impact International Journals and Publications

© IIJP 2026 | Volume 2, Issue 3, pp. 952-967 | ISSN: 2636-4484

KNN IMPUTATION AND HARD VOTING ENSEMBLE FOR HEART DISEASE PREDICTION

Vincent Onochie Udechukwu[1] & Dr. Hafsa Kabir Ahmad[2]
Bayero University, Kano, Nigeria,
[1]**Email:** vincent.udechukwu417@gmail.com/**ORCID:** https://orcid.org/0009-0007-4265-2073
[2]**Email:** hkahmad.cs@buk.edu.ng/ORCID: https://orcid.org/0000-0001-6031-4405

ARTICLE INFO
Article No.: 0486
Accepted Date: 24/07/2026
Published Date: 10/08/2026
Type: Research

ABSTRACT
Heart disease (HD) remains one of the leading causes of morbidity and mortality worldwide, posing a significant public health challenge. Early and accurate identification of individuals at risk is essential for reducing disease progression, preventing complications, and improving patient outcomes. Machine learning (ML) has emerged as an effective approach for medical diagnosis by identifying complex patterns within clinical data and supporting evidence-based clinical decision-making. Consequently, numerous ML algorithms have been applied to cardiovascular disease prediction to enhance the accuracy and reliability of conventional diagnostic methods. However, limited research has examined the combined application of K-nearest neighbour (KNN) data imputation and a hard voting ensemble (HVE) model for HD prediction using the HD dataset from IEEE Dataport. This study evaluates the effectiveness of KNN data imputation and a HVE model in improving the predictive performance of HD classification. The study utilized 1,190 patient records from the IEEE Dataport HD dataset. Following data preprocessing, the dataset was divided into training and testing sets using an 80:20 ratio, resulting in 885 training samples and 238 testing samples. Four widely used classifiers, namely Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Random Forest (RF), and Extreme Gradient Boosting (XGBoost), were trained and evaluated. A HVE model was subsequently developed by combining the predictions of the individual classifiers to improve overall classification performance. The proposed framework achieved a classification accuracy of 93.70%, demonstrating that the HVE effectively enhanced predictive performance while reducing model bias. Furthermore, KNN data imputation effectively addressed missing values, contributing to improved classification accuracy and model robustness. Overall, the proposed framework provides a reliable approach for HD prediction and offers practical support for early diagnosis, clinical decision-making, and preventive healthcare.

Keywords: Heart disease prediction, machine learning, K-nearest neighbour imputation, hard voting ensemble, IEEE Dataport, classification.

Introduction

Heart disease (HD) affects millions of individuals worldwide and continues to exert a substantial burden on healthcare systems in terms of cost, morbidity, and mortality (Dubey et al., 2024). Cardiac failure, in particular, is widely recognized as one of the most prevalent and critical global health challenges (Gour et al., 2024). Broadly, HD encompasses a range of pathological conditions involving structural or functional abnormalities of the heart that impair normal blood circulation (Ozcan & Peker, 2023). According to the World Health Organization, cardiovascular diseases remain the leading cause of death globally, underscoring their profound public health impact (Ghasemieh et al., 2023; Kadhim & Radhi, 2023; Maidul Islam et al., 2024; Shaheenur Islam Sumon et al., 2022). In current clinical practice, the prediction and diagnosis of HD largely rely on physicians' assessments of patients' medical histories, reported symptoms, and findings from physical examinations. To address these research gaps, this study investigates the combined application of K-nearest neighbour (KNN) data imputation and a hard voting ensemble (HVE) learning approach for HD prediction using HD dataset from IEEE Dataport. Specifically, the study aims to: (i) evaluate the effectiveness of KNN data imputation in handling missing values and improving data quality; (ii) assess the predictive performance of four machine learning(ML) classifiers, namely Support Vector Machine (SVM), K-Nearest Neighbour (KNN), Random Forest (RF), and Extreme Gradient Boosting (XGBoost); (iii) develop and evaluate a majority-based HV ensemble model by integrating the predictions of the individual classifiers; and (iv) compare the predictive performance of the proposed HVE model with those reported in existing studies. The main contribution of this study lies in demonstrating the effectiveness of integrating KNN data imputation with an HVE framework to improve the performance and robustness of HD prediction, thereby providing a reliable decision-support approach for early diagnosis and clinical practice.

Literature Review

HD remains one of the leading causes of mortality worldwide, prompting extensive research into the application of ML techniques for early diagnosis and risk prediction. The availability of the IEEE Dataport comprehensive HD dataset, which combines records from the Cleveland, Hungarian, Switzerland, Long Beach VA, and Statlog databases, has enabled researchers to develop and evaluate a wide range of predictive models. Existing studies have primarily focused on ensemble learning, individual ML classifiers, preprocessing techniques, and hybrid learning frameworks to improve prediction accuracy. A substantial body of research has demonstrated the superiority of ensemble learning over individual ML models. Dubey et al. (2024) evaluated several classifiers, including Logistic Regression (LR), Decision Tree (DT), RF, SVM, XGBoost, Extra Trees (ET), and KNN, together with ensemble approaches such as majority voting, bagging, boosting, and stacking. Their results showed that the stacking ensemble achieved the highest prediction accuracy of 93.67%, outperforming all individual classifiers. Similarly, Maidul Islam et al. (2023) developed a stacked ensemble consisting of nine ML algorithms and reported an accuracy of 91.06%, while Liu et al. (2022) proposed a stacking framework that combined ten classifiers, including SVM, KNN, LR, RF, ET, Gradient Boosting DT, XGBoost, LightGBM, Catboost, and Multilayer Perceptron, achieving an accuracy of 89.86%. Likewise, Tezerjan and Mollabagher (2026) integrated LR, C5.0 Decision Tree, and Neural Networks into a hybrid model that attained an accuracy of 93.44%. In contrast to the predominance of stacking-based approaches, Akkur (2023) proposed a HVE model for HD prediction using the IEEE Dataport dataset. The study employed GridSearchCV to optimize the hyperparameters of the constituent classifiers and utilized SHapley Additive exPlanations (SHAP) to enhance model interpretability. The proposed HVE achieved a prediction accuracy of 92.78%, Collectively, these findings demonstrate that ensemble learning generally provides higher predictive performance. However, the overwhelming reliance on stacking architectures introduces increased computational complexity, longer training time,

and reduced model interpretability. Consequently, there remains limited evidence regarding the effectiveness of ensemble strategies, particularly optimized HVE, for HD prediction using the IEEE Dataport dataset.

