International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

A HYBRID K-MEANS AND MACHINE LEARNING APPROACH FOR HEART DISEASE PREDICTION: A COMPARATIVE STUDY ON BINARY AND TERTIARY RISK CLASSIFICATION

Dr. M. Molavi-Arabshahi *[1], and Hasan Mahdi Alqssari [2]

[1] School of mathematics and computer science, Iran University of science and technology, Tehran, Iran; mlavi@iust.ac.ir

[2] School of mathematics and computer science, Iran University of science and technology, Tehran, Iran; saedhasan73@yahoo.com

* Correspondence: mlavi@iust.ac.ir

Abstract

Heart disease is one of the most dangerous and deadly diseases in the world, due to the large number of deaths caused by this disease annually. This paper proposes a hybrid approach that combines K-means with nine machine learning techniques for predicting heart disease and forecasting future heart disease occurrences. Two datasets, 1,026 records each, were generated: a binary infected-uninfected dataset and a three-class (infected/uninfected/likely infected) dataset, both labeled according to the WHO clinical criteria. The feature space was transformed, the data were clustered using the K-means algorithm, and classification was performed. The trained machine learning algorithms, utilizing the clustered data, included a Decision Tree, a random forest, XGBoost, and Gradient Boosting, with the aid of clinical parameters such as age, blood pressure, and cholesterol level. The Decision Tree achieved the highest accuracy of 89.27% on the three-class prediction problem, outperforming Gradient Boosting and Random Forest. The hybrid K-means and machine learning algorithm exhibits good predictive performance and potential clinical value for identifying patients at risk of future heart disease, facilitating early intervention.

Keywords: Heart Disease Prediction; K-means Clustering; Decision Tree; Supervised Learning; Risk Classification; Machine Learning.

1. Introduction

Heart disease has been the abiding killer of people on earth, as over 17 million deaths are caused by the disease every year, according to the World Health Organization [1]. Cardiovascular diseases represent a serious economic and social burden regardless of the progress in medicine. The risk of heart disease should therefore be predicted correctly and at an early age so that timely intervention can be carried out and preventable deaths can be avoided.

Artificial intelligence (AI) and machine learning (ML) have become innovative solutions in various fields such as the healthcare sector [2,3], agriculture [4], plant diseases [5,6], geography [7,8], and more, providing data-driven diagnostic assistance and decision-making op-

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

portunities. ML algorithms can be used to identify subtle patterns that cannot be perceived by the human eye by analyzing complex clinical measurements such as blood pressure, cholesterol, age, and ECG signals [9]. These methods have been extensively utilized to diagnose heart diseases using models such as Decision Trees, Random Forests, Logistic Regression, and Support Vector Machines, enabling early risk prediction and improved clinical outcomes.

It has been demonstrated that supervised machine learning (ML) models can effectively classify heart diseases and have been employed successfully in numerous studies. For instance, Anuradha et al. [10] reported high accuracy using conventional classifiers on the UCI repository. However, the majority of prior studies have been confined to binary classification—distinguishing only between diseased and healthy individuals—while neglecting patients who may develop the disease in the future [11]. Ahmed et al. [12] proposed a hybrid K-means approach integrated with a set of ML models to predict future diabetes onset. Such future susceptibility remains a critical challenge, as it involves modeling intermediate levels of risk that are typically absent in traditional datasets.

This paper proposes a hybrid predictive framework that combines unsupervised K-means clustering with supervised ML models to improve risk prediction for heart disease. The main contributions of this study are as follows:

• To create a new risk class based on clinical thresholds (e.g., high blood pressure and cholesterol levels), enhancing clustering analysis through the introduction of a novel grouping strategy.

• To perform a comparative analysis of nine ML classifiers—including Decision Tree, Random Forest, XGBoost, and Gradient Boosting—on binary (two-class) and tertiary (three-class) prediction problems.

• To evaluate a hybrid architecture integrating clustering-based preprocessing with supervised classification for enhanced prediction of future heart disease risk.

The remainder of this paper is organized as follows: Section 2 reviews the related literature on ML-based heart disease prediction. Section 3 presents the proposed methodology, including dataset description, preprocessing, clustering, and model training. Section 4 reports and discusses the experimental results. Finally, Section 5 concludes the paper and outlines future research directions.

2. Related Work

3.1. Studies on Binary Classification

Devansh Shah et al. [13] presented a study in October 2020 that examined multiple supervised learning algorithms, including Random Forest, Decision Tree, K-Nearest Neighbor (KNN), and Naive Bayes, to classify heart diseases using the Cleveland Clinic dataset from the UCI repository. The dataset consisted of 303 cases and 14 clinical features. Their results demonstrated that the K-Nearest Neighbor algorithm achieved the highest accuracy.

Lakshmi C.N. et al. [14] employed supervised ML models such as KNN, Decision Tree, Support Vector Machine (SVM), Naive Bayes, and Random Forest on the same Cleveland

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

dataset (303 records, 14 features). Their results revealed that different algorithms predicted heart disease with varying levels of severity depending on the selected features.

Amna Kanwal et al. [15] proposed a model combining several ML techniques to accurately predict heart disease. Feature selection methods, including Relief and LASSO, were applied to identify the most relevant attributes, and classifiers such as Decision Tree, Naive Bayes, KNN, SVM, and Logistic Regression were compared. The Decision Tree achieved the highest accuracy of 85.21%.

Rachana Sani and H.S. Guruprasad [16] analyzed machine learning algorithms, including Decision Tree, Logistic Regression, Random Forest, and Neural Networks, based on accuracy, precision, and recall. Their study aimed to determine the most suitable algorithm for each metric and proposed a machine learning methodology derived from the specified dataset variables.

Guarneros-Nolasco et al. [17] applied ten ML algorithms to four public cardiovascular datasets (Cleveland, Framingham, Faisalabad, South African Heart), using 70/30 train–test splits and 10-fold cross-validation. They found that Decision Tree, Logistic Regression, and SVM achieved the best accuracies (80–83%) in binary heart disease diagnosis.

