

A Machine Learning Framework for Heart Disease Prediction With Feature Selection and Hyperparameter Optimization Across Benchmark and Real-World Datasets

Sunanda Budiha^{1, 2, ✉}, Sheetalrani R. Kawale², Nitin Agarwal³

1. Computer Science, Government First Grade College, Bagalkot, IND

2. Computer Science, Karnataka State Akkamahadevi Women's University, Vijayapura, IND

3. Consulting Physician & Cardiologist, Ayush MultiSpeciality Hospital and Research Centre, Vijayapura, IND

Received: May 07, 2026 | Review began: May 25, 2026 | Review ended: June 28, 2026 | Published: July 03, 2026

© Copyright 2026

This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Abstract

Heart disease is a major cause of death worldwide, and thus there is a need to have good, sound, and clinically usable predictive models. In this paper, a machine learning model is developed for the prediction of heart diseases using both a benchmark dataset (UCI-Statlog) and a real-world dataset obtained from the Ayush Multi-Specialty Hospital and Research Centre (AMSHRC), Vijayapura, Karnataka, India. Five machine learning methods - Decision Tree, Random Forest, Logistic Regression, XGBoost, and AdaBoost - were tested under three different experimental conditions: without feature selection, Chi-Square feature selection, and Analysis of Variance F-Test feature selection. Hyperparameters were optimized using Grid Search Cross-Validation. The Random Forest model achieved 88% accuracy without feature selection and 86% accuracy with Chi-Square selection, and 85% with Analysis of Variance F-test feature selection methods applied to the UCI-Statlog dataset, which means that feature selection did not improve the performance of the model on this dataset. Across all experimental configurations, XGBoost achieved the highest accuracy of 96% and an area under the curve of 0.99 on the AMSHRC dataset, while Logistic Regression exhibited fairly consistent predictive performance despite increasing clinical complexity in the dataset, highlighting the robustness of the evaluated models. The proposed framework provides a realistic and generalizable decision-support model to determine the risk of heart diseases in the early stages. In conclusion, this study is based on the prediction of heart diseases using machine learning, feature selection, and real-world data along with grid search cross-validation and classification.

Categories: AI applications, AI/ML-based decision support systems, Machine Learning (ML)

Keywords: heart disease prediction, cardiovascular disease, feature selection, grid search cross-validation, decision support systems

Introduction

Cardiovascular diseases, particularly heart disease, are the leading cause of mortality worldwide. According to the World Health Organization, heart disease accounts for the highest number of deaths globally, causing more than 20 million deaths annually [1]. The increasing trends of heart-related diseases in both developed and developing countries highlight the need for improved diagnostic and predictive interventions to reduce associated health risks.

Heart disease is a category of diseases that involve the structure and functioning of the heart, including coronary artery disease, heart failure, cardiac arrhythmias, and congenital heart defects. Key risk factors for heart disease are unhealthy dietary habits, physical inactivity, tobacco use, excessive alcohol consumption, high cholesterol levels, hypertension,

How to cite this article:

Budiha S, Kawale S R, Agarwal N (July 03, 2026) A Machine Learning Framework for Heart Disease Prediction With Feature Selection and Hyperparameter Optimization Across Benchmark and Real-World Datasets. Cureus J Comput Sci 3 : es44389-026-00150-7. DOI <https://doi.org/10.7759/s44389-026-00150-7>

diabetes, and genetic predisposition [2,3]. These factors contribute to the gradual accumulation of plaque within the arteries, a condition known as atherosclerosis, which restricts blood flow and may cause myocardial infarction or stroke [4].

Clinical diagnosis of heart disease is commonly performed using a combination of tests and diagnostic tools such as electrocardiography (ECG), echocardiography (ECHO), tread mill test (TMT), chest X-rays, blood tests, and coronary angiography (CAG) [5]. These tests are essential; however, they are not always conclusive. Variations in patient symptoms, underlying causes, and interpretation of results may lead to delayed or uncertain diagnoses, placing the patient at risk unless immediate treatment is provided.

