage recipes. The meal planning section enables users to plan meals using weekly or monthly views, helping them organize their meals more effectively. When meals are added or edited, the system automatically updates the grocery list. This integration reduces manual work and improves planning efficiency.

https://doi.org/10.31436/ijpcc.v12i2.721

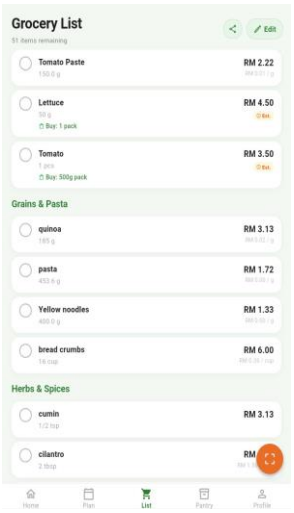


Fig. 9. List Page

Fig. 9 presents the list page displaying the user's grocery items in an organized manner. Grocery items can be viewed as a complete list or grouped by specific recipes. Each item includes quantity and estimated cost, allowing users to monitor their spending in relation to their budget. Users can also edit or remove items and share the grocery list for convenient use during shopping [27].

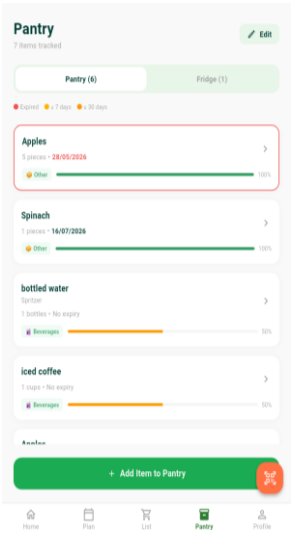


Fig.10. Pantry Page

Fig. 10 displays the pantry page allowing users to track available items stored in their pantry and fridge. Items are displayed together with quantity and expiry date information, enabling users to monitor food availability and avoid wastage. Alerts are provided for items that are nearing expiry or running low [28]. Users can update pantry records manually or through the smart scan feature.

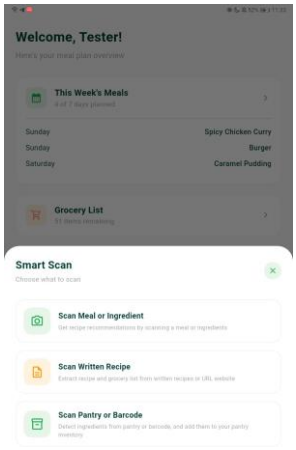


Fig. 11. Smart Scan Feature

Fig. 11 shows the smart scan feature that is designed to reduce manual data entry by allowing users to scan food items, recipes, or pantry contents. The system processes the scanned input to identify ingredients, extract recipe details, or update pantry inventory automatically. This feature improves user convenience and supports faster interaction with the application across multiple pages [29].

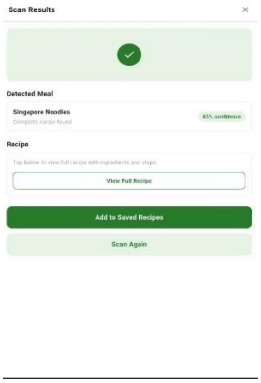
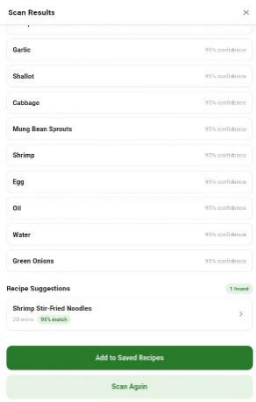


Fig. 12. Scan meal Result Page



https://doi.org/10.31436/ijpcc.v12i2.721

Fig. 13. Scan Ingredients Result Page

Fig 12 and 13 shows the scan meal and ingredients page that opens a live camera preview with two switchable modes at the top, Detect Meal and Detect Ingredients. In Detect Meal mode, the user points the camera at a prepared dish or captures it from their gallery, and the AI returns the full recipe including the name, prep and cook time, servings, estimated budget, tools required, ingredients with amounts, and step-by-step cooking instructions. In Detect Ingredients mode, the same camera flow is used but the AI instead identifies all visible raw ingredients in the image and lists them with their quantities and units. Once analysis is complete, the results are displayed on the Scan Results screen where the user can save the recipe to their library or add it directly to their meal plan.