Another category of studies has concentrated on evaluating the predictive capability of individual ML algorithms. Ozcan and Peker (2023) employed the Classification and Regression Tree (CART) algorithm to predict HD and determine the relative importance of clinical variables, reporting an accuracy of 87%. Kalla and Chandrasekaran (2023) investigated the relationship between clinical symptoms and HD using chi-square analysis and LR, achieving an accuracy of 85.12%. Similarly, Wazid et al. (2022) developed a secure Internet of Medical Things (IoMT)-based framework in which SVM achieved the highest prediction accuracy of 87.57%. These studies demonstrate that conventional ML algorithms can provide reliable diagnostic performance. Nevertheless, their predictive capabilities remain limited compared with ensemble methods because individual classifiers are unable to exploit the complementary strengths of multiple learning algorithms. In addition, most of these studies evaluated only a small number of models without investigating whether combining multiple optimized classifiers could further improve predictive performance.

Several researchers have also investigated the influence of preprocessing techniques on HD prediction. Magboo and Magboo (2023) compared four missing value imputation methods, namely mean, median, KNN, and Multiple Imputation by Chained Equations (MICE), in combination with Recursive Feature Elimination (RFE). Their findings indicated that AdaBoost with median imputation and RFE achieved the highest prediction accuracy of 84% on the IEEE Dataport dataset, while the four imputation techniques produced relatively similar predictive performance. Doppala et al. (2022) combined feature selection with ensemble learning and achieved an accuracy of 93.39% on the IEEE Dataport dataset, demonstrating that appropriate feature engineering can substantially improve classification performance. Although these studies confirmed the importance of preprocessing, they primarily evaluated imputation methods or feature selection independently and provided limited insight into how KNN-based missing value imputation influences the performance of optimized ensemble learning models.

Recent studies have further explored hybrid and deep learning approaches to improve cardiovascular disease prediction. Mahmud et al. (2023) proposed a ML metamodel that combined RF, Gaussian Naïve Bayes, and DT with KNN as the final estimator, obtaining an accuracy of 87%. Similarly, Rustam et al. (2022) introduced a convolutional neural network (CNN)-based feature extraction framework integrated with a soft voting ensemble comprising LR, SVM, and Stochastic Gradient Descent classifiers, achieving an accuracy of 93%. These approaches demonstrate that integrating deep learning with ensemble learning can improve prediction performance. However, they generally require extensive computational resources, sophisticated parameter tuning, and larger training datasets, thereby limiting their practical deployment in resource-constrained healthcare settings.

Previous studies consistently demonstrate that ensemble learning generally provides higher predictive performance. Several important research gaps remain. First, most existing studies have concentrated on stacking which is a computationally intensive ensemble techniques, while comparatively little attention has been given to optimized HVE, that are computationally efficient, and easier to interpret. Second, although several studies have investigated missing value imputation, few have systematically examined the impact of KNN-based imputation on the predictive performance of optimized HVE models using the IEEE Dataport HD dataset. Third, many published studies provide insufficient methodological detail regarding the sequence of preprocessing operations, particularly whether missing value imputation and outlier removal were performed before or after partitioning the dataset into training and testing sets. Performing these preprocessing steps before data partitioning

introduces data leakage, allowing information from the test data to influence model training and resulting in overly optimistic performance estimates. To overcome these limitations, the present study first partitions the dataset into training and testing sets before applying KNN-based missing value imputation and outlier removal exclusively to the training data. The learned preprocessing parameters are subsequently applied to the test data, thereby eliminating data leakage and ensuring a statistically valid evaluation of predictive performance. Furthermore, the study integrates HVE learning to develop a computationally efficient and robust HD prediction model using the IEEE Dataport dataset.

Methodology

Dataset Description

The HD dataset employed in this study is a consolidated compilation of five well-established datasets obtained from the IEEE Dataport repository, namely the Cleveland, Hungarian, Switzerland, Long Beach Veterans Affairs, and Statlog (Heart) datasets (Magboo & Magboo, 2023; Shaheenur Islam Sumon et al., 2024; Siddhartha, 2020). A detailed summary of the dataset characteristics and selected features is presented in **Table 1.**

Table 1 Heart Disease Dataset

Attribute	Variable Code	Unit	Data Type
Age	age	In Years	Numerical
Sex	sex	1,0	Binary
Chest Pain Type	cp	1,2,3,4	Nominal
Resting Blood Pressure	resting_bp_s	mm Hg	Numerical
Serum Cholesterol	cholesterol	mg/dL	Numerical
Fasting Blood Sugar	fasting_bs	1,0 > 120 mg/dl	Binary
Resting ECG Results	resting_ecg	0,1,2	Nominal
Maximum Heart Rate	max_heart_rate	71-202	Numerical
Achieved Exercise-Induced Angina	exercise_angina	0,1	Binary
ST Depression (Old peak)	old peak	depression	Numerical
ST Segment Slope	ST_slope	0,1,2	Nominal
Target Variable	target	0,1	Binary

Preprocessing

The dataset was loaded into Google Colab, Python 3.2 for subsequent processing. Preliminary inspection revealed that the resting blood pressure, serum cholesterol, and ST slope attributes contained zero values that did not represent physiologically plausible measurements but instead denoted missing or unrecorded observations. These values were therefore recoded as missing values (NaN) prior to any further preprocessing. Because this recoding step involved no parameter estimation, it was applied to the entire dataset. Subsequently, the dataset was partitioned into training and testing subsets using a stratified train–test split, with 80% of the records allocated to the training set and the remaining 20% reserved for testing. All subsequent preprocessing procedures involving parameter estimation were performed exclusively on the training subset, and the fitted transformations were subsequently applied to the testing subset. This strategy was adopted to prevent information from the testing data from influencing model development and evaluation.

1. Outliers Removal

Outlying observations in the continuous variables, namely age, resting blood pressure, serum cholesterol, maximum heart rate, and oldpeak, were identified using the Interquartile Range (IQR) method. For each variable, the first (Q1) and third (Q3) quartiles were computed from the training subset only, and the IQR was calculated as the difference between Q3 and Q1. Lower and upper outlier thresholds were then defined as $Q1 - 1.5 \times IQR$ and $Q3 + 1.5 \times IQR$, respectively. These thresholds, derived exclusively from the training subset, were subsequently applied to both the training and testing subsets, and observations falling outside the established bounds were removed. Records containing missing values at this stage, resulting from unrecorded measurements that had not yet been imputed, were retained and were not treated as outliers. Following the recoding process, serum cholesterol contained 172 missing values, whereas ST slope and resting blood pressure each contained one missing value. These missing values were subsequently imputed using the KNN imputation algorithm with the number of neighbors set to 5. The imputer was fitted exclusively on the training subset, after which the learned parameters were applied to impute missing values in both the training and testing subsets. Restricting model fitting to the training data ensured that the imputation process was not influenced by information contained in the testing subset, thereby preventing data leakage.