Artificial intelligence is a revolutionary technology in the healthcare industry, offering great potential for early diagnosis and disease prediction [6,7]. Machine learning (ML) and deep learning (DL) algorithms can be used to capture complex patterns in large datasets and uncover latent associations that traditional statistical methods may not capture [8,9]. These models can assess risk at the individual level, enabling clinicians to intervene earlier and more effectively.

The quality and diversity of training data are critical determinants of the reliability of such models. Many existing studies have been conducted on small or biased samples [10-23], limiting their generalizability. Although some studies demonstrate high predictive ability, the results obtained using such algorithms may vary when applied to different datasets or in clinical settings. Furthermore, most studies rely on publicly available datasets such as UCI and Statlog, which are valuable for model assessment but may not fully capture the complexities and diversity present in standard clinical practice.

Larger datasets have been employed in more recent papers [11,13,14,16,18,20,24,25] to address these shortcomings, though relatively few have been validated in real-world clinical studies. This paper attempts to address them by evaluating ML models on a benchmark dataset (UCI-Statlog) and a new real-world clinical dataset (Ayush Multi-Specialty Hospital and Research Centre (AMSHRC)). Both feature selection and non-feature selection methods are implemented to assess the predictive performance under different conditions to assess different ML algorithms.

The clinical dataset AMSHRC used in this paper has also been employed in our earlier research on predicting heart disease using ML models, DL architectures, and feature-group-aware gated neural fusion techniques [26-28]. In contrast to those studies, the present work focuses on feature selection and hyperparameter optimization through a grid search approach, along with comparative evaluation using both benchmark and clinical datasets.

The novelty of this work is not the development of a new ML algorithm. Instead, it stems from the systematic evaluation of multiple optimized ML models on benchmark (UCI Statlog) and real-world clinical (AMSHRC) datasets within a unified framework for preprocessing, feature selection, and hyperparameter optimization. The study further investigates whether feature selection could improve prediction accuracy and evaluates the robustness of the models across datasets with varying clinical complexities. The relatively small performance variances across feature selection settings imply that the assessed models were resistant to changes in the selected feature set.

The main contributions of this paper are the following:

- A systematic comparison of benchmark and real-world clinical datasets to assess the difference in model performance in a controlled and real-world clinical environment.
- Integration of demographic, physiological, behavioral, and diagnostic information obtained via standard clinical practice for the prognostication of cardiovascular disease.
- A comparative analysis of feature selection methodologies, including Chi-Square Test and Analysis of Variance (ANOVA) F-Test, demonstrates that feature selection models had strong performance across multiple selection techniques, indicating resilience to variations in the chosen features.
- Objective comparison of several ML algorithms with a common preprocessing pipeline and hyperparameter tuning using grid search cross-validation.

How to cite this article:

Budihal S, Kawale S R, Agarwal N (July 03, 2026) A Machine Learning Framework for Heart Disease Prediction With Feature Selection and Hyperparameter Optimization Across Benchmark and Real-World Datasets. *Cureus J Comput Sci* 3 : es44389-026-00150-7. DOI <https://doi.org/10.7759/s44389-026-00150-7>

