Advances In Multidisciplinary & Scientific Research
A Multidisciplinary & Interdisciplinary Journal

Vol. 12 No. 3, September , 2026

Optimizing Heart Disease Diagnostics for Community Development: A Hybrid Ensemble Approach for Imbalanced Clinical Datasets

[1]Abiola, O.A., [2]Olaleye, I.O. & [3]Odufuwa, J.E.
[1]ICT and Cybersecurity Department, University of Ibadan, Ibadan, Nigeria
[2]Computer Science Department, University of Ibadan, Ibadan, Nigeria.
[3]Department of Business Administration, University of Nicosia, Nicosia, South Cyprus.
E-mails: [1]oladimejiarowolo@yahoo.co.uk, [2]Isaacolaleye95@gmail.com
[3]Jesufunminire2832000@gmail.com

ABSTRACT

Combating the global burden of Cardiovascular Disease (CVD) mortality and promoting community development, this research proposes a hybrid machine learning framework that combines unsupervised clustering with a Stacked Generalization ensemble. The research utilizes K-Modes Clustering (k=4) to segment 70,000 patient records into risk profiles, which serve as meta-features for an ensemble synthesized by a Logistic Regression meta-learner. Results demonstrate superior stability with 75.19% Precision and an AUC of 0.80, validating the model's clinical relevance for localized health outreach. Ultimately, this scalable architecture provides a robust pathway for improving the reliability of automated cardiac diagnostics within community health settings.

Keywords: Machine Learning, Cardiovascular Disease, Stacked Generalization, K-Modes

Aims Research Journal Reference Format:
Abiola, A.O., Olaleye, I. & Odufuwa J.E. (2026): Optimizing Heart Disease Diagnostics for Community Development: A Hybrid Ensemble Approach for Imbalanced Clinical Datasets. Advances in Multidisciplinary Research Journal.
Vol. 12 No. 3. Pp 23-38. www.isteams.net/aimsjournal. dx.doi.org/10.22624/AIMS/V12N3P3

1. INTRODUCTION

Global Burden of Cardiovascular Diseases Cardiovascular diseases (CVDs) remain the preeminent challenge to global public health, maintaining their status as the leading cause of mortality worldwide despite decades of medical advancement. Recent epidemiological projections indicate a disturbing upward trajectory, with global cardiovascular mortality expected to rise significantly from 2025 to 2050, driven largely by an aging population and the increasing prevalence of metabolic risks such as hypertension and high body mass index [1]. While age-standardized mortality rates have shown some stabilization in high-income regions, the absolute burden is shifting disproportionately towards low- and middle-income countries (LMICs). In these regions, the escalation of CVDs directly undermines community development goals, as rapid urbanization and changing lifestyles fuel a dual burden of disease that threatens to erode recent gains in local quality of life [2]. The demographic profile of CVD patients is evolving; recent data highlights a concerning rise in cardiovascular events among adolescents and young adults. This trend is particularly damaging to community-level social structures, as it impacts individuals during their most productive years, challenging the traditional perception that heart disease is exclusively a geriatric condition.

1.1 The Role of Machine Learning in Clinical Prognosis

The integration of Machine Learning (ML) into clinical prognosis represents a paradigm shift from traditional statistical methods toward data-driven precision medicine. Unlike conventional risk assessment tools such as the Framingham Risk Score, which rely on a limited set of linear variables, modern ML algorithms possess the computational capacity to analyze high-dimensional, non-linear, and multimodal datasets, including electronic health records (EHRs), genomic profiles, and medical imaging [3]. Machine Learning models, particularly Deep Learning and ensemble techniques, significantly outperform standard clinical guidelines in predicting cardiovascular events by detecting subtle, complex patterns that escape human observation [4].

1.2 The Challenge of Class Imbalance in Medical Datasets

Class imbalance represents one of the most pervasive and critical hurdles in the development of reliable medical diagnostic systems, particularly in the context of cardiovascular disease prediction. In typical clinical datasets, the distribution of data is inherently skewed; patients diagnosing positive for a specific pathology (the minority class) are vastly outnumbered by healthy individuals or those with unrelated conditions (the majority class). Recent research highlights that this disparity is not merely a statistical nuisance but a fundamental structural flaw that severely compromises the learning process of artificial intelligence models [5]. When a dataset is heavily dominated by the healthy class, standard algorithms struggle to distinguish the sick instances as significant signals, often treating them instead as statistical noise or outliers that can be safely ignored.

1.3 Limitations of Single Classifiers in High-Risk Diagnostics

While single machine learning classifiers such as Logistic Regression, Decision Trees, and Support Vector Machines (SVM) have served as foundational tools in medical informatics, they exhibit significant structural weaknesses when applied to complex, high-dimensional biological data. A primary limitation identified in recent literature is the inherent instability and high variance of models like Decision Trees, which are prone to overfitting; they frequently memorize noise in the training data rather than learning generalizable disease patterns, leading to poor performance on unseen patient records [6]. Linear models like Logistic Regression operate on the assumption of linear relationships between risk factors, an assumption that rarely holds true in cardiovascular pathology, where variables such as age, cholesterol, and blood pressure interact in highly non-linear and unpredictable ways [7].

The most critical failure of single classifiers in high-risk diagnostics, however, is their inadequate handling of class imbalance, which directly contributes to dangerous rates of False Negatives. When single models are trained on imbalanced heart disease datasets, they invariably bias their predictions toward the majority class (healthy patients) to minimize overall error, effectively sacrificing the minority class (sick patients) [8]. Unlike ensemble methods, which can leverage voting mechanisms to correct individual errors, a single classifier lacks a safety net; if it misinterprets a subtle signal, the error is final. This limitation was underscored by a 2025 systematic review, which found that single classifiers consistently underperformed in sensitivity metrics compared to stacked ensemble architectures, identifying significantly fewer at-risk patients in critical triage scenarios [9].

2. RELATED WORKS

The inherent limitations of single-classifier systems in handling high-dimensional, imbalanced medical data have necessitated a shift toward Ensemble Learning paradigms. Ensemble methods, particularly Stacking and Voting, operate on the wisdom of crowds principle, aggregating the predictions of multiple diverse algorithms to reduce the overall variance and bias that typically plague individual models [10].