3.2. Studies Utilizing Ensemble and Advanced Methods

Priyanka Gupta and Dambarudhar Seth [18] applied Multilayer Perceptron (MLP) and various ML classifiers to two datasets (UCI Cardiology and Framingham). After hyperparameter tuning and feature importance analysis, Random Forest showed superior accuracy of 97.13% for the Framingham dataset.

Md. Julker Nayeem et al. [19] applied Naive Bayes, KNN, and Random Forest classifiers to predict heart disease. They enhanced model performance by removing unnecessary features and handling missing values using mean imputation. Random Forest achieved the highest classification accuracy (95.63%).

Suman Halder [20] developed a model to predict heart failure probability using clinical attributes such as age, gender, cholesterol, thalach, and ejection fraction. Among Naive Bayes, Random Forest, Logistic Regression, and SVM, Random Forest achieved the best performance with 98.05% accuracy, 100% sensitivity, and 96.36% specificity.

Alsubai et al. [21] proposed a quantum-enhanced heart failure detection model using the UCI Cleveland dataset. Their instance-based Quantum Machine Learning (QML) and Quantum Deep Learning (QDL) frameworks achieved 83.6% and 98% accuracy, respectively, surpassing traditional ensemble baselines.

Table 1 form highlights the progression of research and clarifies the gap that motivates the present study—specifically the lack of tertiary (three-class) risk prediction and the limited use of unsupervised clustering in prior work.

Table 1. Summary of Representative Studies on Heart Disease Prediction Using Machine Learning

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Study	Dataset Used	Algorithms Used	Best Accuracy (%)	Classification Type	Key Limitation
Shah et al. [13]	Cleveland (303 records, 14 features)	RF, DT, KNN, NB	**KNN ≈ highest**	Binary	Limited to small dataset; no future-risk prediction
Lakshmi C.N. et al. [14]	Cleveland (303, 14 features)	KNN, DT, SVM, NB, RF	Varying across features	Binary	Model performance heavily dependent on selected features
Kanwal et al. [15]	Heart disease dataset	DT, NB, KNN, SVM, LR + feature selection (Relief, LASSO)	**DT = 85.21%**	Binary	Did not consider multi-class or risk-level prediction
Sani & Guruprasad [16]	Heart failure dataset	DT, LR, RF, NN	Metric-dependent (precision/recall varies)	Binary	Focuses on metric comparison; no risk stratification
Guarneros-Nolasco et al. [17]	Four datasets: Cleveland, Framingham, Faisalabad, South African Heart	10 ML algorithms	**80–83%** (DT, LR, SVM)	Binary	No unsupervised learning; only retrospective diagnosis
Gupta & Seth [18]	UCI Cardiology + Framingham	MLP, RF, KNN, SVM, NB	**RF = 97.13%**	Binary	High accuracy but no intermediate risk class

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Nayeem et al. [19]	Heart disease dataset	NB, KNN, RF	**RF = 95.63%**	Binary	Uses basic imputation; no clustering or risk labeling
Suman Halder [20]	Heart failure clinical dataset	NB, RF, LR, SVM	**RF = 98.05%**, Sensitivity 100%, Specificity 96.36%	Binary	Tailored to heart failure only; lacks multiclass modeling
Alsubai et al. [21]	Cleveland dataset	QML, QDL	**QDL = 98%**	Binary	Quantum models still experimental; no 3-class prediction

3.3. Gaps and Motivation for Our Work

Although the binary and ensemble models achieved remarkable accuracies [13,14,15,16,17,18,19,20,21], most focus solely on retrospective diagnosis without addressing future risk or intermediate disease stages. Few studies have integrated unsupervised approaches to reveal latent groupings valuable for preventive care. To address this gap, our work proposes a hybrid K-means with ML framework capable of distinguishing three risk levels—healthy, likely infected, and infected—thereby extending traditional diagnostic models toward predictive risk stratification and early-warning cardiology.

3. Methodology

The present work utilizes a comprehensive methodology combining K-means clustering with nine different machine-learning techniques to forecast future heart disease instances accurately. The methodology is specifically developed to improve the flexibility and accuracy of prediction models by using both unsupervised and supervised learning approaches. To enrich the methodology section and provide a high-level overview of the research novelty, Figure 1 presents a conceptual framework illustrating the flow of patient clinical data through the proposed hybrid system. The diagram summarizes the transitions from data collection, preprocessing, and K-Means clustering to feature augmentation and final supervised learning classification. This figure helps readers quickly understand how the hybrid methodology extends beyond traditional binary prediction and introduces a three-level risk stratification mechanism.

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

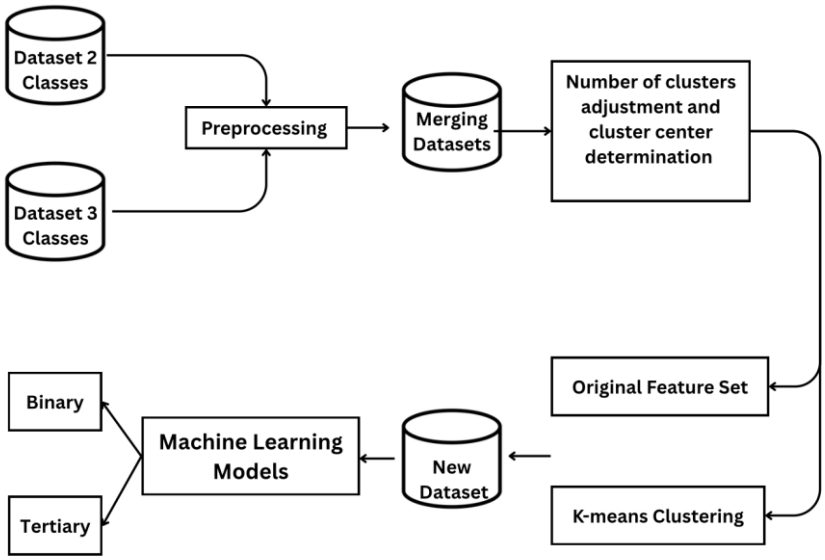

Figure 1. Conceptual Framework of the Proposed Hybrid Model.