Categorical Variables

Categorical variables, including chest pain type, ST slope, resting electrocardiographic results, fasting blood sugar, sex, and exercise-induced angina, were transformed using one-hot encoding. The encoding scheme was derived solely from the training subset and subsequently applied to the testing subset. Column alignment was enforced to ensure that categories absent from either subset were represented by zero-valued columns, thereby maintaining consistent feature dimensions across both datasets.

Continous Variables

Continuous variables were standardized using the MinMax Scaler. The scaling parameters, namely the mean and standard deviation, were estimated exclusively from the training subset and subsequently applied to both the training and testing subsets. This approach ensured that no information from the testing data contributed to the estimation of transformation parameters. Four supervised ML algorithms SVM, XGBoost, RF, and KNN were trained using the preprocessed training data. Hyperparameter optimization for each classifier was performed using GridSearchCV with five-fold cross-validation on the training subset, employing classification accuracy as the optimization criterion. The optimal hyperparameter configuration identified for each classifier was subsequently evaluated on the independent testing subset. Following the development of the individual classifiers, an ensemble model was constructed using a HVE that combined the predictions of the four

optimized classifiers. The voting weight assigned to each classifier was determined through an exhaustive search over a predefined set of candidate weight combinations. Each candidate configuration was evaluated using five-fold cross-validation accuracy on the training subset, and the weight combination yielding the highest average cross-validation accuracy was selected. The final optimized HVE was then evaluated on the testing subset using accuracy, receiver operating characteristic area under the curve (ROC–AUC), and the classification report metrics, including precision, recall, and F1-score.

Exploratory Data Analysis

It is the initial process of examining a dataset to understand its characteristics, identify patterns, and detect data quality issues.

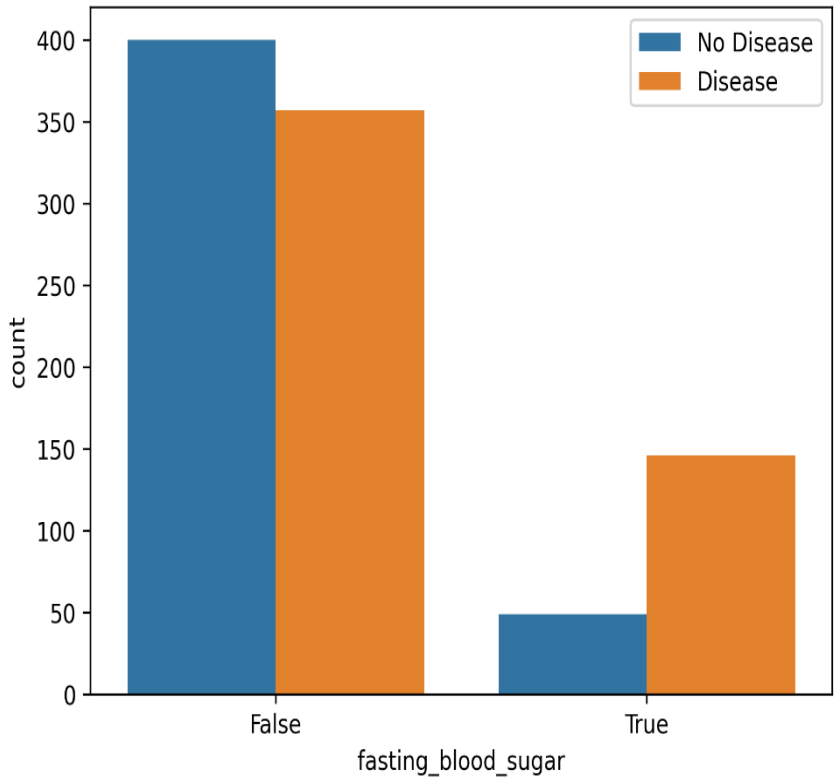


Figure 1. fasting blood sugar with the target variable

Figure 1 shows that most individuals had fasting blood sugar ≤120 mg/dL (False) regardless of HD status. However, elevated fasting blood sugar (True) was more common among individuals with HD than those without, suggesting a possible association with HD.

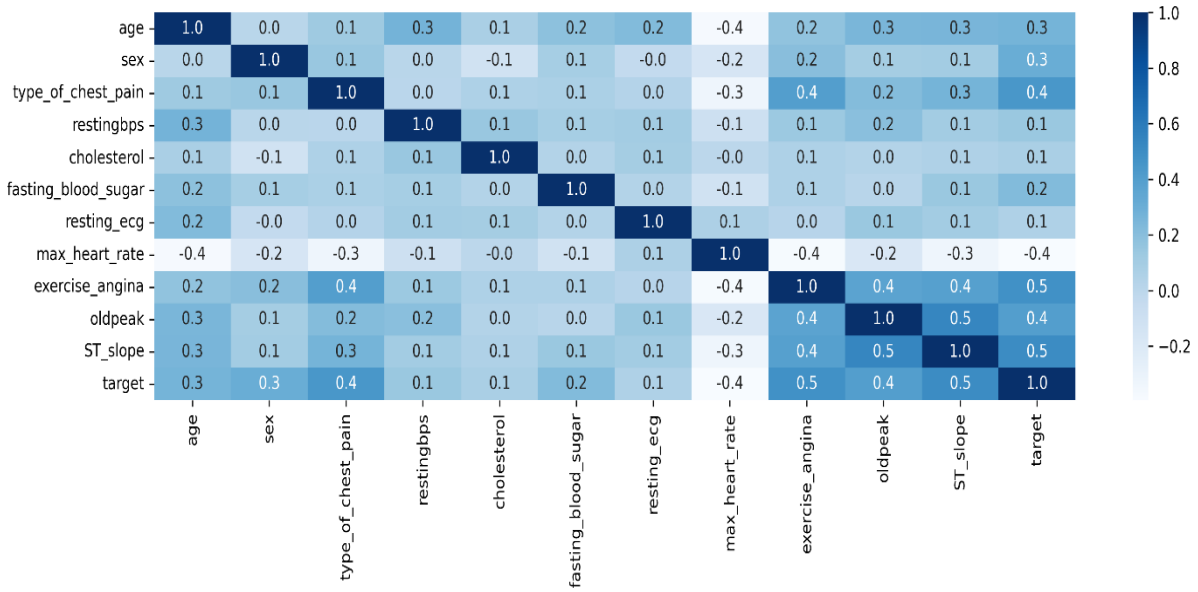


Figure 2. Correlation Matrix

Figure 2 presents the correlation matrix of the HD dataset. The target variable shows the strongest positive correlations with exercise angina (0.5), ST slope (0.5), chest pain type (0.4), and oldpeak (0.4), while maximum heart rate has the strongest negative correlation (-0.4), indicating its inverse relationship with HD.