Related work

Several ML and artificial intelligence frameworks for predicting heart disease have been developed in recent years using UCI, Statlog, and other benchmark datasets. For instance, Babu et al. [10] introduced a Quantum-Enhanced Machine Learning framework comprising Quantum Bagging Ensemble (QBE), Quantum Support Vector Machine (SVM), and Quantum Artificial Neural Network models, tested on a combined Kaggle Heart Disease dataset consisting of 1,190 instances [11]. Chi-Square feature selection yielded the highest accuracy of 96.1% for the QBE model, outperforming Quantum SVM (95.0%) and Quantum Artificial Neural Network (92.9%). Omkari and Shaik [12] studied heart disease prediction using both large and small Kaggle datasets [13,14], incorporating additional clinical features such as Mean Arterial Pressure (MAP), Body Mass Index (BMI), and PP. Feature selection was performed using Chi-Square, Mutual Information, and ANOVA F-Test methods, while the performance of SVM, Decision Tree (DT), Random Forest (RF), and Multi-Layer Perceptron classifiers was analyzed. The authors proposed a Two-Layer Voting framework, optimized via grid search cross-validation, which achieved accuracies of 99.0% and 99.3% with hard and soft voting schemes, respectively. Vijayaraj and Pasupathi [15] proposed a hybrid Harris-Hawks optimization technique along with context-aware feature selection and analyzed K-Nearest Neighbors (KNN), Logistic Regression (LR), SVM, and RF models on a dataset of 270 instances [16]. The H-HGO with SVM and LR achieved an accuracy of 94.74%.

Mahajan et al. [17] introduced a Statistical Feature Selection (SFS) stacking approach, in which feature selection methods - ANOVA F-Test, Information Gain, Gini Index, and Chi-Square - were stacked to select the best features from the UCI dataset [18], comprising 920 samples from Cleveland (303 samples), Hungary (294 samples), Switzerland (123 samples) and VA Long Beach (200 samples). They proposed Ensemble Stacked-Learning methods, i.e., SFS-Stacking method with Cross-Validation Classifier (SFS-SCVC) and SFS-SVM. The SFS-SCVC achieved the highest accuracy of 97.81%, with SFS-SVM having the lowest accuracy at 96.71%. Ram Kumar et al. [19] tested various ML and DL models on the Cleveland dataset containing 303 samples. They applied three ML models - KNN, SVM-Linear, SVM-Radial Basis Function - and two DL models - Artificial Neural Network and Convolutional Neural Network. Their work did not select the best features, but all features were factored in prediction. Results indicated that SVM-Linear achieved the highest accuracy at 74.2% among the tested models. Cenitta et al. [20] introduced a Hybrid-Residual Attention-Enhanced Long Short-Term Memory (LSTM) model, which integrated attention-residual mechanisms with stacked LSTM layers for prediction. Using the Cleveland dataset (303 samples), their approach attained an accuracy of 97.71%.

Addisu et al. [21] combined Transfer Learning with a Conventional Visual Geometry Group-16 (VGG16) Convolutional Neural Network-based feature extraction method to predict heart disease. To address the issue of limited dataset size, they employed a Conditional-Tabular Generative Adversarial Network to generate synthetic samples. The extracted features were then used with various ML models, including DT, KNN, LR, RF, SVM, and Gradient Boosting (GB). On the dataset described in [14], the highest accuracy (92%) was achieved by VGG16 with RF. Teja and Rayalu [22] used the dataset discussed in [11] and evaluated several ML models, including Model Averaged Neural Network, Multivariate Adaptive-Regression Splines, Bayesian Generalized Linear Method, Boosted Generalized Linear Model, Flexible Discriminant Analysis, Conditional Inference Tree, XGBoost, Bagged Tree, Adaptive Boosting (AB), Naive Bayes, GB, and LR. Only seven features were manually selected to train the models, with XGBoost achieving the highest accuracy of 93%. Ganie et al. [23] also used the dataset from [11] and tested 15 ML base models: Stochastic Gradient Descent, Categorical Boosting, Light-Gradient Boosting Machine, Extra Trees, Linear Discriminant Analysis, RF, DT, SVM, Naive Bayes, Multi-Layer Perceptron, KNN, LR, GB, AB, and XGBoost. From these, six models, CatBoost, XGBoost, Light-Gradient Boosting Machine, Extra Trees, RF, and LR, were selected to develop stacking and voting models. The study employed hyperparameter tuning and k-fold cross-validation. The results revealed that voting and stacking methods achieved 98% accuracy in prediction. Table 1 provides a summary of the literature review.