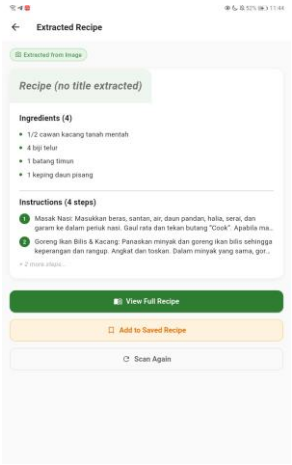


Fig. 14. Scan Written Recipe Result Page

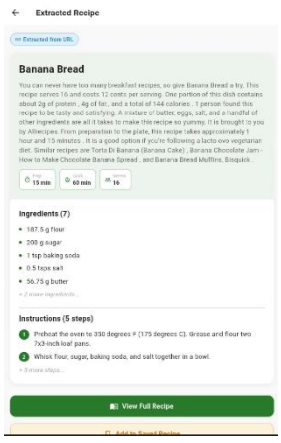


Fig. 15. Scan URL Recipe Result Page

Fig 14 and 15 presents the scan-written recipe page that allows users to extract a full recipe from either a photo of a handwritten or printed recipe, or by pasting a link to any recipe website. When an image is used, the app encodes it and sends it to the AI edge function which reads the visible text and returns the recipe title, description, ingredients, instructions, prep time, cook time, and servings. When a URL is pasted, the edge function fetches and parses the webpage on the server side, bypassing the restrictions that recipe websites often place on mobile requests and returns the same structured recipe data. The extracted recipe is shown in a preview screen where the user can review and confirm before it is saved to their recipe library.

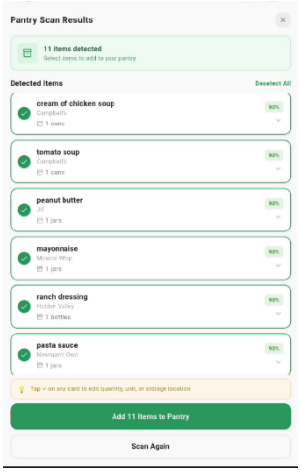


Fig. 16. Scan Pantry Result Page

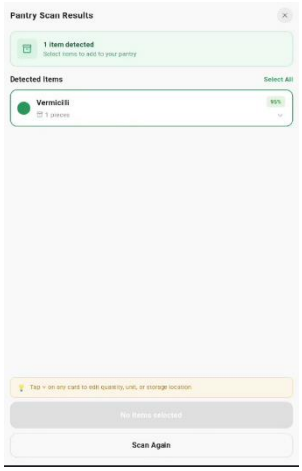


Fig. 17. Scan Barcode Result Page

Fig 16 and 17 illustrates the scan pantry and barcode page serves as a single camera screen for adding items to the pantry using two different detection methods. In Scan and Recognise mode, the user takes a photo or uploads an image

https://doi.org/10.31436/ijpcc.v12i2.721

of their pantry shelf or fridge contents, and the AI identifies all visible food items and returns them as a list with name, category, estimated quantity, unit, storage location, and confidence score. The user then selects which detected items to save to their pantry. In Scan Barcode mode, the camera continuously scans for product barcodes using real-time detection; once a barcode is found, the app looks it up against the Open Food Facts database to retrieve the product name, brand, category, and quantity. If the barcode is not found in the database, a bottom sheet appears giving the user the option to add the item manually or scan a different barcode.

IV. Project Development, Implementation and Evaluation

A. *System integration*

GrocerEase is built using Flutter (Dart) as the cross-platform mobile frontend, targeting both Android and iOS platforms. Supabase serves as the backend-as-a-service (BaaS), providing a PostgreSQL relational database for all persistent data storage and Supabase Auth for secure user registration and login. Table II summarises the technology stack.

TABLE II.