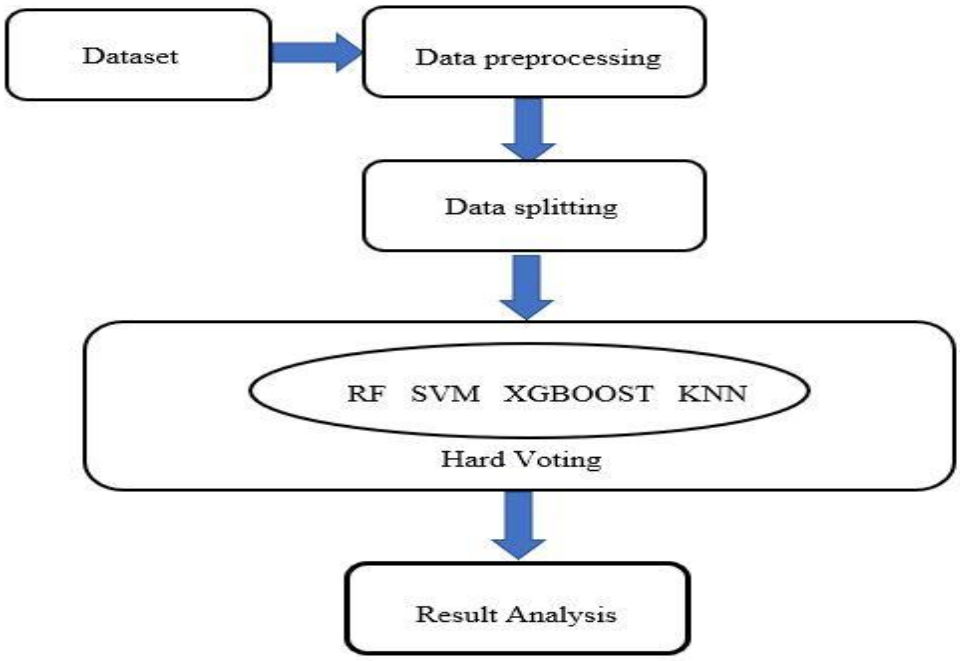


Figure 3 Work flow Diagram

KNN Imputation

Data imputation refers to the systematic process of replacing missing entries in a dataset with statistically informed estimates to preserve data completeness and analytical validity (Umar et al., 2025). In this study, KNN imputation technique was employed to address missing values within the dataset. KNN imputation is widely recognized for its flexibility, as it can accommodate both categorical and numerical variables without imposing stringent

distributional assumptions on the underlying data (Batista & Monard, 2003). The method belongs to the class of hot-deck imputation techniques originally developed to compensate for nonresponse in sample surveys by estimating missing values from similar observed instances (Zhang, 2012). Its effectiveness is particularly evident under the Missing at Random (MAR) assumption, where local relationships among observed variables can be exploited to generate reliable imputations (Sasu et al., 2025). In the present study, KNN imputation was implemented using five nearest neighbors (k = 5) and the Euclidean distance metric to quantify the similarity between observations in the feature space. Furthermore, distance-weighted averaging was employed during the imputation process, assigning greater influence to neighboring observations located closer to the incomplete instance while reducing the contribution of more distant neighbors. This weighting strategy improves the representativeness of the imputed values by prioritizing highly similar observations. The choice of k = 5 provides a balance between minimizing estimation variance and preserving local data characteristics, thereby reducing the likelihood of overfitting or excessive smoothing (Joel et al., 2025). Operationally, the algorithm computes the Euclidean distance between observations with complete and incomplete records, identifies the five most similar neighbors, and estimates missing values using a distance-weighted aggregation of their observed values (Altamimi et al., 2024). By exploiting the intrinsic relationships among observations, the KNN imputation procedure preserves the underlying data distribution, minimizes information loss, and enhances the quality of the input data for subsequent ML modelling, thereby improving predictive performance and model generalizability.

Performance Metrics (PM)

In ML, PMs constitute quantitative indicators used to assess the predictive accuracy and generalization capability of a model when applied to previously unseen data. These metrics provide an objective basis for comparing different models and evaluating their suitability for a given classification task.

Accuracy: It is defined as the proportion of correctly classified instances relative to the total number of instances evaluated. It represents the overall correctness of the model's predictions and is mathematically expressed in Equation (1).

$$\frac{True\ positive + True\ Negative}{True\ \ positive + True\ negative + false\ positive + false\ negative} \quad (1)$$

Precision: It is a PM that measures the proportion of true positive predictions among all instances classified as positive by the model. It reflects the reliability of positive predictions and is particularly important in clinical decision-making scenarios where false positives may lead to unnecessary interventions or resource utilization. It is mathematically defined as the ratio of correctly predicted positive instances to the total number of predicted positive instances, as expressed in Equation (2).

$$\frac{True\ positive}{True\ \ positive + false\ positive} \quad (2)$$

Recall: It is also referred to as sensitivity or true positive rate, quantifies the model's ability to correctly identify actual positive instances within the dataset. It measures the proportion of positive cases that are accurately detected among all true positive instances and is particularly critical in medical diagnostic applications, where failing to identify affected patients can have serious consequences. It is computed using the formula presented in Equation (3).

$$\frac{True\ positives}{True\ \ positive + false\ negatives} \quad (3)$$

F-Score: It is defined as the harmonic mean of precision and recall, offering a single, balanced metric that accounts for both false positives and false negatives. By integrating these two

complementary measures, the F-Score provides a more informative assessment of a model's classification performance, particularly in situations involving class imbalance. The F-Score is mathematically expressed in Equation (4).

$$F - Score = 2 * \frac{Precision * Recall}{Precision + Recall} \quad (4)$$

Receiver Operating Characteristic (ROC): The Curve is a widely used graphical tool for evaluating the discriminative performance of classification models across varying decision thresholds. The ROC curve plots the true positive rate (TPR), also referred to as recall, against the false positive rate (FPR), thereby illustrating the trade-off between sensitivity and specificity. In this study, the TPR is computed using Equation (3), while the FPR is derived from Equation (5). This visualization provides a comprehensive assessment of a model's ability to distinguish between positive cases and negative instances, effectively capturing its capacity to maximize correct detections while minimizing false alarms.

$$FPR = \frac{FP}{FP + TN} \quad (5)$$

Area Under the Curve (AUC) quantifies the overall discriminative capability of a classification model by measuring the area beneath the Receiver Operating Characteristic (ROC) curve across all possible decision thresholds. The AUC value typically ranges from 0 to 1, where values closer to 1 signify superior model performance and a strong ability to effectively differentiate between positive and negative classes. Consequently, a higher AUC reflects greater robustness and reliability of the classifier in distinguishing true cases from non-cases (Chandrasekhar & Peddakrishna, 2023).