How to cite this article:

Budihal S, Kawale S R, Agarwal N (July 03, 2026) A Machine Learning Framework for Heart Disease Prediction With Feature Selection and Hyperparameter Optimization Across Benchmark and Real-World Datasets. *Cureus J Comput Sci* 3 : es44389-026-00150-7. DOI <https://doi.org/10.7759/s44389-026-00150-7>

Reference	Models	Preprocessing	Feature selection	Total samples	Highest accuracy
[10]	BE, SVM, ANN, QBE, QSVM, QANN	Yes	Chi-Square	1,190	QBE: 96.1%
[12]	DT, RF, MLP, TLV Hard Voting (TLV-HV), TLV Soft Voting (TLV-SV)	Yes	ANOVA F-Test, Mutual Information, Chi-Square	1,025	TLV-SV: 99.3%
[15]	KNN, LR, SVM, RF	No	CAM	270	SVM, LR: 94.74%
[17]	LR, DT, RF, KNN, SVM, XGB, SFS-SVM, SFS-SCVC	Yes	ANOVA F-Test, Information Gain, Gini Index, Chi-Square	920	SFS-SCVC: 97.81%
[19]	KNN, SVM-Linear, SVM-RBF, ANN, CNN	Yes	No	303	SVM-Linear: 74.2%
[20]	HRAE-LSTM	Yes	No	303	HRAE-LSTM: 97.71%
[21]	DT, KNN, LR, RF, SVM, GB	Yes	VGG16	1,025	RF: 92%
[22]	MANN, MARS, BGLM, BGGLM, FDA, CIT, XGB, BT, AB, NB, GB, LR, KNN and RF	No	Manual Feature Selection	1,190	XGB: 93%
[23]	SGD, CB, LGBM, ET, LDA, RF, DT, SVM, NB, MLP, KNN, LR, GB, AB, XGB, Stacking, Voting	Yes	No	1,190	Stacking, Voting: 98%

TABLE 1: Summary of the literature review.

BE, Bagging Ensemble; SVM, Support Vector Machine; ANN, Artificial Neural Network; QBE, Quantum Bagging Ensemble; QSVM, Quantum Support Vector Machine; QANN, Quantum Artificial Neural Network; DT, Decision Tree; RF, Random Forest; MLP, Multi-Layer Perceptron; TLV-HV, Two-Layer Voting–Hard Voting; TLV-SV, Two-Layer Voting–Soft Voting; KNN, K-Nearest Neighbors; LR, Logistic Regression; XGB, Extreme Gradient Boosting; SFS-SVM, Statistical Feature Selection–Support Vector Machine; SCVC, Stacking Cross-Validation Classifier; RBF, Radial Basis Function; CNN, Convolutional Neural Network; HRAE-LSTM, Hybrid Residual Attention-Enhanced–Long Short-Term Memory; GB, Gradient Boosting; MANN, Model Averaged Neural Network; MARS, Multivariate Adaptive Regression Splines; BGLM, Bayesian Generalized Linear Model; BGGLM, Boosted Generalized Linear Model;

How to cite this article:

Budihal S, Kawale S R, Agarwal N (July 03, 2026) A Machine Learning Framework for Heart Disease Prediction With Feature Selection and Hyperparameter Optimization Across Benchmark and Real-World Datasets. Cureus J Comput Sci 3 : es44389-026-00150-7. DOI <https://doi.org/10.7759/s44389-026-00150-7>

FDA, Flexible Discriminant Analysis; CIT, Conditional Inference Tree; BT, Bagged-Tree; AB, Adaptive Boosting; NB, Naïve Bayes; SGD, Stochastic Gradient Descent; CB, CatBoost; LGBM, Light Gradient Boosting Machine; ET, Extra Trees; LDA, Linear Discriminant Analysis; ANOVA, Analysis of Variance

Table 1 summarizes selected existing studies that have applied artificial intelligence and ML to predict heart disease. Despite their promising predictions, some drawbacks remain. First, much of the existing literature uses benchmark datasets with limited clinical diversity. There are fewer studies that use real-world clinical data to validate their models. Furthermore, the effect of the feature selection and model optimization has not been tested across different datasets. The proposed work addresses these issues by combining UCI and Statlog datasets for more diversity, collecting a new AMSHRC real-world clinical dataset, applying similar systematic preprocessing, performing extensive feature selection, and rigorously evaluating performance with and without feature selection for both datasets.