TECHNOLOGY STACK

Component	Technology/Package	Purpose
Front end	Flutter / Dart	Cross-platform mobile UI (Android & iOS)
Database	Supabase (PostgreSQL)	Cloud relational database for persistent data
Authentication	Supabase Auth	Secure user registration and login
Navigation	Go router	Declarative routing between apps screen
State management	Provider	Application state handling and UI reactivity
Smart scan	Camera image picker, http	Image capture and external AI API for food/recipe recognition

B. System Output

I. Administrator

GrocerEase apps does not include a dedicated admin panel in this prototype iteration. System administration is performed via the Supabase dashboard, where database tables, user authentication records, and API access can be managed directly.

II. User(s)

Authenticated users access the system through five modules via a bottom navigation bar:

- Home (Dashboard): Displays a summary of the weekly meal plan, grocery list progress, pantry expiry alerts and monthly budget usage, with quick-access navigation to each module.
- Plan – Meal Plan & Recipes: Supports weekly and monthly meal scheduling. When a meal is added, the system automatically aggregates its ingredients and updates the grocery list. Users can save and customise recipes with proportional quantity scaling per serving.
- Grocery List: Displays required ingredients (consolidated across all planned recipes, with pantry items deducted). Items show quantity and estimated cost. Users can edit items or share the list externally.
- Pantry: Tracks pantry and fridge contents with quantity and expiry dates. Push notifications alert users to items nearing expiry (within three days) or running low. Items can be added manually or via Smart Scan.
- Smart Scan: Provides three modes: (1) scan a food item for recipe recommendations, (2) scan a written recipe to extract ingredients via OCR and (3) scan pantry contents to auto-populate inventory. This feature reduces manual data entry across all modules.

C. System testing

I. User Acceptance Testing (UAT)

UAT was conducted with six volunteer participants (students and working adults) to verify that GrocerEase meets all defined functional requirements. Participants performed predefined tasks independently while testers recorded actual outcomes against expected results.

II. UAT Test Cases

TABLE III presents the UAT test cases developed based on the functional requirements of GrocerEase. Each test case covers a specific system feature, specifying the expected result and the actual outcome observed during testing. All ten test cases were executed by participants and returned a Pass status, confirming that the system behaves as intended across all primary modules.

TABLE III.
UAT Test Cases

ID	Test Scenario	Expected Result	Status
T C-01	User Registration & Login	User registers and logs in successfully	Pass

https://doi.org/10.31436/ijpcc.v12i2.721

T C-02	Weekly Meal Plan Creation	Meals assigned to each day; plan saved correctly	Pass
T C-03	Auto Grocery List Generation	Grocery list generated from meal plan with correct ingredients	Pass
T C-04	Ingredient Quantity Aggregation	Duplicate ingredients merged with summed quantities	Pass
T C-05	Pantry Availability Filter	Items in pantry excluded from grocery list	Pass
T C-06	Pantry Expiry Alert	Push alert triggered for items expiring within 3 days	Pass
T C-07	Smart Scan – Food Item	Food identified; relevant recipes recommended	Pass
T C-08	Smart Scan – Written Recipe	Ingredients extracted via OCR and added to grocery list	Pass
T C-09	Smart Scan – Pantry	Multiple pantry items detected and added to inventory	Pass
T C-10	Budget Monitoring	Monthly spending tracked; remaining balance displayed	Pass

III. Enhancement

Based on UAT feedback, the following improvements were made to the prototype:

- Expiry alert threshold made user-configurable (default: 3 days).
- Grocery list enhanced with category grouping (e.g., Produce, Dairy) for easier in-store use.
- Smart Scan result screen updated to allow ingredient review and editing before submission.

V. Conclusions

All four study objectives were successfully addressed in the prototype. GrocerEase automatically generates grocery lists from selected recipes, aggregates and optimises ingredient quantities based on servings, filters out pantry-available items to prevent duplicate purchases and provides a structured five-module interface for organised food planning. UAT results confirmed this, with all 10 test cases passing.

The study has four constraints in terms of scanning accuracy, budget data and platform. The apps using ML Kit where the performance depends on image quality, poor lighting and irregular handwriting reduce OCR precision. Grocery item prices are pre-defined estimates which real time market pricing was outside the study scope. The application was validated only on Android mobile devices, therefore, its compatibility with iOS devices is unexplored.

I. Future Enhancement (Modules)

A dedicated Administrator Dashboard should be developed for in-app user and content management. A Community module enabling recipe and meal plan sharing among users would further enhance engagement.