Hyperparameter optimization (HO): It constitutes a critical stage in the development of robust ML models, as it directly influences generalization capability and predictive reliability. This process entails a systematic and structured exploration of the hyperparameter space to identify parameter configurations that yield optimal model performance. Unlike model parameters learned during training, hyperparameters govern the learning behavior and structural characteristics of algorithms and must be defined prior to model fitting.

In the context of ML, HO aims to determine the most effective combination of hyperparameter values that maximizes performance on a validation dataset, thereby reducing overfitting and enhancing model stability (Budholiya et al., 2022). Appropriate tuning of these hyperparameters ensures that learning algorithms operate under conditions that are well aligned with the underlying data distribution and task-specific objectives.

In this study, HO was systematically conducted for all evaluated models using the IEEE Dataport dataset. The selected hyperparameter configurations, summarized in Table 2, were derived through GridSearchCV. These optimized settings were adopted to enhance classification accuracy, improve convergence behavior, and strengthen the overall predictive performance of the proposed ML framework.

CLASSIFIERS	GRIDSEARCHCV HYPERPARAMETER TUNING VALUES
SVM	'C': 100, 'gamma': 'auto', 'kernel': 'rbf'
RF	'max_depth': None, 'max_features': 'sqrt', 'min_samples_leaf': 1, 'min_samples_split': 2, 'n_estimators': 100
XGBoost	'learning_rate': 0.1, 'max_depth': 8, 'min_child_weight': 3, 'n_estimators': 180, 'subsample': 0.7
KNN	'metric': 'manhattan', 'n_neighbors': 9, 'weights': 'distance'

Table 2 Hyperparameters Optimization values for classifiers

Data Classification

In this study, four supervised ML classification algorithms were implemented using Python to enhance the prediction of HD. Each classifier was trained, evaluated, and comparatively analyzed to determine its effectiveness in identifying the likelihood of HD among patients. The selected models represent a diverse set of learning paradigms, enabling a balanced assessment of distance-based, margin-based, and ensemble-driven approaches. A concise description of each classification technique is presented below.

1 k-Nearest Neighbors (KNN): The KNN algorithm is a non-parametric, instance-based learning method that retains the entire training dataset and classifies unseen samples based on their similarity to neighboring instances. Classification is performed by assigning the class most frequently represented among the k nearest data points in the feature space (Dubey et al., 2024). Owing to its conceptual simplicity and flexibility, KNN does not rely on an explicit hypothesis function or prior model assumptions, making it relatively easy to implement and interpret (Maidul Islam et al., 2024).

2 Support Vector Machine (SVM): SVM is a supervised learning algorithm widely applied to classification and regression tasks. It operates by constructing an optimal decision boundary, known as a hyperplane, that maximizes the margin between data points of different classes (Behera et al., 2022). By focusing on the most informative samples referred to as support vectors, SVM achieves strong generalization performance, particularly in high-dimensional feature spaces (Magboo & Magboo, 2023).

3 Random Forest (RF): RF is an ensemble-based supervised learning method that aggregates the predictions of multiple decision trees to generate a final output. Each tree is trained on a randomly selected subset of the data using bootstrap sampling, while feature randomness is introduced during node splitting. This bagging-based strategy reduces variance, mitigates overfitting, and enhances predictive accuracy (Chandrasekhar & Peddakrishna, 2023).

4 Extreme Gradient Boosting (XGBoost): XGBoost is an advanced gradient boosting algorithm designed to improve predictive performance through sequential model training. Unlike bagging-based methods, boosting constructs decision trees iteratively, with each new model correcting the errors of its predecessors. XGBoost incorporates regularization mechanisms and efficient optimization techniques, enabling it to handle complex nonlinear relationships and large-scale datasets effectively.

5 Hard voting Ensemble (HVE): Ensemble method improve the overall accuracy of the model by combining the predictions of multiple classifiers using the HV technique. In this study, a HVE strategy was adopted, wherein each classifier independently produces a prediction, and the final class label is determined by majority voting across all models. Each model operates in a unique way and may extract different information from the data, relying on a single classifier alone may not be sufficient (Shareef et al., 2023)

Result and Discussion

The ML models, SVM, RF, XGBoost, and KNN were trained on the training set and subsequently assessed using the test set to evaluate their predictive effectiveness. To ensure optimal model performance, Hyperparameter tuning was conducted prior to final evaluation. Model optimization involved iterative tuning of hyperparameters to identify configurations that maximized predictive accuracy. In this study, the GridSearchCV technique was employed to exhaustively search the predefined hyperparameter space for each classifier. The optimal values obtained for the prediction models are summarized in Table 3. Following optimization, the evaluated models demonstrated strong predictive performance. Accuracy scores of 85.71%, 94.54%, 93.70%, and 93.28% were achieved by the SVM, RF, XGBoost and KNN models, respectively, indicating the effectiveness of the preprocessing, imputation and optimization strategies in enhancing HD prediction performance.

Table 3 Performance Measures of the Models

Models	Accuracy	Precision	Recall	F1 score	ROC
SVM	85.7	87.10	85.70	86.40	92.0
RF	94.50	95.20	94.40	94.80	97.8
XGBOOST	93.70	93.70	94.40	94.10	97.0
KNN	93.30	95.10	92.10	93.50	96.5
HVE	93.70	95.10	92.90	94.0	----

Subsequently, the predictive performance was further enhanced through the implementation of a HVE strategy that integrated the four individual ML models: SVM, RF, XGBoost, and KNN. By aggregating the class predictions of these base learners and determining the final output via majority voting, the ensemble capitalized on the complementary strengths of the individual classifiers. This ensemble-based approach resulted in an overall classification accuracy of 93.70%, demonstrating its effectiveness in delivering more robust and reliable HD predictions compared with the standalone model.