3.1. Dataset Description

3.1.1. First Dataset (Two-Label Classification)

The Heart Disease Dataset is an extensive compilation of medical and demographic data used to predict, detect, and diagnose heart disease in individuals. This dataset is widely employed in data analysis and machine learning research to build predictive models that support the early diagnosis of cardiovascular diseases. It includes a comprehensive range of features commonly associated with cardiovascular health, forming a robust foundation for both statistical exploration and predictive modeling.

The dataset contains demographic information such as age and sex, as well as key clinical measurements and health indices. For instance, age is a continuous variable representing the patient's age in years, while sex is a binary variable (1 indicates male, 0 indicates female). Core clinical parameters include resting blood pressure, serum cholesterol concentration, and maximum heart rate achieved during exercise — all of which provide valuable insights into cardiovascular condition. Additionally, the dataset includes qualitative (categorical) variables such as chest pain type and thalassemia status, which represent different chest pain categories and hereditary blood conditions, respectively.

The dataset comprises a total of 1026 patient records, of which 499 correspond to individuals without heart disease and 527 to those diagnosed with heart disease. Thus, the target variable is balanced, ensuring unbiased learning and evaluation across both classes. The distribution of the two classes is illustrated in Figure 2.

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

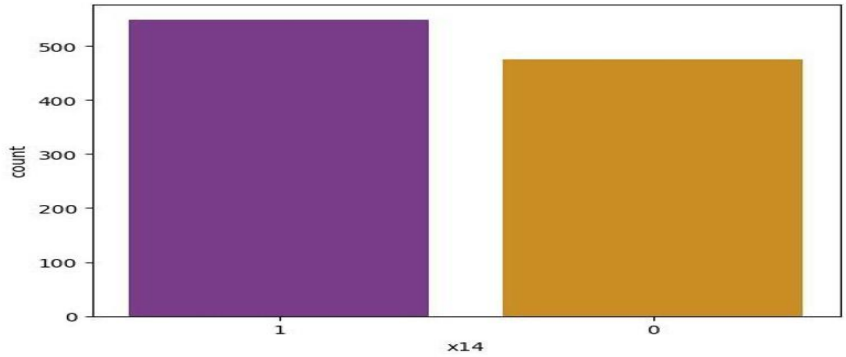

Figure 2. Number of classes for first dataset.

3.1.2. Second Dataset (Three-Label Classification)

The second dataset used in this study was obtained from the Kaggle machine learning repository and corresponds to the well-known Cleveland Heart Disease dataset, which is widely employed for predictive cardiovascular diagnosis using supervised ML algorithms. It comprises a total of 1026 patient records and includes multiple demographic and clinical attributes. Most prior studies have utilized a subset of fourteen features for prediction, including demographic parameters (age, gender) and vital health indicators such as resting blood pressure, serum cholesterol, fasting blood sugar, and other relevant cardiovascular metrics.

With guidance from reports published by the World Health Organization [22,23], related scientific literature [24,25,26], and consultations with specialized cardiologists at the Cardiac Surgery and Catheterization Hospital in Diwaniyah, the dataset was categorized into three classes:

(1) not infected (coded as 0),

(2) infected (coded as 1), and

(3) likely to be infected in the future (coded as 2).

This three-level labeling was based on key clinical thresholds, including systolic blood pressure greater than 140 mmHg, total cholesterol exceeding 200 mg/dL, and additional medical indicators of cardiovascular stress. The resulting class distribution is visualized in Figure 3.

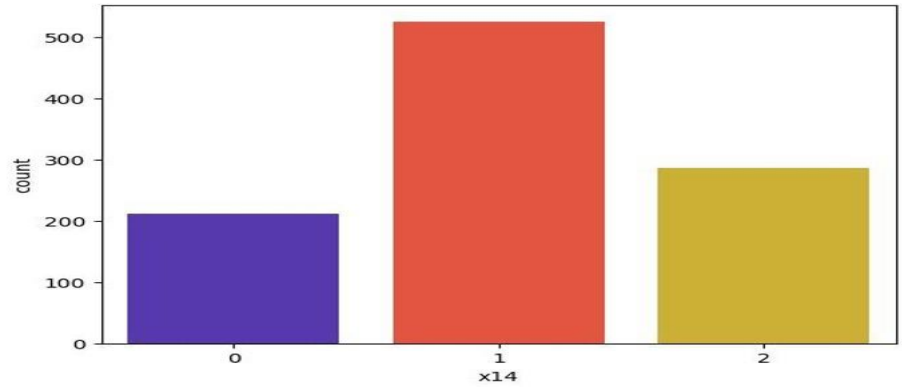

Figure 3. Number of classes for second datasets.

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

3.2. Preprocessing

Preprocessing is a crucial stage in preparing the Heart Disease Dataset for efficient analysis and modeling. Its primary objective is to cleanse and transform the raw data into a suitable format for machine learning algorithms. This stage ensures the reliability, consistency, and interpretability of the dataset through several essential steps:

• Initial Data Inspection:
This step involves loading and exploring the dataset, examining variable types and ranges, and identifying missing or abnormal values that require correction. Missing values can be handled through imputation methods such as replacing them with the mean, median, or using more sophisticated approaches like the K-Nearest Neighbors imputer.

• Feature Encoding and Transformation:
Since many ML algorithms require numerical input, categorical features must be converted into numerical form. For example, the categorical feature ChestPainType can be one-hot encoded, while binary variables such as Sex and FastingBloodSugar can be verified for accuracy and consistency.

• Feature Scaling:
Features with large numeric ranges (e.g., MaxHeartRate, RestingBloodPressure) are scaled to ensure equal influence on model performance. Standardization (zero mean, unit variance) or normalization (range 0–1) helps prevent any single feature from dominating the learning process.

• Managing Outliers:
Outliers can distort model performance and must be addressed through removal, transformation (e.g., logarithmic or square root), or capping. For instance, abnormally high or low SerumCholesterol readings are validated to distinguish genuine cases from data entry errors.

• Dataset Partitioning:
To evaluate model performance objectively, the dataset is divided into training and testing subsets—typically 80% for training and 20% for testing—ensuring generalization on unseen data.

• Feature Selection:
This step enhances model efficiency by retaining only the most influential attributes using correlation analysis or feature-importance metrics, thereby reducing dimensionality and overfitting.