II. Future Enhancement (Functions)

Barcode scanning for faster pantry entry, nutritional analysis for planned meals and real-time grocery pricing through supermarket API integration are key functional improvements for future releases.

III. Future Enhancement (Features)

- Online Grocery Ordering: Direct integration with supermarket delivery services from within the app.
- Multi-User Household Mode: Shared accounts for collaborative meal planning and grocery management.
- Food Waste Analytics: Dashboard tracking discarded expired items to support waste-reduction habits.

GrocerEase addresses a clear gap in existing mobile applications by integrating meal planning, grocery management, pantry tracking, and budget monitoring within a single platform. Guided by Agile methodology and user survey findings, the prototype successfully demonstrated all core features. UAT results validated that all functional requirements were met. Despite constraints related to scope and platform coverage, the project achieves its intended objectives and establishes a solid foundation for a fully deployable, commercially viable grocery planning solution.

Acknowledgment

Both authors, Alin Farhain Binti Abdul Rajat @ Abdul Razak and Nur Eilyana Eirdyna Binti Ismail, would like to express their deepest appreciation to their project supervisor, Dr. Shuhaili Binti Talib, for her continuous guidance, valuable suggestions, close supervision and constant moral encouragement throughout the duration of this project. Her dedication and support have been greatly instrumental in the successful completion of this work.

Conflict of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors Contribution Statement

Alin Farhain Abdul Rajat @ Abdul Razak contributed to the conceptualization, literature review, system analysis,

system design, mobile application development, implementation, testing, data analysis, and preparation of the manuscript. Nur Eilyana Eirdyna Ismail contributed to the system development, prototype implementation, testing, validation, and manuscript review. Shuhaili Talib supervised the research, provided guidance on the project methodology and system design, reviewed the manuscript, and contributed to the interpretation of the findings. All authors read and approved the final version of the manuscript.

Data Availability Statement

The data supporting the findings of this project include the questionnaire responses collected during the requirement gathering phase, system design documents, and User Acceptance Testing (UAT) results used in the development and evaluation of the GrocerEase mobile application. The data are not publicly available as they contain research materials intended solely for academic purposes. However, relevant data and supporting materials may be made available by the corresponding author upon reasonable request.

Ethics Statement

This study involved the collection of questionnaire responses for system requirement analysis and the participation of volunteer users in User Acceptance Testing (UAT). Participation was voluntary, and all responses were collected anonymously. No personally identifiable information was used during the development, testing, or evaluation of the GrocerEase application. The study was conducted in accordance with institutional research ethics guidelines to ensure participant privacy, confidentiality, and responsible research practices.

References

[1] V. Stefan, E. van Herpen, A. A. Tudoran, and L. Lähteenmäki, "Avoiding food waste by Romanian consumers: The importance of planning and shopping routines," Food Quality and Preference, vol. 28, no. 1, pp. 375–381, 2013, doi: 10.1016/j.foodqual.2012.11.001.

[2] N. Punitha, A. Sibi, S. Francis, and N. Sradhaya, "AI powered meal planner," International Journal of Research Publication and Reviews, vol. 6, no. 5, pp. 3755–3766, May 2025. [Online]. Available: https://ijrpr.com/uploads/V6ISSUE5/IJRPR45071.pdf

[3] Marketing Scoop, "15 meal planning apps to make a nutritionist's job easier," Apr. 2, 2024. [Online]. Available: https://www.marketingscoop.com/tech/15-meal-planning-apps-to-make-a-nutritionists-job-easier/ [Accessed: Jul. 31, 2026].

[4] SaaSHub, "Eat This Much vs. Mealime: Compare differences and reviews," n.d. [Online]. Available: https://www.saashub.com/compare-eat-this-much-vs-mealime [Accessed: Jul. 31, 2026].

[5] S. Atejioye, "Review of meal-planning, grocery list and expense apps," SIIT, Oct. 13, 2025. [Online]. Available: https://siit.co/blog/review-of-meal-planning-grocery-list-expense-apps/48686 [Accessed: Jul. 31, 2026].

[6] Dietium, "Meal planning apps comparison—Expert review 2025," 2025. [Online]. Available: https://dietium.com/meal-planning-apps-comparison/ [Accessed: Jul. 31, 2026].