Materials And Methods

The methodology of this work consists of eight significant subsections. The section *Dataset Preparation* describes the preparation of the datasets, including the combination of the UCI-Statlog dataset and the introduced AMSHRC dataset. The section *Flow of the Proposed Work* presents the overall analytical pipeline applied in this work. The section *Exploratory Data Analysis* discusses data distribution and characteristics. The preprocessing steps, including missing value imputation, outlier handling, and encoding, are described in the section *Preprocessing*. The *Feature Selection* section is dedicated to selecting relevant features using the Chi-Square test and ANOVA F-Test. The section *ML Models* provides an overview of the different ML models that have utilized Hyperparameter Grid Search Cross-Validation (HGSCV), as presented in the section *Hyperparameter Grid Search Cross-Validation*. Finally, the model performance is assessed in the section *Performance Evaluation* using conventional classification metrics.

Dataset preparation

Heart disease data were analyzed on both benchmark and real-world clinical datasets. The benchmark data were obtained from the UCI Machine Learning Repository and the Statlog Heart Disease dataset [18,24]. The UCI dataset contains Cleveland (303 samples), Hungary (294 samples), Switzerland (123 samples), and Long Beach VA (200 samples) databases, totaling 920 samples, while the Statlog dataset contains 270 samples. The target variables were defined consistently across datasets to build a consistent dataset. In the UCI dataset, the target property (num) used a value of 0 to indicate the absence of heart disease and values greater than 0 to indicate the presence of heart disease. For the binary classification task, all values greater than 0 were assigned 1. In the Statlog dataset, original labels of 1 (absence) and 2 (presence) were recoded to 0 and 1, respectively, for compatibility with the UCI dataset. Both datasets were standardized and integrated into a single dataset of 1,190 samples, with 561 healthy individuals and 629 patients with heart disease. The integration was made possible by common clinical attributes and consistent target definitions. The merged dataset included clinical features such as age, sex, chest pain type, cholesterol level, resting blood pressure, fasting blood sugar, ECG results, maximum heart rate, presence of angina while exercising, ST depression, slope, thal, and heart disease presence. The UCI dataset contained missing values, whereas the Statlog data collection did not contain missing values.

Apart from publicly available benchmark datasets, a clinical dataset was created from a real-world clinical setup at AMSHRC, Vijayapura, Karnataka, India. The dataset contained 2,000 anonymized patient records spanning ages 20 to 95 and comprising 17 attributes along with the target variable. Out of 2,000 records, 1,221 were classified as normal and 779 as heart disease cases. It included demographic, physiological, lifestyle, and diagnostic variables associated with cardiovascular disease risk assessment, such as age, blood pressure, pulse rate, oxygen saturation, cholesterol status, diabetes, hypertension, smoking or alcoholism, ECG, ECHO, TMT, and CAG parameters. Diagnostic data such as ECG, ECHO, TMT, and CAG were included to evaluate the model performance under routine clinical situations. However, these factors may have information directly related to the final diagnosis. The category “not performed” was retained for diagnostic investigations such as ECG, ECHO, TMT, and CAG, as some of the tests were either not clinically required for a

How to cite this article:

Budihal S, Kawale S R, Agarwal N (July 03, 2026) A Machine Learning Framework for Heart Disease Prediction With Feature Selection and Hyperparameter Optimization Across Benchmark and Real-World Datasets. Cureus J Comput Sci 3 : es44389-026-00150-7. DOI <https://doi.org/10.7759/s44389-026-00150-7>

particular patient or were contraindicated or deferred based on the patient's clinical condition and physician assessment. This category was retained to reflect actual clinical practice and to avoid assumptions about diagnostic information that was not accessible. The prediction problem was formulated as a binary classification problem in which $Y = 0$ indicated the absence of heart disease and $Y = 1$ indicated the presence. In the AMSHRC dataset, the final diagnosis recorded in hospital records by consulting cardiologists served as the basis for the target label. Heart disease status was determined independently of the ML approach and was based on extensive clinical evaluation in line with regular hospital practice. The AMSHRC dataset has also been utilized in our previous studies on heart disease prediction [26-28], and patient confidentiality was ensured by removing all identifiable information before analysis

Flow of the proposed work

Figure 7 presents an overview of the proposed approach for predicting heart disease. It starts with data collection and preprocessing, which includes imputing missing values using the KNN algorithm and detecting outliers using the Interquartile Range (IQR). The preprocessed data are analyzed using different feature selection approaches using Chi-Square and ANOVA F-Test methods. Then, DT, RF, LR, XGBoost, and AdaBoost algorithms are trained and optimized using the HGSCV method. Model performance are evaluated using accuracy, precision, recall, F1-score, and Area Under the Curve-Receiver Operating Characteristic (AUC-ROC) metrics under three experimental settings: first, without any feature selection approach; second, with Chi-Square feature selection; and third, with ANOVA F-Test feature selection.

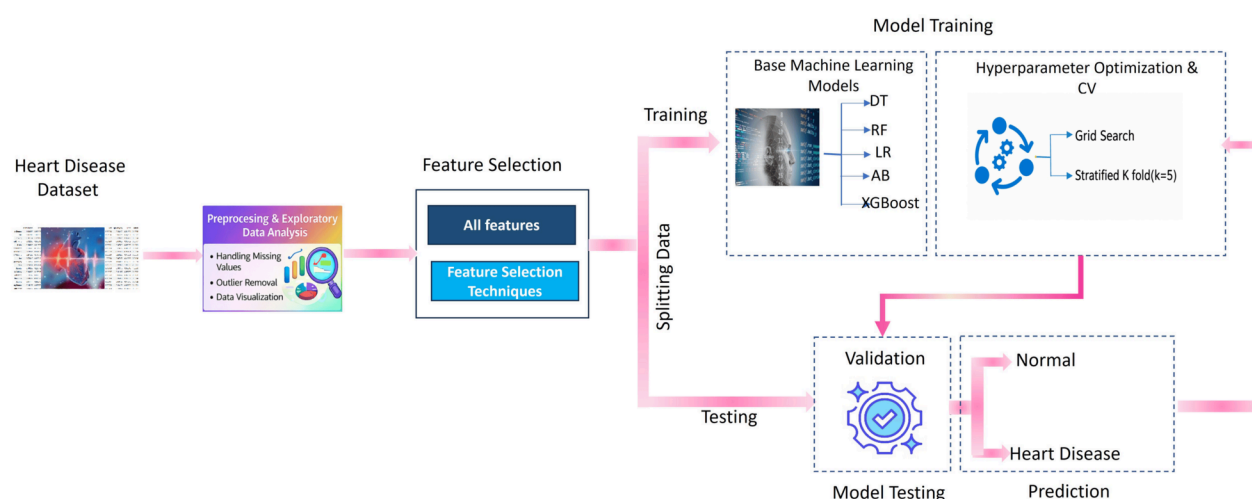


FIGURE 1: Proposed methodology.