Overall, preprocessing transforms unstructured data into a clean, logical, and consistent form suitable for accurate heart disease prediction models.

3.3. K-Means Clustering

K-Means is a partition-based clustering algorithm that iteratively assigns data points to clusters by minimizing intra-cluster variance [27]. It is widely used in exploratory data analysis due to

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

its simplicity, scalability, and computational efficiency. In the context of heart disease prediction, K-Means groups patients based on clinical similarities, enabling the identification of latent patterns and risk clusters.

The goal of K-Means is to form clusters where points within each group are highly similar, while points in different clusters are dissimilar. Each cluster center (C) represents the mean of all data points assigned to that group, and the Euclidean distance between a data point (P) and its cluster center is calculated as:

$$d(C, P) = \|C - P\| \quad (1)$$

In this study, the dataset was divided into three clusters corresponding to the three health states. Among the 1026 records, 212 patients were assigned to Cluster 1, 526 to Cluster 0, and 287 to Cluster 2. Effective feature selection significantly improves clustering performance by emphasizing the most informative variables.

Thus, K-Means serves not only as a clustering tool but also as a preprocessing stage to simplify data and provide structured inputs for subsequent supervised models.

3.4. Machine Learning Models

Machine learning (ML), a subfield of artificial intelligence (AI), enables computers to learn patterns from data and make predictions without explicit programming. Recent advances in computational power have allowed ML algorithms to be applied in healthcare for disease prediction, diagnosis, and personalized treatment [28,29]. In this study, nine supervised ML models were implemented for the predictive diagnosis of heart disease, as outlined below.

• Random Forest (RF):
An ensemble learning method that builds multiple decision trees and combines their outputs to improve predictive accuracy and reduce overfitting [30]. RF is robust to noise and provides insights into feature importance.

• Decision Tree (DT):
A tree-based model that recursively splits data based on feature thresholds [31]. It is interpretable and well-suited for identifying nonlinear relationships between clinical indicators and disease outcomes.

• XGBoost:
An optimized gradient-boosting framework that integrates regularization for better generalization [32]. It delivers strong performance with efficient computation and is highly effective for structured medical data.

• Gradient Boosting (GB):
A sequential ensemble method that adds weak learners to correct errors of previous models [33]. It improves prediction accuracy through iterative refinement.

• K-Nearest Neighbors (KNN):
A simple, non-parametric algorithm that classifies a sample based on the majority label of its k nearest neighbors [34]. It is intuitive and effective for small to medium-sized datasets.

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

• Support Vector Machine (SVM, RBF Kernel):
A robust classifier that separates classes by maximizing the margin between them [35]. The radial basis function (RBF) kernel enables nonlinear decision boundaries and requires careful tuning of C and γ.

• Logistic Regression (LR):
A statistical model commonly used for binary classification [36]. It estimates the probability of heart disease occurrence and provides interpretable coefficients for clinical features.

• Multinomial Naive Bayes (MNB):
A probabilistic classifier based on Bayes' theorem [37]. It assumes feature independence and performs efficiently with categorical health indicators.

• Support Vector Machine (Linear Kernel):
A linear version of SVM that identifies the optimal hyperplane for classification [38]. It is computationally efficient and well-suited for linearly separable datasets.

Each of these models was trained and evaluated on both binary (two-class) and ternary (three-class) versions of the dataset to assess their predictive performance.

4. Results and Discussion

4.1. K-Means Clustering Results

Based on the proposed methodology, nine machine learning techniques were integrated with the K-Means clustering algorithm, including Decision Tree, K-Nearest Neighbors (KNN), Random Forest, Support Vector Machine (linear and RBF kernels), XGBoost, Gradient Boosting, Logistic Regression, and Multinomial Naive Bayes. The K-Means algorithm was employed to group the patients into clusters prior to supervised classification in order to enhance data organization and predictive accuracy.

To determine the optimal number of clusters, the Elbow Method was applied. This method plots the relationship between the number of clusters (k) on the x-axis and the within-cluster variance (inertia or sum of squared distances) on the y-axis. At the beginning, when k is small, the variance is large because the data points are widely dispersed within a few clusters. As k increases, the variance decreases sharply since the data become more finely partitioned, reducing intra-cluster dispersion.

The elbow point corresponds to the stage where the rate of variance reduction slows significantly—indicating that adding more clusters offers minimal improvement. In this study, the elbow was observed at k = 3, as shown in Figure 4. Therefore, the optimal number of clusters for this dataset is three, corresponding to the following patient groups:

1. Uninfected: Individuals with normal clinical readings and no signs of heart disease.
2. Infected: Individuals diagnosed with heart disease based on medical indicators.
3. Likely to be infected: Individuals at elevated risk of developing heart disease in the future.

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

This tripartite clustering aligns with the clinical understanding of cardiovascular risk stratification and provides a structured foundation for subsequent supervised model training.

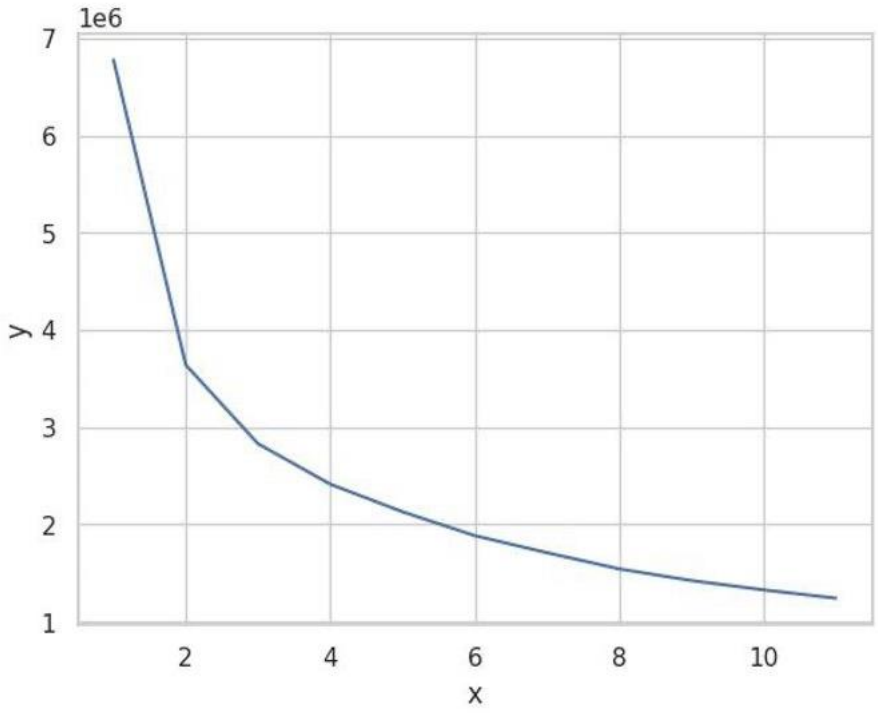

Figure 4. Elbow Method showing optimal 3 for K-Means clustering.