[7] Eat This Much, Inc., "Eat This Much—Meal planner," Apple App Store, n.d. [Online]. Available: https://apps.apple.com/my/app/eat-this-much-meal-planner/id981637806 [Accessed: Jul. 31, 2026].

[8] MealBoard, "MealBoard—Meal planner," Apple App Store, n.d. [Online]. Available: https://apps.apple.com/us/app/mealboard-meal-planner/id333425918 [Accessed: Jul. 31, 2026].

[9] K. Beck et al., "Manifesto for Agile software development," Agile Alliance, 2001. [Online]. Available: https://agilemanifesto.org/ [Accessed: Jul. 31, 2026].

[10] I. Sommerville, Software Engineering, 10th ed. Boston, MA, USA: Pearson, 2016.

[11] J. Highsmith, Agile Project Management: Creating Innovative Products, 2nd ed. Boston, MA, USA: Addison-Wesley, 2009.

[12] J. Rubin and D. Chisnell, Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests, 2nd ed. Indianapolis, IN, USA: Wiley, 2008.

[13] R. S. Pressman and B. R. Maxim, Software Engineering: A Practitioner's Approach, 9th ed. New York, NY, USA: McGraw-Hill Education, 2020.

[14] J. W. Creswell and J. D. Creswell, Research Design: Qualitative, Quantitative, and Mixed Methods Approaches, 5th ed. Thousand Oaks, CA, USA: SAGE Publications, 2018.

[15] B. A. Kitchenham and S. L. Pfleeger, "Personal opinion surveys," in Guide to Advanced Empirical Software Engineering, F. Shull, J. Singer, and D. I. K. Sjøberg, Eds. London, U.K.: Springer, 2008, pp. 63–92, doi: 10.1007/978-1-84800-044-5_3.

[16] A. Bryman, Social Research Methods, 5th ed. Oxford, U.K.: Oxford University Press, 2016.

[17] D. A. Norman, The Design of Everyday Things, rev. and expanded ed. New York, NY, USA: Basic Books, 2013.

[18] M. Fowler, UML Distilled: A Brief Guide to the Standard Object Modeling Language, 3rd ed. Boston, MA, USA: Addison-Wesley, 2004.

[19] G. Booch, J. Rumbaugh, and I. Jacobson, The Unified Modeling Language User Guide, 2nd ed. Boston, MA, USA: Addison-Wesley, 2005.

[20] R. S. A. Kareem, T. Tilford, and S. Stoyanov, "Fine-grained food image classification and recipe extraction using a customized deep neural network and NLP," Computers in Biology and Medicine, vol. 175, Art. no. 108528, Jun. 2024, doi: 10.1016/j.compbiomed.2024.108528.

[21] K. J. P. Ortiz, P. N. P. Bautista, M. V. D. Dimailig, and A. C. D. Llamzon, "Recipe recommendation system using IoT-based food inventory management of perishables for household food waste reduction," Chemical Engineering Transactions, vol. 106, pp. 361–366, 2023, doi: 10.3303/CET23106061.

[22] E. Yourdon, Modern Structured Analysis. Englewood Cliffs, NJ, USA: Prentice Hall, 1989.

[23] K. E. Kendall and J. E. Kendall, Systems Analysis and Design, 8th ed. Upper Saddle River, NJ, USA: Pearson Education, 2011.

[24] S. Few, Information Dashboard Design: Displaying Data for At-a-Glance Monitoring, 2nd ed. Burlingame, CA, USA: Analytics Press, 2013.

[25] J. Preece, Y. Rogers, and H. Sharp, Interaction Design: Beyond Human–Computer Interaction, 4th ed. Chichester, U.K.: Wiley, 2015.

[26] J. Parfitt, M. Barthel, and S. Macnaughton, "Food waste within food supply chains: Quantification and potential for change to 2050," Philosophical Transactions of the Royal Society B: Biological Sciences, vol. 365, no. 1554, pp. 3065–3081, Sep. 2010, doi: 10.1098/rstb.2010.0126.

[27] J. Nielsen, Usability Engineering. San Francisco, CA, USA: Morgan Kaufmann, 1993.