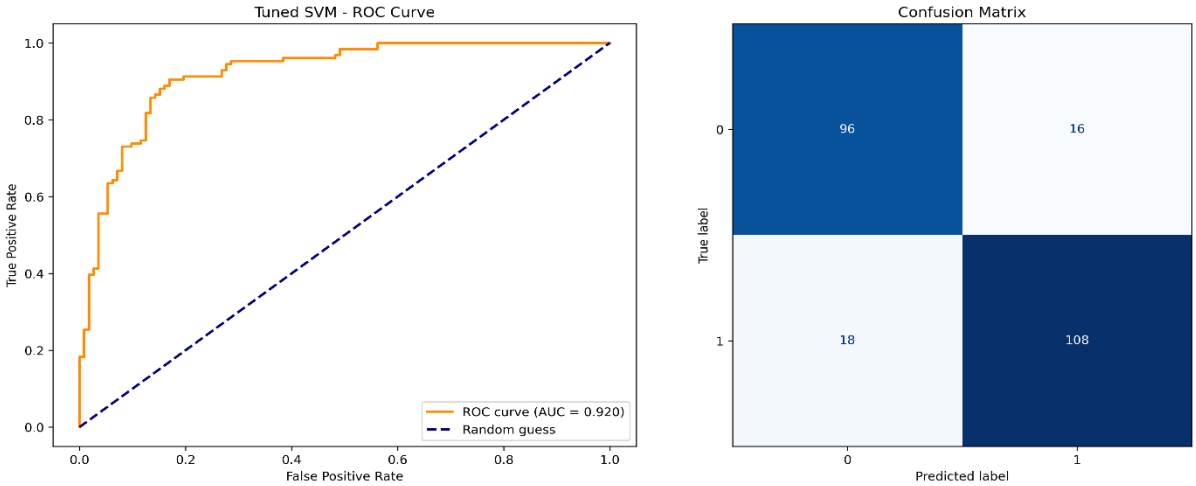


Figure 4 ROC Curve and Confusion matrix for SVM

The SVM classifier demonstrated a strong ability to distinguish between patients with and without HD, as illustrated by the ROC curve presented in Figure 4. The model exhibited balanced classification performance, effectively identifying both positive and negative cases with no significant bias toward either class. Its high recall, precision, and true negative rate indicate consistent predictive capability and reliable generalization on the test set. These results demonstrate that the SVM is a dependable base learner for the proposed ensemble model.

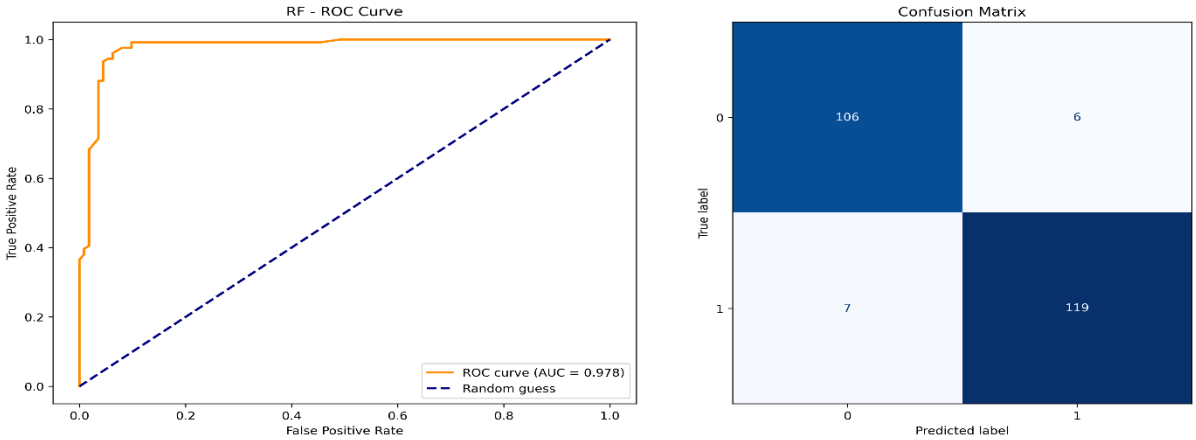


Figure 5 ROC Curve and Confusion matrix for RF

The RF model performed strongly on the test set, as illustrated by the ROC curve and confusion matrix presented in Figure 5. The confusion matrix indicates that the model achieved a balanced classification performance by accurately identifying both patients with and without HD while maintaining relatively few misclassifications. The high recall, precision, and true negative rate demonstrate the model's effectiveness in detecting positive cases without substantially increasing false alarms. Overall, these results show that RF exhibited robust and balanced predictive performance, making it one of the strongest base learners considered for the ensemble.

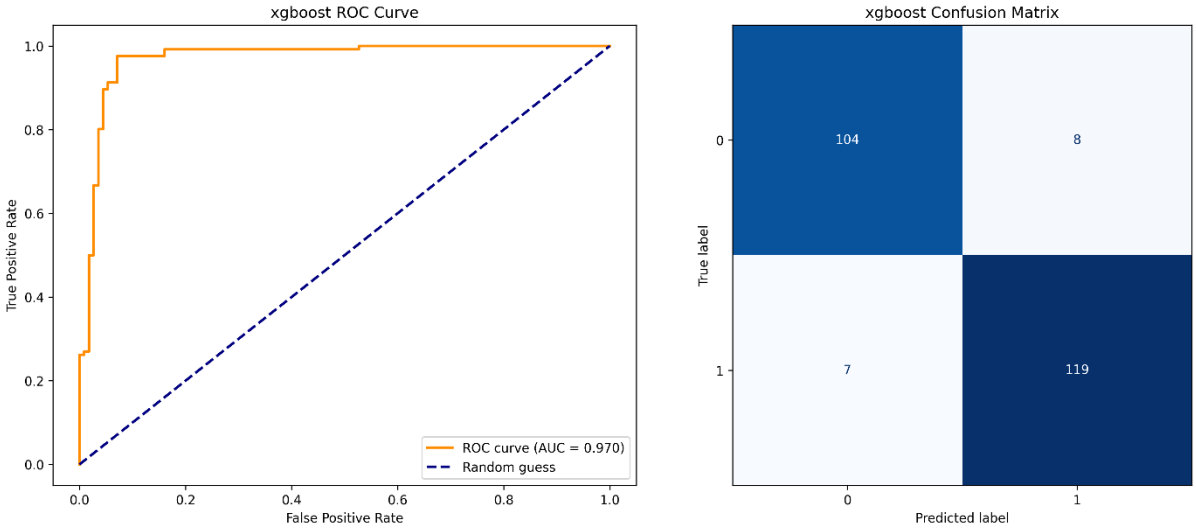


Figure 6 ROC Curve and Confusion matrix for XGBoost

The XGBoost model demonstrated strong classification performance on the test set. It achieved high predictive accuracy while maintaining a balanced ability to identify both patients with HD and those without the disease. The model exhibited high recall (sensitivity) and true negative rate (specificity), indicating effective detection of positive cases with relatively few missed diagnoses and accurate identification of negative cases. Furthermore, the high area under the ROC curve (AUC) confirms the model's excellent discriminative ability, demonstrating that it can reliably distinguish between patients with and without HD across different classification thresholds.