4.2. Machine Learning Results

Nine supervised machine learning algorithms were evaluated on the test set, each represented by a bar in the accuracy chart. The horizontal axis (X) lists the models—KNN, SVM (RBF), Gradient Boosting, XGBoost, Random Forest, Decision Tree, Logistic Regression, SVM (Linear), and Multinomial Naive Bayes—while the vertical axis (Y) shows their corresponding accuracies, ranging between 0.80 and 1.00 (80–100%).

Among all algorithms, the Decision Tree achieved the highest accuracy, approximately 0.89 (89%), demonstrating the strongest classification performance, as shown in Figure 5. Random Forest, Gradient Boosting, and XGBoost followed closely with accuracies between 0.87 and 0.88, confirming their reliability in predicting heart disease.

Conversely, KNN produced the lowest accuracy (below 0.85), suggesting that it may not be optimal for this dataset without parameter optimization. The remaining algorithms—SVM (RBF), SVM (Linear), Logistic Regression, and Multinomial Naive Bayes—performed moderately, all with accuracies below 0.85.

Overall, the Decision Tree outperformed all other models, effectively partitioning the data and yielding the highest predictive accuracy. Ensemble and boosting-based algorithms (Random Forest, XGBoost, and Gradient Boosting) also achieved competitive results, while simpler models such as SVM and KNN may require further hyperparameter tuning to improve performance.

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

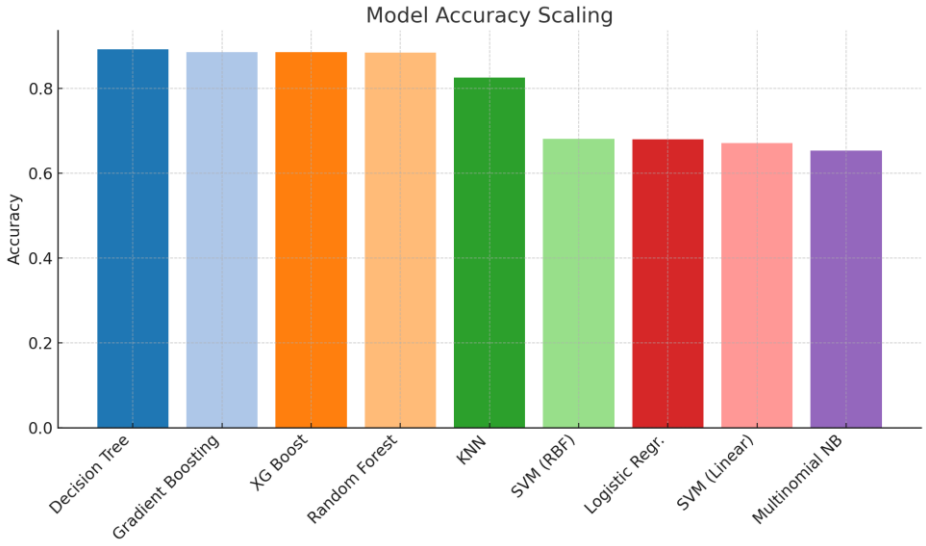

Figure 5. Results of classification models.

4.3. Machine Learning Accuracy and Training Time

Machine learning techniques were applied to evaluate and test the heart disease dataset using nine different algorithms: KNN, SVM (Linear), XGBoost, Naive Bayes, Logistic Regression, Decision Tree, Random Forest, SVM (RBF), and Gradient Boosting. The performance of each model was measured in terms of accuracy and training time, as summarized in Table 2 and Figure 6.

The results show that the Decision Tree algorithm achieved the highest accuracy of 0.8926 (89.26%) with a training time of only 0.01 seconds, making it both fast and highly effective for this dataset . Gradient Boosting and XGBoost achieved similar accuracies of 0.8861 (88.61%), though their training times were longer—0.74 and 0.52 seconds, respectively. Random Forest followed closely with an accuracy of 0.8845 (88.45%) and a training time of 0.20 seconds.

KNN achieved moderate performance, with an accuracy of 0.8260 (82.60%) and a rapid training time of 0.01 seconds, indicating efficiency in computation. SVM (RBF) and Logistic Regression recorded accuracies of 0.6813 (68.13%) and 0.6796 (67.96%), respectively, with training times of 0.24 and 0.08 seconds. SVM (Linear) performed slightly worse, achieving 0.6715 (67.15%) accuracy in 0.52 seconds. The Multinomial Naive Bayes classifier had the lowest performance, with an accuracy of 0.6536 (65.36%) and a training time of 0.01 seconds.

In summary, Decision Tree, Gradient Boosting, and XGBoost were the most accurate models, with the Decision Tree standing out as the best balance between accuracy and speed. Algorithms such as KNN and Naive Bayes trained extremely fast but at the cost of lower precision, while SVM-based and logistic regression models underperformed, achieving less than 70% accuracy. If the objective is to maximize predictive accuracy, the Decision Tree, Gradient Boosting, and XGBoost algorithms are the most suitable choices. However, if computational efficiency and rapid model training are prioritized, Decision Tree and KNN offer the best compromise between speed and accuracy.

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Table 2. Performance comparison of machine learning algorithms on the three-class classification task.