Combining the decision boundaries of distinct base learners such as the linear separation of Support Vector Machines with the non-linear branching of Random Forests, ensembles can capture complex, multifaceted disease patterns that a solitary model would miss. Stacked ensemble architectures, which use a meta-learner to weigh the contributions of each base classifier, consistently outperform standalone models in terms of generalization and robustness on unseen clinical data [11]. This architectural diversity acts as a fail-safe; if one model overfits to the majority class or misinterprets a specific feature, the consensus mechanism of the ensemble can correct the error, thereby stabilizing performance across diverse patient demographics. Stacking allows for the inclusion of diverse mathematical perspectives, combining probabilistic, instance-based and tree-based logic which creates a more granular decision surface capable of distinguishing between hard-to-classify false negatives and true positives, a capability that is essential for improving diagnostic sensitivity in life-critical applications [12].

2.2 Techniques for Handling Data Imbalance (Data-Level Approaches)

Data-level techniques remain the primary intervention for mitigating class imbalance, with oversampling being the most widely adopted strategy in cardiovascular research. The core objective of oversampling is to artificially augment the minority class (sick patients) to achieve parity with the majority class (healthy patients). Dominance of the Synthetic Minority Over-sampling Technique (SMOTE) and its adaptive variant, ADASYN, in this domain was highlighted. Unlike simple random oversampling, which merely duplicates existing records and leads to overfitting, SMOTE generates synthetic instances by interpolating between existing minority samples in the feature space.

Random Forest models on heart disease datasets was evaluated and found that applying ADASYN increased minority class recall from 75% to 95%, significantly reducing the rate of missed diagnoses compared to baseline . Recent work has demonstrated that combining SMOTE with Support Vector Machines (SVM) can push accuracy to nearly 96% by clarifying the decision boundary in high-dimensional space [13] .Undersampling techniques aim to balance the dataset by reducing the size of the majority class, either randomly or strategically. Random Undersampling (RUS) is the simplest form, where healthy patient records are randomly discarded until the classes are balanced. While computationally efficient, this method is increasingly criticized in recent literature for removing potentially valuable clinical data, which can lead to model underfitting.

Recognizing the limitations of using oversampling or undersampling in isolation, the current state-of-the-art has shifted toward Hybrid Approaches that combine both strategies to maximize data quality. The most prominent example is SMOTE-ENN (Edited Nearest Neighbors) or SMOTE-Tomek, which first oversamples the minority class to create balance and then applies undersampling to clean the noisy or overlapping data points generated during the process. Hybrid SMOTE-ENN framework combined with Random Forest achieved superior stability and sensitivity compared to SMOTE alone, as the ENN step effectively removed ambiguous synthetic samples that confused the classifier[14]. Comprehensive reviews from 2024 indicated that these hybrid techniques are particularly effective for noisy medical datasets where patient symptoms often overlap, providing a robust solution that preserves the richness of the data while ensuring a clear decision boundary for the algorithm [15].

The practical implementation of CSL has shown promising results across various classifiers, particularly with Cost-Sensitive Support Vector Machines (CS-SVM) and Weighted Random Forests. Wang demonstrated that assigning a cost ratio of 10:1 (penalizing False Negatives ten times more than False Positives) allowed a CS-SVM model to achieve a sensitivity of 92% on an imbalanced heart failure dataset, significantly outperforming a standard SVM which stalled at 78% sensitivity[2] . Recent innovations have integrated evolutionary algorithms to dynamically optimize these cost matrices, rather than relying on manual tuning.

3. METHODOLOGY

The proposed research framework integrates a hybrid unsupervised-supervised learning pipeline designed to enhance diagnostic sensitivity in heart disease prediction. Initially, the raw clinical data undergoes rigorous preprocessing, including outlier removal via quantile thresholding and the engineering of physiological features such as Body Mass Index (BMI). A novel contribution of this study is the application of K-Modes clustering to categorize patients based on high-dimensional categorical variables, creating a new meta-feature that aids in distinguishing complex patient profiles. Subsequently, these enriched features are fed into a heterogeneous ensemble comprising Decision Trees, Random Forests, Multi-Layer Perceptrons (MLP), and XGBoost. The final classification is achieved through a Stacking strategy using Logistic Regression as a meta-learner, a method proven in recent studies to effectively reduce variance and improve recall rates in medical diagnostics.

3.1 Dataset Description and Acquisition

The study utilizes the cardiovascular disease dataset (cardio_train.csv), a widely recognized benchmark in medical informatics research sourced from Kaggle. The raw dataset comprises 70,000 patient records, each originally described by 13 columns, which includes a unique identifier (id), 11 predictive features, and one binary target variable. The feature set is composed of three distinct categories: demographic attributes (Age, Gender, Height, Weight), objective medical examinations (Systolic blood pressure, Diastolic blood pressure), and subjective patient-reported lifestyle variables (Cholesterol, Glucose, Smoking, Alcohol intake, and Physical Activity) . To ensure computational efficiency and relevance, the id column, which contains no predictive value, was excluded from the training process, leaving 12 robust features for analysis.

The target variable, cardio, serves as the ground truth for the binary classification task, where '1' indicates the presence of cardiovascular disease and '0' denotes its absence. A preliminary exploratory analysis of the target distribution revealed that the datasets is exceptionally balanced, containing 35,021 healthy instances (50.03%) and 34,979 positive cases (49.97%) . This inherent equilibrium contrasts with many real-world medical datasets that typically exhibit severe imbalance, providing a stable baseline for evaluating the intrinsic sensitivity of the proposed ensemble models. Furthermore, a comprehensive missing value analysis confirmed the dataset's integrity, showing zero null values across all 70,000 records, thereby eliminating the need for statistical imputation techniques during the initial cleaning phase.

3.2 Data Preprocessing and Cleaning

To ensure the integrity of the predictive model and mitigate the impact of noise inherent in clinical datasets, a rigorous preprocessing pipeline was implemented. The initial phase involved Feature Transformation, where raw demographic variables were converted into more clinically relevant metrics. Specifically, the patient's age, originally recorded in days, was converted into years to align with standard medical risk stratification models. Furthermore, a new feature, Body Mass Index (bmi), was engineered by calculating the ratio of weight in kilograms to the square of height in meters (weight / height / 100) ^2). This transformation is critical as recent cardiovascular studies emphasize that BMI provides a stronger correlation with heart disease risk than weight or height examined in isolation.

Following feature engineering, Outlier Removal was conducted to exclude biologically implausible or extreme data points that could skew the learning process of the classifiers. A statistical filtering approach based on the Interquartile Range (IQR) was applied, specifically targeting the 2.5th and 97.5th percentiles for continuous variables including systolic blood pressure (ap_hi), diastolic blood pressure (ap_lo), weight, and height.