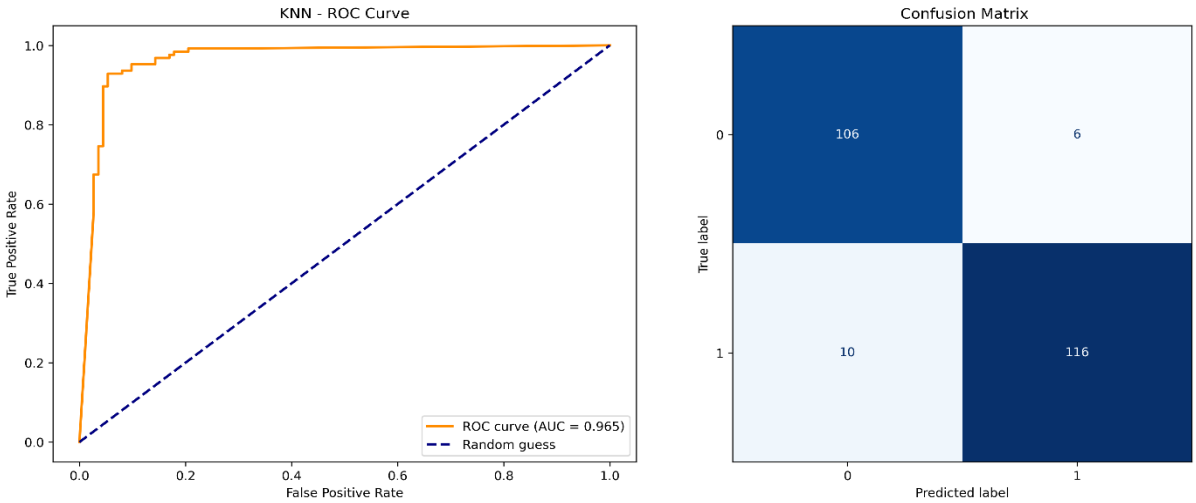


Figure 7 ROC Curve and Confusion matrix for KNN

The KNN classifier demonstrated strong discriminative performance on the test dataset, as illustrated by the ROC curve and confusion matrix in Figure 6. The model effectively distinguished between patients with and without HD while maintaining balanced classification performance across both classes. Its high precision, recall, true negative rate, and F1-score indicate reliable predictive capability with relatively few misclassifications. Overall, these findings demonstrate that the KNN classifier provides accurate and consistent performance for HD prediction.

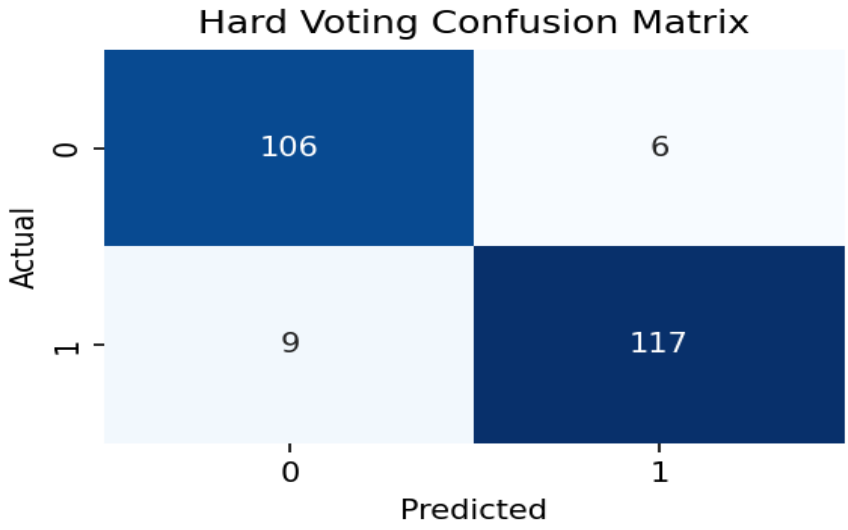


Figure 8 Confusion matrix for HVE Model.

The confusion matrix (CM) presented in Fig. 8 illustrates the classification performance of the proposed ensemble model. The model correctly identified 106 positive cases, underscoring its strong capability in detecting patients with HD. In terms of negative classification, 117 instances were accurately predicted, indicating robust performance in recognizing non-diseased cases. Misclassification errors were minimal, with 9 false positives, where negative instances were incorrectly labelled as positive, and 6 false negative, in which a positive case was misclassified as negative. A comprehensive evaluation of the proposed approach was subsequently conducted and benchmarked against conventional ML techniques. Comparative performance results with relevant existing methods are summarized in Table 4. The findings

demonstrate that the proposed ensemble framework consistently performed comparable to or exceeding the limited prior work using HVE on this specific dataset reported in the HD prediction studies, thereby highlighting its effectiveness and practical relevance for clinical decision-support applications.

Table 4 Comparison of Proposed Model with Related Works

AUTHORS	METHODOLOGY USED	DATASET	ACCURACY (%)
(Magboo & Magboo, 2023)	Ensemble approach	IEEE Dataport	84.0
(Ozcan & Peker, 2023)	Classification and regression tree algorithm	IEEE Dataport	87.0
(Akkur, E.,2023)	Hard voting	IEEE Dataport	92.78
(Dubey et al., 2024)	Stacking	IEEE Dataport	93.67
(Tezerjan And Mollabagher., 2026)	Hybrid model	IEEE Dataport	93.44
Proposed Method	**KNN Imputation + Hard voting**	IEEE Dataport	**93.70**

Conclusion and Future Work

This study presents a comprehensive evaluation of ML techniques for HD prediction, with a primary emphasis on improving performance through an optimized HVE classifier. A range of classification algorithms was systematically evaluated, and the experimental results demonstrated that the proposed HVE model achieved a classification accuracy of 93.70%. The performance of the ensemble model can be attributed to its ability to combine the complementary strengths of individual classifiers, thereby improving prediction stability and reducing model-specific bias. Despite these promising findings, the study has several limitations. First, the proposed model was developed and evaluated using a single combined HD dataset, which may limit the generalizability of the results to other populations and clinical environments. Second, the model was not validated using an independent external test dataset; therefore, its performance in real-world clinical settings remains to be established. In addition, the computational cost and inference time associated with the ensemble model were not investigated, which may be relevant when considering deployment in resource-constrained healthcare environments.

Future research should address these limitations by validating the proposed approach on external multicenter datasets with diverse patient populations to assess its robustness and generalizability. Further studies should also investigate computational efficiency and model scalability, while exploring advanced feature selection techniques, data imbalance handling strategies, and emerging learning architectures to further improve predictive performance and facilitate practical clinical implementation.