Model	Accuracy scaling	Training time (sec)
Decision Tree	0.8926	0.01
Gradient Boosting	0.8861	0.74
XG Boost	0.8861	0.52
Random Forest	0.8845	0.2
KNN	0.8260	0.01
SVM (RBF)	0.6813	0.24
Logistic Regr.	0.6796	0.08
SVM (Linear)	0.6715	0.52
Multinomial NB	0.6536	0.02

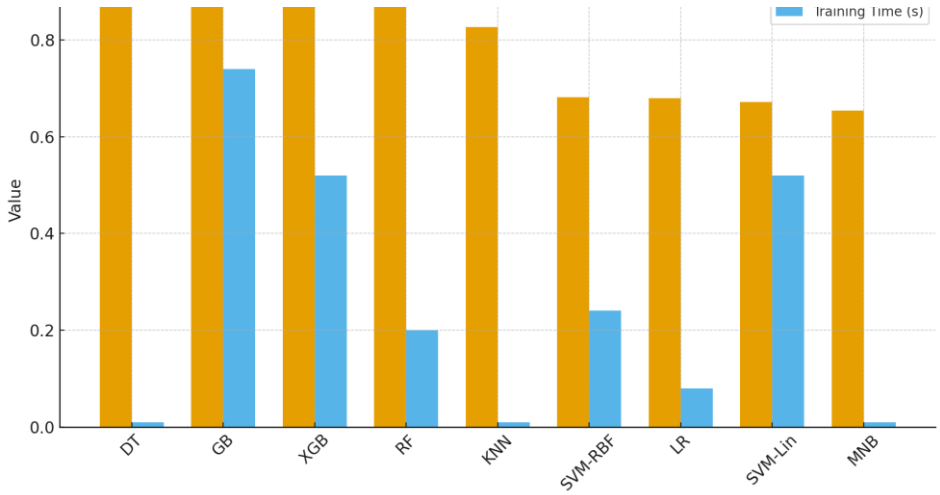

Figure 6. Accuracy and Training Time Comparison Across Models.

4.4. Discussion

A hybrid framework combining K-Means clustering with machine learning techniques was implemented to enhance the classification of heart disease cases. Nine algorithms were evaluated, including KNN, SVM (Linear), SVM (RBF), Multinomial Naive Bayes, Gradient Boosting, Logistic Regression, Decision Tree, XGBoost, and Random Forest. The integration of K-Means improved data organization and revealed latent patient groupings prior to supervised classification.

The experimental results demonstrated that the Decision Tree algorithm achieved the highest accuracy (0.89), outperforming all other models in the three-class classification task. Ensemble and boosting-based models such as Gradient Boosting, XGBoost, and Random Forest also

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

achieved competitive accuracies, confirming their robustness for structured medical datasets. These findings suggest that tree-based learning methods are particularly effective for heart disease prediction due to their capability to handle nonlinear interactions and mixed data types.

As illustrated in Figure 7, the confusion matrix of the Decision Tree model shows strong diagonal dominance, with 106, 110, and 59 correct predictions for the three classes, indicating reliable classification performance and minimal misclassification.

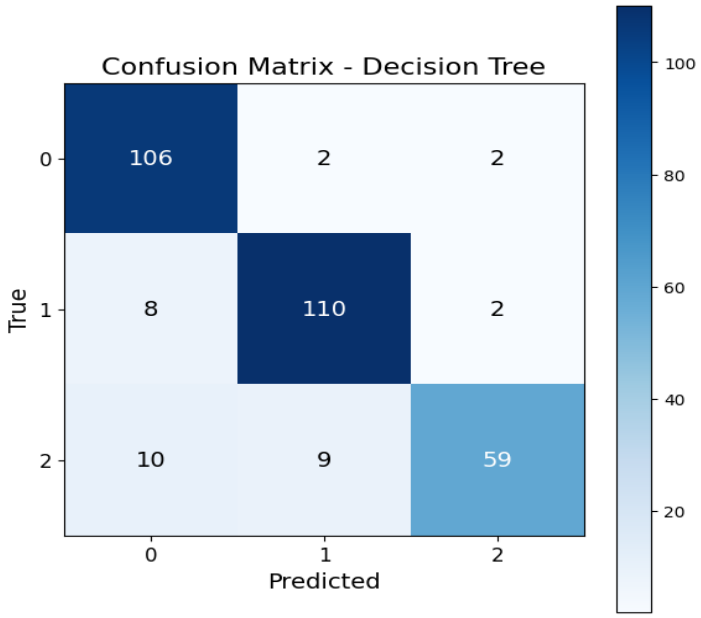

Figure 7. Confusion Matrix for Decision Tree.

The class-wise evaluation shown in Figure 8 provides a clearer understanding of how the Decision Tree model performs across the three target categories. For the not infected class, the model achieved a precision of 0.855 and a very high recall of 0.964, indicating that it correctly identified almost all true negative cases, with minimal false classifications. The corresponding F1-score of 0.906 confirms a balanced and reliable performance, which is consistent with the confusion matrix where 106 out of 110 samples were correctly predicted.

Similarly, the infected class demonstrated strong performance, with precision and recall values of 0.909 and 0.917, respectively, resulting in the highest F1-score 0.913. This indicates that the model is highly capable of detecting infected cases while maintaining a low false-positive rate, an important factor in medical diagnosis. The confusion matrix supports this result, showing 110 correct classifications with only a few misclassifications.

In contrast, the likely class showed a different pattern. Although it achieved the highest precision 0.937, meaning predictions for this class are generally correct, the recall dropped to 0.756, indicating that the model missed a notable number of actual “likely” cases. The F1-score of 0.836 reflects this imbalance, and the confusion matrix confirms that 19 samples were misclassified into other classes.

Overall, these results show that the Decision Tree model performs very well on clearly defined classes not infected and infected, while the intermediate likely category remains more chal-

Received: August 11, 2025

International Journal of Applied Mathematics
Volume 38 No. 12s, 2025
ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

lenging due to overlapping characteristics. This highlights the importance of class-level analysis in multi-class medical prediction tasks.