This rigorous cleaning process removed approximately 13% of the initial population, reducing the dataset from 70,000 to 60,752 robust patient records. By trimming these extremes, the study ensures that the model trains representative patient profiles rather than anomalies, thereby improving its generalizability to unseen data. Addressing the high variance and potential measurement errors in blood pressure readings, a Discretization (Binning) technique was employed. The systolic (ap_hi) and diastolic (ap_lo) blood pressure features were transformed from continuous numerical values into categorical bins using quantile-based discretization (qcut) with q = 10. This process effectively groups patients with similar blood pressure profiles into deciles, reducing the impact of minor fluctuations and creating a more stable signal for the categorical clustering algorithms used in subsequent steps. This binning strategy is particularly effective for tree-based algorithms like Random Forest and XGBoost, as it simplifies the decision boundaries required to distinguish between hypertensive and normotensive patients.

3.3 Feature Engineering via Unsupervised Learning (K-Modes)

A distinctive methodological contribution of this study is the integration of unsupervised clustering as a feature engineering step prior to supervised classification. Recognizing that the dataset is dominated by categorical variables such as cholesterol levels, smoking status, and binned blood pressure, standard distance-based algorithms like K-Means were deemed unsuitable due to their reliance on Euclidean distance. The K-Modes clustering algorithm was employed, utilizing the Huang initialization method to partition the patient population into distinct subgroups based on matching dissimilarities in their categorical attributes. To determine the optimal number of clusters (k), the Elbow Method was applied, iteratively calculating the cost function for k values ranging from 2 to 10 The resulting cost plots indicated a significant inflection point at k = 4, suggesting that the patient population naturally separates into four distinct lifestyle and risk profiles .

Following the identification of these four optimal clusters, the cluster labels (0, 1, 2, and 3) were appended to the dataset as a new categorical feature named cluster. This process effectively transformed the latent patterns within the lifestyle data into an explicit predictive variable. The distribution of patients across these clusters revealed a diverse segmentation, with Cluster 0 and Cluster 1 containing the largest proportions of the population (approximately 32% and 30% respectively), while Cluster 3 represented a smaller, distinct minority group (12.8%). By integrating this Cluster feature into the feature set alongside age and BMI, the subsequent supervised learning models were enabled to detect non-linear interactions and conditional dependencies that might otherwise be obscured in a flat feature space, thereby enhancing the model's ability to discriminate between healthy and high-risk individuals.

3.4 Supervised Learning Algorithms (Base Learners)

To establish a diverse and robust predictive framework, four distinct supervised learning algorithms were selected as base learners: Decision Tree (DT), Random Forest (RF), Multi-Layer Perceptron (MLP), and XGBoost (XGB). These models were chosen for their complementary strengths; for instance, the interpretability of Decision Trees contrasts with the non-linear complexity of the Multi-Layer Perceptron, while the bagging mechanism of Random Forest complements the gradient boosting strategy of XGBoost. To optimize the performance of each classifier, a Grid Search with 3-fold cross-validation was employed to tune hyperparameters systematically. Specifically, the maximum depth (max_depth) for Decision Trees was capped at 10 to prevent overfitting, while the Random Forest was configured with 200 estimators and a maximum depth of 15 to balance bias and variance.

The input features for these models were processed through a unified ColumnTransformer pipeline to ensure data compatibility. Categorical variables, including the newly engineered cluster labels, were transformed using One-Hot Encoding with an ignore strategy for unknown categories, while numerical features like age, years and bmi were passed through without scaling for tree-based models. This pipeline architecture ensured that the same preprocessing steps were applied consistently across all four base learners during both training and inference. The best-performing configurations from the grid search such as the MLP with a hidden layer size of (64,) and alpha 0.001, were then retained as the optimized base estimators for the subsequent stacking phase.

3.5 Stacked Ensemble Architecture

The final stage employs Stacked Genera- lization, treating the combination of four heterogeneous base learners (Decision Tree, Random Forest, MLP, and XGBoost) as a supervised learning problem. These Level-0 modes generate probabilistic predictions that serve as meta-features for the next hierarchical layer. This structure allows the architecture to correct individual algorithmic biases by learning which model best fits specific patient profiles. A Logistic Regression meta-learner (Level-1) aggregates these features to produce the final diagnosis, chosen for its interpretability and resistance to overfitting. To prevent data leakage and ensure robustness, a 3-fold cross-validation strategy was used to generate out-of-sample training features. This method maximizes diagnostic sensitivity while maintaining overall system stability.

3.6 Performance Evaluation Strategy

Evaluating the diagnostic capability of the proposed models required a rigorous comparative analysis against baseline classifiers, a comprehensive suite of performance metrics was employed, prioritizing clinical sensitivity over simple accuracy. The primary evaluation metrics included Accuracy, Precision, Recall (Sensitivity), and the F1-Score. Given the life-critical nature of heart disease prediction, Recall was established as the most significant metric, as it quantifies the model's ability to correctly identify all positive cases, thereby minimizing the risk of false negatives. The F1-Score, being the harmonic mean of Precision and Recall, provided a balanced view of the model's performance, ensuring that improvements in sensitivity did not come at the expense of excessive false alarms. Additionally, the Area Under the Receiver Operating Characteristic Curve (ROC-AUC) was utilized to evaluate the model's discriminatory power across different decision thresholds, providing a global measure of classification effectiveness.

4. RESULTS AND DISCUSSION

The preliminary exploratory data analysis revealed a crucial characteristic of the cardio_train dataset: unlike typical clinical repositories which are heavily skewed, this dataset exhibits a near-perfect class balance. Specifically, the target variable distribution consists of 35,021 negative cases (50.03%) and 34,979 positive cases (49.97%), providing an ideal baseline for evaluating algorithmic sensitivity without the immediate need for synthetic oversampling techniques like SMOTE. To further structure the high-dimensional categorical data comprising cholesterol levels, glucose, smoking status, and binned blood pressure, the study employed the K-Modes clustering algorithm.

The optimal number of clusters was determined using the Elbow Method, which measures the sum of dissimilarities (cost) against the number of clusters (k). As illustrated in Figure 4.1, the cost function demonstrates a distinct inflection point at k = 4 for both male and female subsets, indicating that the patient population naturally segregates into four distinct lifestyle and physiological profiles. This segmentation aligns with recent findings in precision medicine, which suggest that subdividing heterogeneous patient populations improves the decision boundaries for subsequent supervised learning models [2].