References

Akkur, E. (2023). Prediction of Cardiovascular Disease Based on Voting Ensemble Model and SHAP Analysis. *Sakarya University Journal of Computer and Information Sciences*, *6*(3), 226–238. https://doi.org/10.35377/saucis...1367326

Altamimi, A., Alarfaj, A. A., Umer, M., Alabdulqader, E. A., Alsubai, S., Kim, T. H., & Ashraf, I. (2024). An automated approach to predict diabetic patients using KNN imputation and effective data mining techniques. *BMC Medical Research Methodology*, *24*(1). https://doi.org/10.1186/s12874-024-02324-0

Batista, G. E. A. P. A., & Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, *17*(5–6), 519–533. https://doi.org/10.1080/713827181

Behera, M. P., Sarangi, A., Mishra, D., & Sarangi, S. K. (2022). A Hybrid Machine Learning algorithm for Heart and Liver Disease Prediction Using Modified Particle Swarm Optimization with Support Vector Machine. *Procedia Computer Science*, *218*, 818–827. https://doi.org/10.1016/j.procs.2023.01.062

Budholiya, K., Shrivastava, S. K., & Sharma, V. (2022). An optimized XGBoost based diagnostic system for effective prediction of heart disease. *Journal of King Saud University - Computer and Information Sciences*, *34*(7), 4514–4523. https://doi.org/10.1016/j.jksuci.2020.10.013

Chandrasekhar, N., & Peddakrishna, S. (2023). Enhancing Heart Disease Prediction Accuracy through Machine Learning Techniques and Optimization. *Processes*, *11*(4). https://doi.org/10.3390/pr11041210

Doppala, B. P., Bhattacharyya, D., Janarthanan, M., & Baik, N. (2022). A Reliable Machine Intelligence Model for Accurate Identification of Cardiovascular Diseases Using Ensemble Techniques. *Journal of Healthcare Engineering*, *2022*. https://doi.org/10.1155/2022/2585235

Dubey, M., Tembhurne, J., & Makhijani, R. (2024). Heart Failure prediction on diversified datasets to improve generalizability using 2-Level Stacking. *Multidisciplinary Science Journal*, *6*(3). https://doi.org/10.31893/multiscience.2024028

Ghasemieh, A., Lloyed, A., Bahrami, P., Vajar, P., & Kashef, R. (2023). A novel machine learning model with Stacking Ensemble Learner for predicting emergency readmission of heart-disease patients. *Decision Analytics Journal*, *7*. https://doi.org/10.1016/j.dajour.2023.100242

Gour, S., Randa, R., Pandagre, K. N., Neema, A., Jat, S., & Malakar, D. (2024.). A Comprehensive Review on Predicting Heart Failure with Ensemble Learning: Integrating Environmental Variables and Health Perspectives. In *International Journal of Environmental Sciences* (Vol. 11). https://www.theaspd.com/ijes.php

Joel, L. O., Doorsamy, W., & Paul, B. S. (2025). A comparative study of imputation techniques for missing values in healthcare diagnostic datasets. *International Journal of Data Science and Analytics*, *20*(7), 6357–6373. https://doi.org/10.1007/s41060-025-00825-9

Kadhim, M. A., & Radhi, A. M. (2023). Heart disease classification using optimized Machine learning algorithms. *Iraqi Journal for Computer Science and Mathematics*, *4*(2), 31–42. https://doi.org/10.52866/ijcsm.2023.02.02.004

Kalla, D., & Chandrasekaran, A. (2023). HEART DISEASE PREDICTION USING MACHINE LEARNING AND DEEP LEARNING. *International Journal of Data Mining & Knowledge Management Process*, *13*(03), 1–14. https://doi.org/10.5121/ijdkp.2023.13301

Liu, J., Dong, X., Zhao, H., & Tian, Y. (2022). Predictive Classifier for Cardiovascular Disease Based on Stacking Model Fusion. *Processes*, *10*(4). https://doi.org/10.3390/pr10040749

Magboo, V. P. C., & Magboo, M. S. A. (2023). Cardiovascular disease prediction with imputation techniques and recursive feature elimination. *AIP Conference Proceedings*, *2602*. https://doi.org/10.1063/5.0124079

Mahmud, I., Kabir, M. M., Mridha, M. F., Alfarhood, S., Safran, M., & Che, D. (2023). Cardiac Failure Forecasting Based on Clinical Data Using a Lightweight Machine Learning Metamodel. *Diagnostics*, *13*(15). https://doi.org/10.3390/diagnostics13152540

Maidul Islam, M., Nasrin Tania, T., Akter, S., & Hassan Shakib, K. (2024). An Improved Heart Disease Prediction Using Stacked Ensemble Method. Telecommunications Engineering, vol 490. Springer, Cham. https://doi.org/10.1007/978-3-031-34619-4_8

Ozcan, M., & Peker, S. (2023). A classification and regression tree algorithm for heart disease modeling and prediction. *Healthcare Analytics*, *3*. https://doi.org/10.1016/j.health.2022.100130

Rustam, F., Ishaq, A., Munir, K., Almutairi, M., Aslam, N., & Ashraf, I. (2022). Incorporating CNN Features for Optimizing Performance of Ensemble Classifier for Cardiovascular Disease Prediction. *Diagnostics*, *12*(6). https://doi.org/10.3390/diagnostics12061474

Sasu, G. V., Ciubotaru, B. I., Goga, N., & Vasilățeanu, A. (2025). Addressing Missing Data Challenges in Geriatric Health Monitoring: A Study of Statistical and Machine Learning Imputation Methods. *Sensors*, *25*(3). https://doi.org/10.3390/s25030614

Shareef, A. Q., & Kurnaz, S. (2023). Deep Learning Based COVID-19 Detection via Hard Voting Ensemble Method. *Wireless Personal Communications*. https://doi.org/10.1007/s11277-023-10485-2

Siddhartha, M. (2020). Heart disease dataset (comprehensive) [Data set]. IEEE Dataport. https://ieee-dataport.org/open-access/heart-disease-dataset-comprehensive.

Shaheenur Islam Sumon, M., Sakib Bin Islam, M., Sohanur Rahman, M., Sakib Abrar Hossain, M., Khandakar, A., Hasan, A., Murugappan, M., & H Chowdhury, M. E. (n.d.). *CardioTabNet: A Novel Hybrid Transformer Model for Heart Disease Prediction using Tabular Medical Data*.

Tezerjan, M. Y., & Mollabagher, M. (2026). From Data to Treatment: Early Detection of Heart Disease Using Machine Learning Techniques. In *Journal of Ilam University of Medical Sciences* (Vol. 34, Number 1).

Umar, N., Hassan, S. K., Umar, A., & Ahmed, S. S. (2025). Predicting Heart Diseases by Selective Machine Learning Algorithms. *Journal of Applied Sciences and Environmental Management*, *29*(1), 255–261. https://doi.org/10.4314/jasem.v29i1.32

Wazid, M., Singh, J., Das, A. K., Shetty, S., Khan, M. K., & Rodrigues, J. J. P. C. (2022). ASCP-IoMT: AI-Enabled Lightweight Secure Communication Protocol for Internet of Medical Things. *IEEE Access*, *10*, 57990–58004. https://doi.org/10.1109/ACCESS.2022.3179418

Zhang, S. (2012). Nearest neighbour selection for iteratively kNN imputation. *Journal of Systems and Software*, *85*(11), 2541–2552. https://doi.org/10.1016/j.jss.2012.05.073