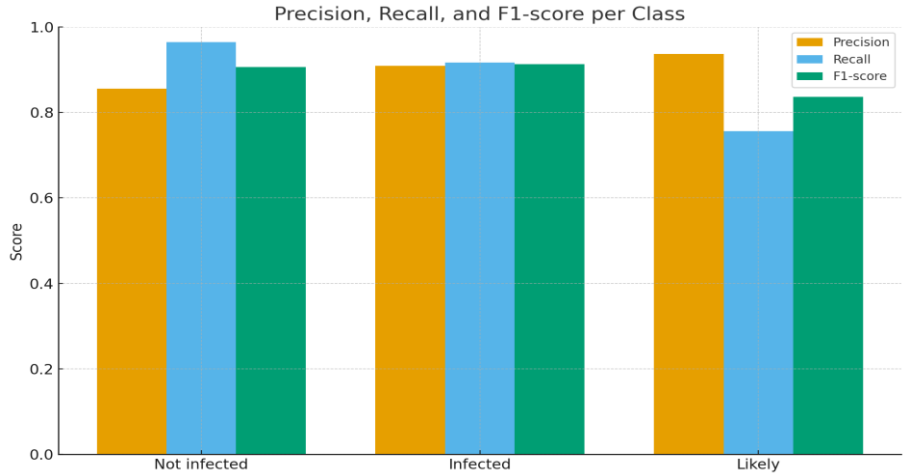

Figure 8. Class-wise Precision, Recall, and F1-score of the Decision Tree model.

5. Conclusions and Recommendations

This study applied a hybrid K-Means and machine learning framework to predict heart disease across three risk categories. Nine algorithms were evaluated in terms of accuracy and training efficiency, leading to the following conclusions and recommendations.

The Decision Tree algorithm demonstrated the best overall performance, achieving an accuracy of 0.8927 (89.27%) with the shortest training time of 0.01 seconds. This result indicates that the Decision Tree model was the most effective in segmenting and classifying the dataset, making it the optimal choice for this type of medical prediction task.

Gradient Boosting and XGBoost also performed well, both achieving an accuracy of 0.8862 (88.62%). Although their training times were slightly longer, these algorithms remain strong candidates when higher accuracy is prioritized over computational speed. Random Forest achieved an accuracy of 0.8845 (88.45%), confirming its robustness for complex and nonlinear data.

In contrast, KNN and Multinomial Naive Bayes recorded the lowest accuracies, 0.8260 (82.60%) and 0.6536 (65.36%), respectively, suggesting that they are less suitable for this dataset in their current form. However, KNN was among the fastest to train (0.01 seconds), highlighting its efficiency for rapid model prototyping.

For future applications, the Decision Tree is recommended as the primary algorithm due to its high accuracy and low computational cost. When maximum precision is required, Gradient Boosting or XGBoost can be employed despite their longer training times. Further improvements can be achieved through hyperparameter tuning of XGBoost, Random Forest, and KNN, as well as optimizing data preprocessing and feature scaling to reduce computation time while maintaining predictive performance.

Received: August 11, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

Repeated evaluation on new or partitioned datasets is encouraged to validate model stability and generalization. Continuous refinement and retraining will ensure that these predictive models remain reliable tools for early diagnosis and risk assessment of heart disease.

.

Author Contributions: Conceptualization, H.M. Alqssari and M. Molavi-Arabshahi; methodology, H.M. Alqssari; software, H.M. Alqssari; validation, H.M. Alqssari and M. Molavi-Arabshahi; formal analysis, H.M. Alqssari; investigation, H.M. Alqssari; resources, M. Molavi-Arabshahi; data curation, H.M. Alqssari; writing—original draft preparation, H.M. Alqssari; writing—review and editing, M. Molavi-Arabshahi; visualization, H.M. Alqssari; supervision, M. Molavi-Arabshahi; project administration, M. Molavi-Arabshahi; funding acquisition, M. Molavi-Arabshahi. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data used in this study are publicly available from the UCI Machine Learning Repository and the Kaggle Heart Disease Dataset. No new data were created or analyzed in this study.

Acknowledgments: The authors would like to thank the School of Mathematics and Computer Science, Iran University of Science and Technology, Tehran, Iran, for their support and valuable academic environment during this research.

Conflicts of Interest: The authors declare no conflicts of interest. funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. World Health Organization (WHO). Cardiovascular Diseases. Available online: https://www.who.int/health-topics/cardiovascular-diseases (accessed on 8 November 2025).
2. Vu, M.T.; Kim, S.H.; Thanh, H.L.N.N.; Roohi, M.; Nguyen, T.H. Trust Region Policy Learning for Adaptive Drug Infusion with Communication Networks in Hypertensive Patients. Mathematics 2025, 13, 136. https://doi.org/10.3390/math13010136.
3. Ashour, L.S.; et al. Non-Overlapping Patch-Based Pre-Trained CNN for Breast Cancer Classification. Iraqi Journal for Computer Science and Mathematics 2025, 6(2). https://doi.org/10.52866/2788-7421.1271.
4. Jawad, B.H.; Alwesabi, O.A.; Abdullah, N.; Mohammed, A.A. Hybrid Artificial Neural Network Activation Function for Agricultural Irrigation. IEEE Access 2025, 13, 93302–93322. https://doi.org/10.1109/ACCESS.2025.3573678.
5. Sumari, P.; et al. Optimizing Lemongrass Disease Detection Using NAS, CNN, and Transfer Learning. Statistics, Optimization & Information Computing 2025, 14(4), 2022–2040. https://doi.org/10.19139/soic-2310-5070-2507.