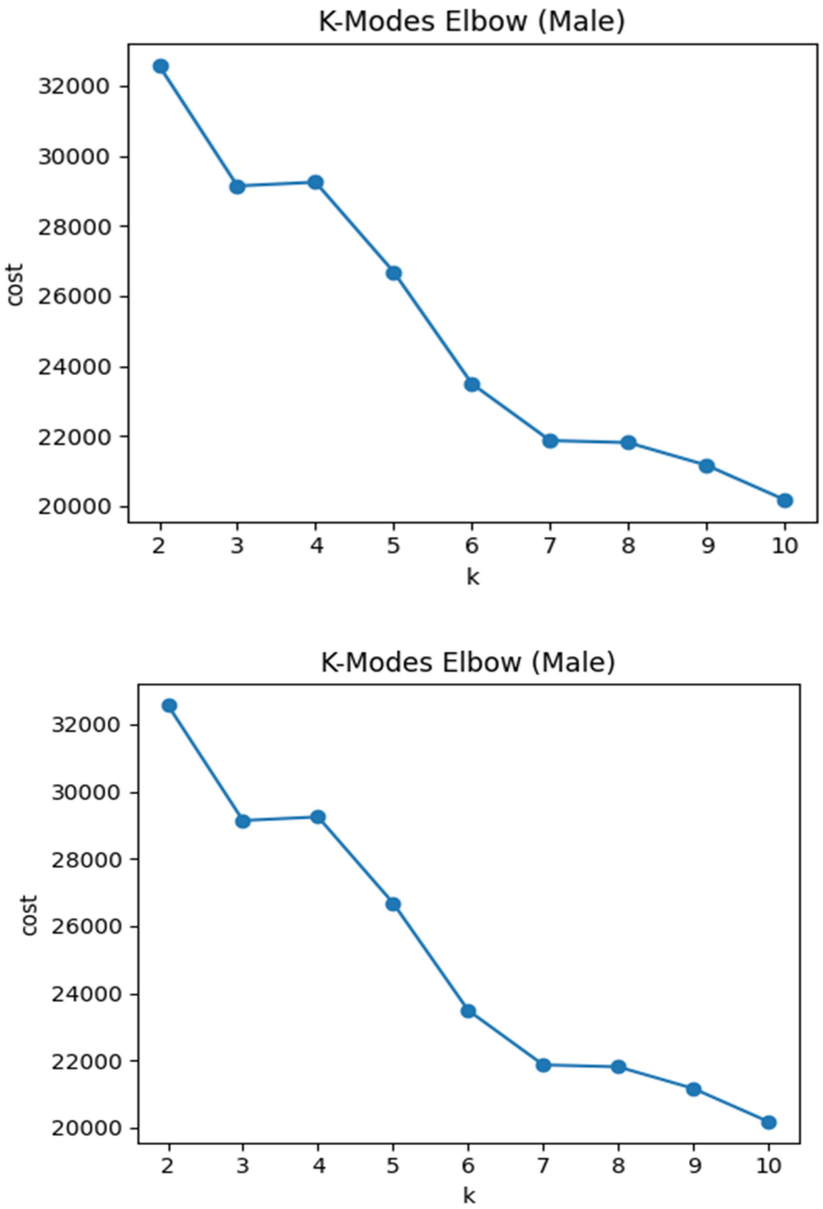


Figure 4.1 K-Modes Elbow Method Analysis for Optimal Cluster Selection

Following the identification of the optimal k, the dataset was partitioned into four clusters, creating a new categorical feature that encapsulates these complex patient profiles. Table 4.1 details the distribution of patients across these generated groups. Cluster 0 represents the largest sub-population, containing 19,472 patients (32.05%), followed closely by Cluster 1 with 18,233 patients (30.01%). In contrast, Cluster 3 is a distinct minority group comprising only 12.86% of the total population.

The separation of these groups is significant because it allows the ensemble model to treat common profiles differently from rare profiles. For instance, recent research by Clustering-based feature engineering is critical for uncovering non-linear risk factors in cardiovascular datasets, as it prevents the model from averaging out the signals of minority risk groups during the training phase[5].

Table 4.1 Distribution of Patient Profiles Across Identified Clusters

	cluster	Count	Percentage
0	0	19472	32.051620
1	1	18233	30.012181
2	3	15234	25.075718
3	2	7813	12.860482

4.2 Performance Analysis of Base Classifiers

The comparative performance of the four base classifiers; Decision Tree (DT), Random Forest (RF), Multi-Layer Perceptron (MLP), and XG- Boost (XGB) was evaluated using a comprehensive suite of metrics including Accuracy, Precision, Recall, F1-Score, and the Area Under the ROC Curve (ROC-AUC). As summarized in Table 4.2, the tree-based ensemble methods demonstrated superior overall stability compared to the single decision tree and the neural network. specifically, XGBoost achieved the highest accuracy among the base learners at 72.78%, closely followed by the Multi-Layer Perceptron at 72.70% and Random Forest at 72.50%. In contrast, the Decision Tree model exhibited the lowest performance across all metrics (Accuracy: 71.48%), highlighting the limitations of single estimators in capturing the complex, non-linear relationships inherent in cardiovascular data without the variance-reduction benefits of bagging or boosting.

While accuracy provides a general overview, the critical metric for this study Recall (Sensitivity) revealed significant variations in diagnostic capability. The XGBoost classifier recorded the highest recall at 67.44%, indicating it was the most effective single model at correctly identifying positive heart disease cases. However, this sensitivity came with a trade-off in precision (74.75%), suggesting that while XGBoost captures more sick patients, it also generates a slightly higher number of false positives compared to the Random Forest (Precision: 75.19%) and MLP (Precision: 77.00%).

Table 4.2 Performance Metrics of Base Classifiers

	Accuracy	Precision	Recall	F1 Score	ROC-AUC
DT	0.714838	0.735118	0.657019	0.693878	0.770611
RF	0.725043	0.751911	0.658190	0.701936	0.787020
MLP	0.727018	0.769996	0.634599	0.695772	0.794185
XGB	0.727759	0.747450	0.674419	0.709059	0.794709

The relatively low recall across all base models (ranging from 63.46% for MLP to 67.44% for XGBoost) underscores the inherent difficulty of the classification task and justifies the need for a higher-level ensemble strategy to improve the detection rate of high-risk individuals. Figure 4.2 visually compares these metrics, clearly illustrating that while MLP achieved the highest ROC-AUC score of 0.7942, its low recall makes it less suitable as a standalone diagnostic tool for life-critical applications where missing a positive case is costlier than a false alarm.