Received: August 11, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

6. Mohammed, A.A.; Alqaseer, H.; Sumari, P.; Zaid, M.M.A. Durian Fruit Disease Detection Using CNN. In Lecture Notes in Networks and Systems; 2025; pp. 97–116. https://doi.org/10.1007/978-981-96-5220-4_9.
7. Zaid, M.M.A.; Mohammed, A.A.; Sumari, P. Road Feature Classification Using CNN. International Journal of Computing and Digital Systems 2025, 17(1), 1–12. https://doi.org/10.12785/ijcds/1571031764.
8. Zaid, M.M.A.; Mohammed, A.A.; Sumari, P. Geographical Land Structure Classification Using CNN. Journal of Computer Science 2024, 20(12), 1580–1592. https://doi.org/10.3844/jcssp.2024.1580.1592.
9. Santhanam, P.; Ahima, R.S. Machine Learning and Blood Pressure. The Journal of Clinical Hypertension 2019, 21(11), 1735–1737. https://doi.org/10.1111/jch.13700.
10. Anuradha, P.; David, V.K. Feature Selection via BoostARoota and CatBoost for Heart Disease. International Journal of Computer Theory and Engineering 2022, 14(1), 2022.
11. Kumar, V.; et al. Addressing Class-Imbalanced Clinical Datasets. Healthcare 2022, 10(7). https://doi.org/10.3390/healthcare100711.
12. Mohammed, A.A.; Sumari, P.; Attabi, K. Hybrid K-Means + PCA for Diabetes Prediction. International Journal of Computing and Digital Systems 2024, 15(1), 1719–1728. https://doi.org/10.12785/ijcds/1501121.
13. Shah, D.; Patel, S.; Bharti, S.K. Heart Disease Prediction Using ML. SN Computer Science 2020, 1, 345. https://doi.org/10.1007/s42979-020-00365-y.
14. Srivastava, J.; et al. Automated Heart Disease Prediction. In IDICAIEI Conference Proceedings; IEEE, 2023; pp. 1–6. https://doi.org/10.1109/IDICAIEI58380.2023.10407008.
15. Kanwal, A.; et al. Detection of Heart Disease Using Supervised ML. VFAST Transactions on Software Engineering 2022, 10(3), 58–70. https://doi.org/10.21015/vtse.v10i3.1106.
16. Sanni, R.R.; Guruprasad, H.S. Performance Metrics for Heart Failure Using ML. Global Transitions Proceedings 2021, 2(2), 233–237. https://doi.org/10.1016/j.gltp.2021.08.028.
17. Guarneros-Nolasco, L.R.; et al. Cardiovascular Disease Prediction Using ML. Mathematics 2021, 9, 2537. https://doi.org/10.3390/math9202537.
18. Gupta, P.; Seth, D. Comparative ML/DL Analysis for Heart Disease. Indonesian Journal of Electrical Engineering and Computer Science 2022, 29(1), 451. https://doi.org/10.11591/ijeecs.v29.i1.pp451-459.
19. Nayeem, M.J.; Rana, S.; Islam, M.R. Heart Disease Prediction Using ML. European Journal of Artificial Intelligence and Machine Learning 2022, 1(3), 22–26. https://doi.org/10.24018/ejai.2022.1.3.13.
20. S.N.P.; K.K.; D.S.; P.P. Impact of AI on Patient Care. Scholars Academic Journal of Pharmacy 2024, 13(2), 67–81. https://doi.org/10.36347/sajp.2024.v13i02.003.

Received: August 11, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

21. Alsubai, S.; et al. Heart Failure Detection Using Quantum ML. Mathematics 2023, 11, 1467. https://doi.org/10.3390/math11061467.
22. Pirillo, A.; Tokgözoğlu, L.; Catapano, A.L. European Lipid Guidelines. Current Atherosclerosis Reports 2024, 26, 133–137. https://doi.org/10.1007/s11883-024-01194-7.
23. Predictors of Unachieved LDL Levels in Ischemic Stroke Patients. Journal of Applied Pharmaceutical Science 2014. https://doi.org/10.7324/JAPS.2014.40417.
24. Davies, T.-L.; et al. Framingham Heart Age vs Real Age. BMJ Open 2013, 3(10), e003245. https://doi.org/10.1136/bmjopen-2013-003245.
25. Visseren, F.L.J.; et al. 2021 ESC Cardiovascular Prevention Guidelines. European Heart Journal 2021, 42(34), 3227–3337. https://doi.org/10.1093/eurheartj/ehab484.
26. Lewington, S.; et al. Blood Pressure & Cholesterol Exposure and CVD Risk. New England Journal of Medicine 2007.
27. Chaudhry, M.; et al. Identifying Patterns Using Unsupervised Clustering. Symmetry 2023, 15(9), 1679. https://doi.org/10.3390/sym15091679.
28. Liu, R.; et al. Platelet Detection with Improved YOLOv3. Cyborg and Bionic Systems 2022, 9780569. https://doi.org/10.34133/2022/9780569.
29. Mohamed, A.A.A.; et al. Colon Disease Diagnosis with CNN and GOA. Diagnostics 2023, 13, 1728. https://doi.org/10.3390/diagnostics13091728.
30. Mienye, I.D.; Sun, Y. Ensemble Learning Survey. IEEE Access 2022, 10, 99129–99149. https://doi.org/10.1109/ACCESS.2022.3191881.
31. Arista, A. DT vs Logistic Regression for COVID-19. Sinkron 2022, 7(1), 59–65. https://doi.org/10.33395/sinkron.v7i1.11310.
32. Huber, F.; et al. XGBoost for Yield Estimation. Computers and Electronics in Agriculture 2022, 202, 107346. https://doi.org/10.1016/j.compag.2022.107346.
33. Mienye, I.D.; Sun, Y. Ensemble Learning Survey (Duplicate). IEEE Access 2022, 10, 99129–99149. https://doi.org/10.1109/ACCESS.2022.3191881.
34. Uddin, S.; et al. KNN Variants for Disease Prediction. Scientific Reports 2022, 12(1), 6256. https://doi.org/10.1038/s41598-022-09997-5.
35. Iqbal, B.M.; et al. Election Sentiment Using SVM. JoSYC 2023, 4(2), 397–404.
36. Sulistianingsih, N.; Switrayana, I.N. SMOTE-Tomek + Logistic Regression. IJEME 2024, 14(3), 22. https://doi.org/10.5815/ijeme.2024.03.02.
37. Odeh, A.H.; et al. Multinomial Naive Bayes for Code Classification. In ACIT 2022 Proceedings; IEEE, 2022. https://doi.org/10.1109/ACIT57182.2022.10003036.
38. Salazar-Rojas, T.; et al. SVM vs Linear Regression for Heavy Metals Assessment. Environmental Pollution 2022, 314, 120227. https://doi.org/10.1016/j.envpol.2022.120227.

Received: August 11, 2025