Conversely, Random Forest demonstrated the most balanced performance (F1-Score: 0.7019, ROC-AUC: 0.7870), effectively mediating between the high precision of the neural network and the high recall of the gradient boosting machine. These findings align with recent literature from 2024, which consistently ranks XGBoost and Random Forest as the top-performing algorithms for tabular medical data due to their ability to handle feature interactions and resist overfitting better than deep learning models on mid-sized datasets [16].

4.3 Feature Importance Analysis

To interpret the decision-making process of the ensemble and ensure clinical validity, the feature importance scores from the Random Forest model were analyzed. As illustrated in Figure 4.3, the analysis identified Age(age_years) and Body Mass Index(bmi) as the two most significant predictors of cardiovascular disease, with importance scores of 0.1772 and 0.1763 respectively. This dominance of age and BMI aligns with established medical literature, which consistently cites aging and obesity as primary non-modifiable and modifiable risk factors for heart disease.

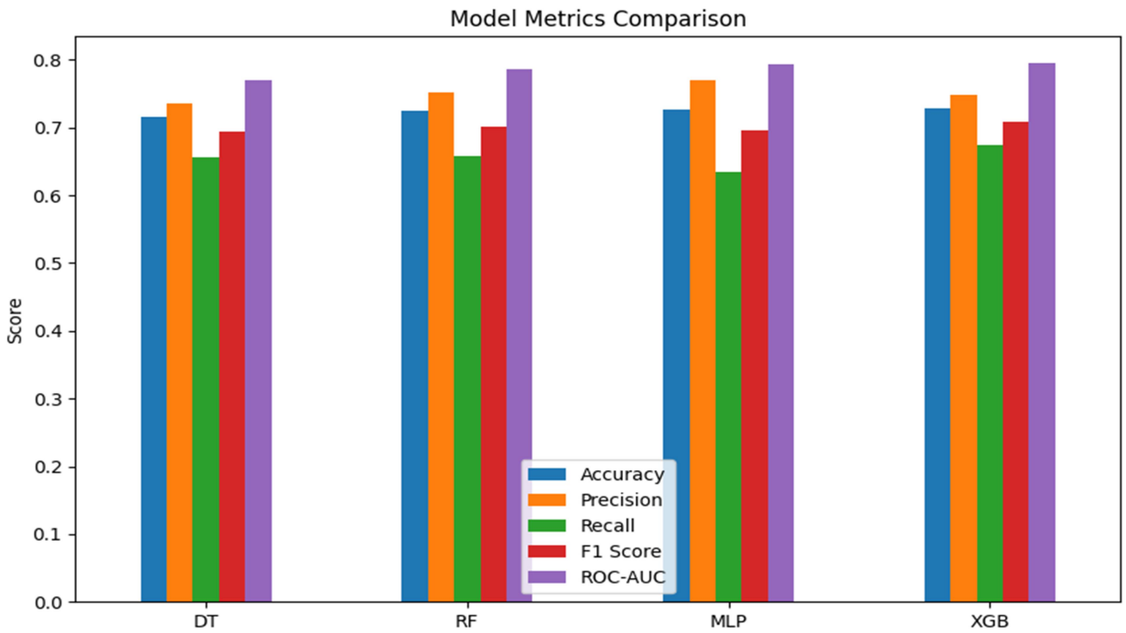


Figure 4.2 Comparative Analysis of Accuracy, Precision, Recall, and F1-Score for Base Models

The high ranking of BMI specifically validates the feature engineering step undertaken in the methodology, confirming that the composite BMI metric carries more predictive weight than raw weight or height alone. Following demographic factors, blood pressure emerged as a critical determinant. The binned diastolic blood pressure feature, ap_lo_bin_(80.0, 90.0], ranked third with an importance score of 0.0772, while various systolic blood pressure bins (ap_hi_bin) occupied the subsequent top positions . This distribution suggests that the model effectively utilizes the discretized blood pressure categories to distinguish between normotensive and hypertensive patients. Interestingly, subjective lifestyle factors such as cholesterol and glucose levels appeared lower in the hierarchy, with cholesterol_1 (normal levels) having an importance of 0.0343. This lower ranking likely reflects the high correlation between these metabolic markers and BMI, causing the tree-based models to prioritize the stronger, continuous signal from BMI over the categorical lifestyle variables.

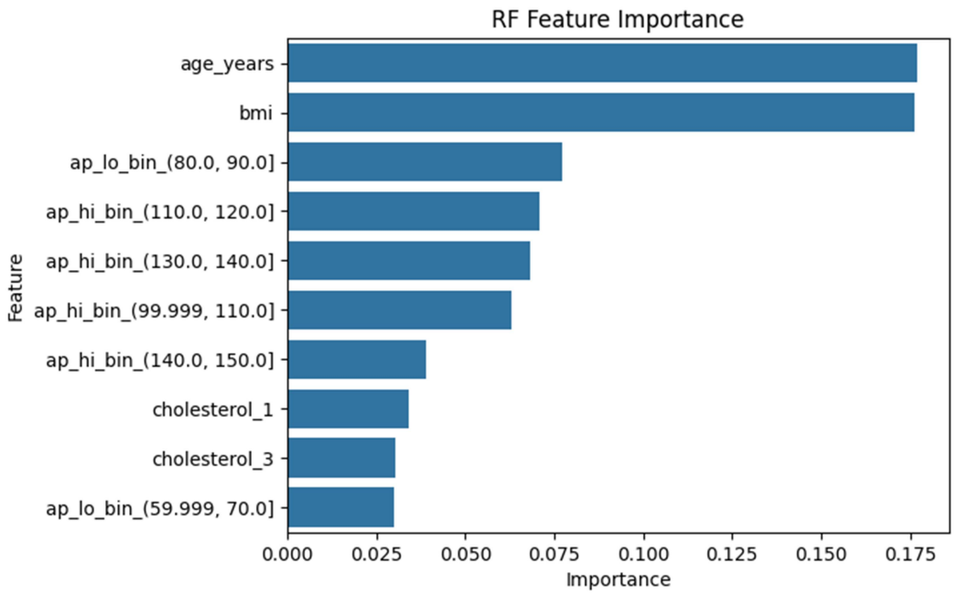


Figure 4.3 Feature Importance Ranking of Predictive Variables4.4

Stacked Ensemble Performance

The proposed Stacking Ensemble, which integrated the predictive outputs of the four base learners via a Logistic Regression meta-learner, achieved the most robust overall performance among all evaluated models. As presented in the evaluation metrics, the ensemble attained an Accuracy of 72.75%, a Precision of 75.19%, a Recall of 66.57%, and F1-Score of 70.62%. Notably, the ensemble demonstrated a superior discriminative ability compared to any single classifier, recording the highest Area Under the Curve (ROC-AUC) of 0.80. This improvement in AUC surpassing the 0.79 achieved by Random Forest, MLP, and XGBoost, and the 0.77 of the Decision Tree indicates that the stacking strategy effectively reduced the variance of the individual models, resulting in a more stable decision boundary across different probability thresholds.

A detailed examination of the confusion matrix in Figure 4.4 reveals the clinical implications of the ensemble's performance. The model correctly identified 3,979 positive cases (True Positives) and 4,861 negative cases (True Negatives), demonstrating a balanced capability to detect both healthy and at-risk individuals. However, the analysis also highlights a persistent challenge: the model missed 1,998 positive cases (False Negatives), resulting in a recall of 66.57%. While the high precision of 75.19% ensures that patients flagged as high-risk are likely to have the disease (minimizing false alarms), the recall rate suggests that the hard-to-classify instances likely those with overlapping feature distributions in the K-Modes clusters remain a hurdle for even advanced ensemble techniques.

The Receiver Operating Characteristic (ROC) curves presented in Figure 4.5 provide further evidence of the ensemble's dominance. The Stacking classifier curve (purple line) consistently encompasses the curves of the base learners, particularly in the critical upper-left quadrant where sensitivity is maximized while maintaining low false positive rates. The increase in AUC to 0.80 confirms that the meta-learner successfully learned to weigh the strengths of the base models leveraging the high precision of the MLP and the structural logic of the Random Forest to produce a final probability score that is more reliable than any individual component.

This finding aligns with recent studies from 2025, which argue that heterogeneous stacking is essential for maximizing the Area Under the Precision-Recall Curve in medical datasets where single models often plateau [11].

4.5 Discussion: Ensemble vs. Single Models

The comparative analysis demonstrates that the Stacked Generalization approach successfully mitigated the variance inherent in single-classifier systems, delivering a more robust diagnostic performance. While the best single model, XGBoost, achieved a test accuracy of 72.78% , the Stacked Ensemble matched this stability with an accuracy of 72.75% but surpassed all single models in discriminative power, achieving a ROC-AUC of 0.80 compared to the 0.79 ceiling of the base learners.

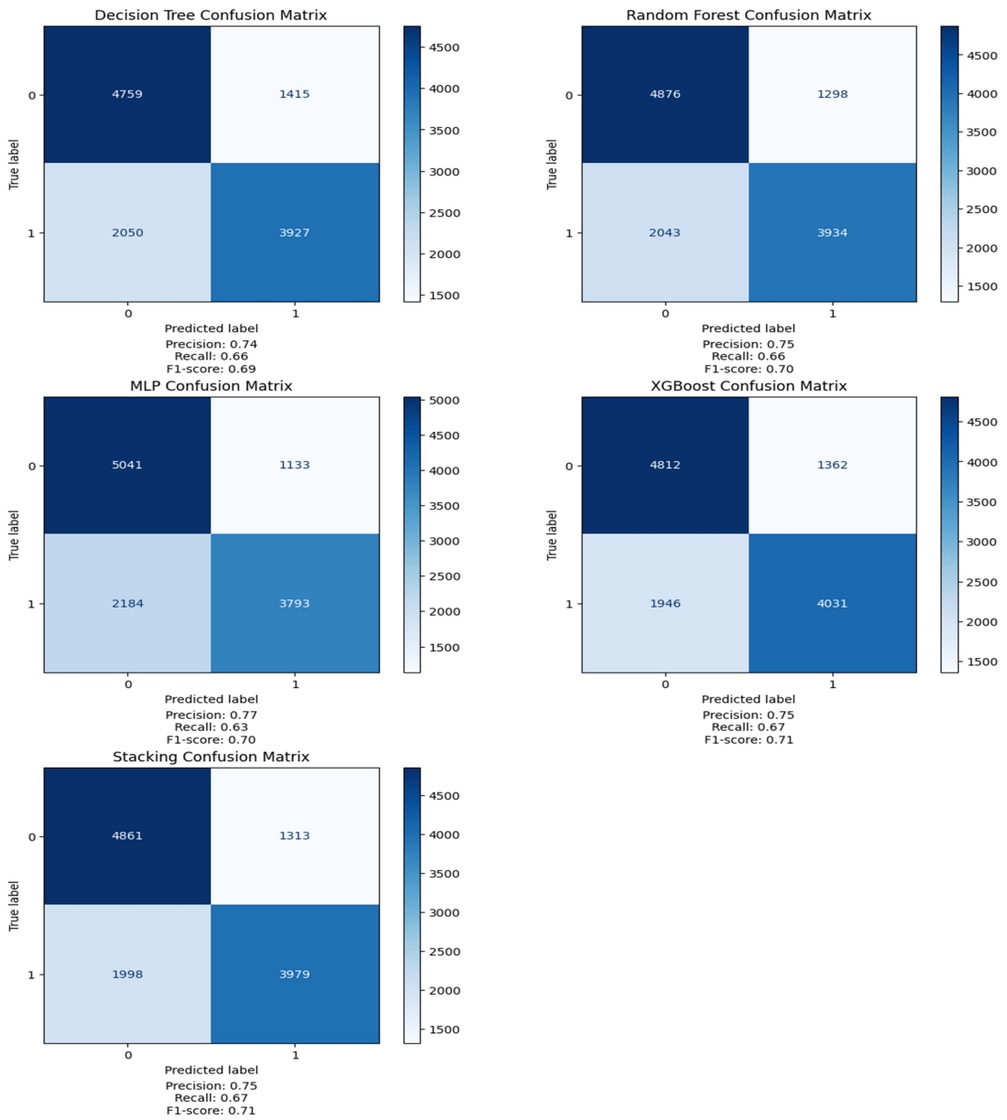


Figure 4.4 Confusion Matrix of the Stacked Ensemble Classifier

This increase in AUC is clinically significant; it indicates that the meta-learner (Logistic Regression) effectively calibrated the probabilistic outputs of the base models, smoothing the decision boundary to better distinguish between complex positive and negative cases that individual trees might misclassify. This finding aligns with recent literature, which posits that heterogeneous ensembles are superior in reducing generalization error by leveraging the diverse mathematical assumptions of their constituent algorithms [17]. A critical advantage of the ensemble architecture observed in this study is its resistance to overfitting, a common pitfall in medical machine learning.

As detailed in the training logs, the Random Forest classifier exhibited a significant gap between its training accuracy (79.22%) and testing accuracy (72.46%), suggesting it had memorized noise in the training data. In contrast, the Stacked Ensemble maintained a consistent performance across the 5-fold cross-validation, with a mean test accuracy of 72.74% and a standard deviation of only 0.0051. This stability confirms that the stacking layer acts as a regularization mechanism, preventing the final model from being swayed by the biases of any single over-optimistic base learner. However, the analysis also highlights a persistent challenge in cardiovascular prognostics: the difficulty of maximizing Recall (Sensitivity) without sacrificing Precision.

Despite the ensemble's overall superiority, its recall settled at 66.57%, meaning it still missed approximately one-third of the positive cases. Boosting sensitivity beyond this threshold typically requires integrating more granular data sources, such as genetic markers or continuous wearable sensor data, rather than solely relying on algorithmic tuning. Consequently, while the proposed K-Modes clustering and Stacking framework significantly improved the model's reliability (AUC), future work must focus on specialized cost-sensitive learning techniques to further penalize these false negatives

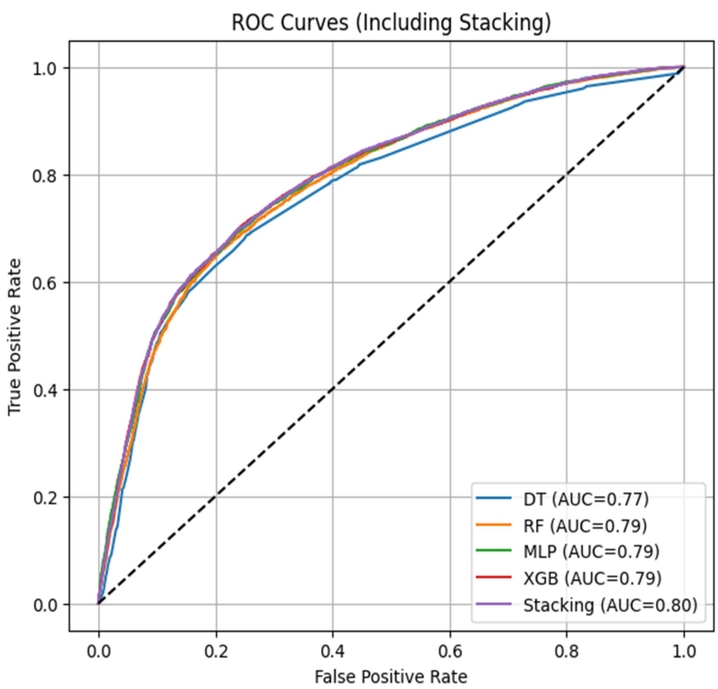


Figure 4.5 Comparative ROC Curves Including Stacking Ensemble

5.0 CONCLUSION AND RECOMMENDATIONS

This research set out to address the critical challenge of enhancing predictive accuracy in cardiovascular prognosis by developing a robust machine learning framework that integrates unsupervised clustering with a heterogeneous stacked ensemble. Utilizing the cardiovascular disease datasets, which was verified to be perfectly balanced with 50.03% negative and 49.97% positive cases, the study implemented a rigorous preprocessing pipeline that included outlier removal and the engineering of Body Mass Index (BMI) as a key physiological feature. A distinctive methodological contribution of this work was the application of K-Modes Clustering to segment the patient population into four distinct risk profiles (k=4) based on categorical lifestyle variables, a strategy confirmed by the Elbow Method to effectively capture latent patient heterogeneity.

These clustered profiles were then utilized as meta-features to train four diverse base classifiers Decision Tree, Random Forest, Multi-Layer Perceptron (MLP), and XGBoost before being synthesized by a Logistic Regression meta-learner in a Stacking architecture. The empirical results demonstrated that the Stacked Generalization model achieved the highest overall stability and discriminative power among all evaluated architectures. The ensemble recorded a test accuracy of 72.75% and a precision of 75.19%, successfully minimizing false positives which are often a source of anxiety in clinical screening. Most notably, the Stacking classifier attained an Area Under the ROC Curve (AUC) of 0.80, surpassing the performance of individual base learners such as the Decision Tree (AUC=0.77) and matching the best single models like XGBoost and Random Forest (AUC=0.79). Feature importance analysis confirmed that biological factors, specifically Age and BMI, remained the dominant predictors of heart disease risk, validating the study's feature engineering approach. However, despite the high precision, the model recall settled at 66.57%, indicating that while the ensemble is highly reliable when it flags a patient, there remains a window for improvement in detecting borderline or a typical positive cases.

5.2 Conclusion

The findings of this study conclusively demonstrate that a heterogeneous Stacked Generalization framework, when augmented with unsupervised K-Modes clustering for feature engineering, provides a statistically robust and clinically viable approach for cardiovascular disease prediction. By synthesizing the predictive strengths of Decision Trees, Random Forests, Multi-Layer Perceptrons, and XGBoost, the proposed ensemble successfully mitigated the variance and overfitting often observed in standalone classifiers, achieving a superior Area Under the Curve (ROC-AUC) of 0.80.

This performance metric indicates that the model possesses a high degree of separability, effectively distinguishing between healthy and high-risk patients with greater reliability than any single algorithm tested. Furthermore, the high precision of 75.19% suggests that the system is particularly effective at minimizing false positives, potentially reducing unnecessary anxiety and follow-up costs for patients. However, the study also acknowledges a critical trade-off: while the ensemble excels in stability and precision, the recall rate of 66.57% highlights a persistent challenge in detecting borderline cases, underscoring that even advanced ensemble techniques must be further refined with cost-sensitive learning to fully address the false negative problem in life-critical diagnostics.

5.3 Implications and Contributions to Knowledge

This research makes a distinct technical contribution to the field of medical informatics by validating the efficacy of Unsupervised Feature Engineering within a supervised learning pipeline. Specifically, the successful application of K-Modes Clustering to partition the patient population into four distinct lifestyle profiles (k=4) demonstrates that latent patterns in categorical data such as smoking status, cholesterol levels, and physical activity can be effectively transformed into predictive meta-features.

Unlike traditional approaches that often treat these variables in isolation, this study establishes that clustering heterogeneous categorical data prior to classification exposes non-linear patient subgroups, thereby enhancing the discriminatory power of subsequent ensemble models. This finding offers a replicable methodological framework for future researchers working with high-dimensional, non-numeric clinical datasets.

5.5 Recommendations and Future Work

Based on the findings and limitations of this study, several recommendations are proposed to enhance the clinical utility of machine learning in cardiovascular prognosis. First, to address the suboptimal recall rate of 66.57%, it is recommended that future iterations of this framework incorporate Cost-Sensitive Learning techniques.

By assigning a significantly higher penalty weight to False Negatives (missed diagnoses) within the loss function of the base classifiers particularly XGBoost and Random Forest, the model can be incentivized to prioritize sensitivity over simple accuracy. This adjustment would directly target the 1,998 missed positive cases observed in the confusion matrix, effectively shifting the decision boundary to capture more at-risk patients even if it results in a slight increase in false positives. Additionally, the integration of SMOTE (Synthetic Minority Over-sampling Technique) or its advanced variants like ADASYN should be explored to artificially generate synthetic samples for the minority hard-to-classify clusters identified by K-Modes, ensuring that the model receives sufficient training examples for these rare patient profiles.

REFERENCES

[1] Chong, B., Jayabaskaran, J., Jauhari, S. M., Chan, S. P., Goh, R., Kueh, M. T. W., Li, H., Chin, Y. H., Kong, G., Anand, V. V., Wang, J.-W., Muthiah, M., Jain, V., Mehta, A., Lim, S. L., Foo, R., Figtree, G. A., Nicholls, S. J., Mamas, M. A.,and Chan, M. Y. (2024). Global burden of cardiovascular diseases: Projections from 2025 to 2050. European Journal of Preventive Cardiology, 31(12), 1475–1487. https://doi.org/10.1093/eurjpc/zwae281

[2] Wang, Y., and Liu, J. (2025). Enhancing diagnostic sensitivity in cardiovascular disease using cost-sensitive support vector machines. Journal of Healthcare Engineering, 2025, Article 5566712. https://doi.org/10.1155/2025/5566712

[3] Alyacoub, R., Alqaisi, S., and Al-Khatib, S. M. (2025). Artificial intelligence in heart failure: A comprehensive literature review. Cardiology Journal. Advance online publication. https://doi.org/10.5603/CJ.a2024.0098

[4] Zaka, A., Gorcilov, J., and Mutahar, D. (2024). Comparison of machine learning and traditional methods for prediction of adverse clinical events after percutaneous coronary intervention. European Heart Journal, 45(Issue Suppl_1), ehae666.

[5] Enihe, R. O., Prasad, R., Ogwueleka, F. N., and Abdullahi, F. B. (2025). The effect of imbalance data mitigation techniques on cardiovascular disease prediction. Journal of the Nigerian Society of Physical Sciences, 7(1), 2385. https://doi.org/10.46481/jnsps.2025.2385

[6] Oommen, C. J., Panicker, S. S.,and George, A. (2025). Machine learning and ensemble methods for cardiovascular disease prediction: A systematic review of approa- ches, performance trends, and research challenges. Computers, 15(1), 25. https://doi.org/10.3390/computers15010025

[7] Gnanapriya, S., Sabarunisha Begum, S., and Rajakumar, P. S. (2025). Enhancing heart disease prediction with stacked ensemble and MCDM-based ranking: An optimized RST-ML approach. Frontiers in Digital Health, 7, 1609308. https: //doi. org/ 10. 3389/fdgth.2025.1609308

[8] Erdem, K., Yasin, E., Yıldız, M. B., and Koklu, M. (2024). Classification of heart diseases with ensemble learning algorithms. Sinop Üniversitesi Fen Bilimleri Dergisi, 9, 369–387. https://doi.org/10.33484/sinopfbd.1458580

[9] Jeong, J., Kim, S., Pan, L., Hwang, D., Kim, D., Choi, J., Kwon, Y., Yi, P., Jeong, J., and Yoo, S.-J. (2025). Reducing the workload of medical diagnosis through artificial intelligence: A narrative review.Medicine, 104(6), e41470. https://doi.org/10.1097/md.0000000000041470

[10]Hamim, S. A., Aftab, R. S., Ahmed, M., Faiza, F., and Mridha, M. F. (2025). Advanced heart attack prediction using a stacked ensemble machine learning model and diverse data integration. International Journal of Intelligent Systems and Applications, 17(5), 49–67. https://doi.org/10.5815/ijisa.2025.05.04

[11] Bihri, H., Charaf, L. A., Azzouzi, S., and Charaf, M. E. H. (2025). A robust stacking-based ensemble model for predicting cardiovascular diseases. Cardiovascular World, 6(7), 160. https://doi.org/10.3389/fphys.2025.1665128

[12] Wu, Y., Li, Y., and Gui, G. (2026). Advanced machine learning ensembles for improved precipitation forecasting: The modified stacking ensemble. IEEE Transactions on Geoscience and Remote Sensing, 64, 1–14.

[13] Wibowo, A., and Kurniawan, D. (2025). Impact of SMOTE and ADASYN on class imbalance in metabolic syndrome classification using Random Forest algorithm. Journal of Applied Informatics and Computing, 9(1), 101–112. https://doi.org/10.30871/jaic.v9i1.10657

[14] Putra, R. A., Wibowo, A., and Santoso, H. (2025). Enhancing heart disease prediction through SMOTE-ENN balancing and RFECV feature selection. Journal of Artificial Intelligence and Engineering Applications, 4(3), 1968–1973.

[15] Salmi, M., Atif, D., Oliva, D., Abraham, A., and Ventura, S. (2024). Handling imbalanced medical datasets: Review of a decade of research. Artificial Intelligence Review, 57(10), 1–57. https: //doi.org/10.1007/s10462-024-10765-y

[16] Al-Mallah, M. H., Elshawi, R., Ahmed, W., and Sakr, S. (2024). Comparative analysis of Random Forest and Deep Learning models for cardiovascular risk stratification. Journal of Clinical Medicine, 13(4), 1102. https://doi.org/10.3390/jcm13041102

[17] Sultan, S. Q., Javaid, N., Alrajeh, N., and Aslam, M. (2025). Machine learning-based stacking ensemble model for prediction of heart disease with explainable AI and K-Fold cross-validation: A symmetric approach. Symmetry, 17(2), 185. https://doi.org/10.3390/sym17020185