DT, Decision Tree; RF, Random Forest; LR, Logistic Regression; XGBoost, Extreme Gradient Boosting; AB, Adaptive Boosting; CV, Cross-Validation

Exploratory data analysis

An exploratory data analysis (EDA) was conducted on both the UCI-Statlog and AMSHRC heart disease datasets to analyze data features and patterns, detect anomalies, and establish relationships among variables. During the initial stage of the analysis, data inspection and missing value checks were performed for both datasets. Missing values were identified in several UCI attributes, specifically the number of major vessels colored by fluoroscopy (ca) and the thallium stress test (thal), whereas no missing values were observed in the Statlog dataset. Univariate analysis was performed using histogram, box plot, and bar plot techniques to examine the distribution of variables such as age, cholesterol, resting blood pressure, and maximum heart rate, which helped identify outliers, skewness, and variability in the data. The distribution of the target variable showed class imbalance in the UCI-Statlog dataset. Relationships between predictor variables and heart disease status were further examined through bivariate analysis, count plots, box plots, and violin plots. Correlation heatmaps established associations among clinical variables, including positive correlations between

How to cite this article:

Budihal S, Kawale S R, Agarwal N (July 03, 2026) A Machine Learning Framework for Heart Disease Prediction With Feature Selection and Hyperparameter Optimization Across Benchmark and Real-World Datasets. Cureus J Comput Sci 3 : es44389-026-00150-7. DOI <https://doi.org/10.7759/s44389-026-00150-7>

thalach and heart disease and negative correlations between age and maximum heart rate. The AMSHRC dataset underwent the same EDA process, with additional emphasis on newly collected clinical and physiological variables, and data quality was verified to ensure consistency in structure and proper encoding of the variables. Continuous variables such as age, weight, pulse rate, SpO₂, and blood pressure values were examined using descriptive statistics and visualization methods. The majority of the values lie within the clinically acceptable ranges, while the distribution of blood pressure values clearly distinguished normal from hypertensive categories. Frequency distributions of categorical variables showed significant relationships between heart disease and chest pain, comorbidities, lifestyle choices, and diagnostic test outcomes.

Preprocessing

Data preprocessing was performed to enhance data quality and develop reliable models. A systematic preprocessing pipeline was used for both the UCI-Statlog and AMSHRC datasets, which involved missing value treatment, outlier detection, feature encoding, and correlation analysis. The AMSHRC and UCI-Statlog datasets contain missing data. The missingness in the UCI-Statlog dataset ranged from 4.96% to 51.3%, whereas the AMSHRC had minimal missing values ranging from 0.05% to 0.55%. The KNN imputation algorithm was implemented in order to estimate missing values without losing any information. The normalization of data was executed before imputation to avoid any scale bias during the estimation of distance. Five nearest neighbors ($k = 5$) were selected based on Euclidean distance, while missing values were estimated as an average of their neighbors. In the case of a minimum number of missing values in the AMSHRC dataset, the same imputation algorithm was applied. The IQR approach was applied to detect the outliers. The IQR was computed according to Equation (1), with Q1 representing the first quartile (25th percentile) and Q3 representing the third quartile (75th percentile). The data points were regarded as outliers when they were below the lower bound and upper bound, and this was checked through Equations (2 and 3).

$$IQR = Q_3 - Q_1 \quad (1)$$

$$Lower\ Bound = Q_1 - 1.5 \times IQR \quad (2)$$

$$Upper\ Bound = Q_3 + 1.5 \times IQR \quad (3)$$

Variables with continuous values such as age, weight, pulse rate, blood pressure, SpO₂, and cholesterol were analyzed based on those criteria. Outliers were treated by applying capping, where data points that fall outside the defined range were replaced by the corresponding boundary point values.

Binary and categorical features were labeled and encoded for modeling. Heatmaps of Pearson correlation were plotted to assess correlations between variables. The correlation matrices of UCI-Statlog and AMSHRC datasets are depicted in Figure 2 and Figure 3, respectively. To prevent information leakage, the datasets were split into 80% training and 20% testing sets. All preprocessing procedures, such as imputation using KNN, outlier removal, and feature selection, were performed only on the training set and then applied to the test set. Hyperparameters were tuned using five-fold grid search cross-validation on the training data.

How to cite this article:

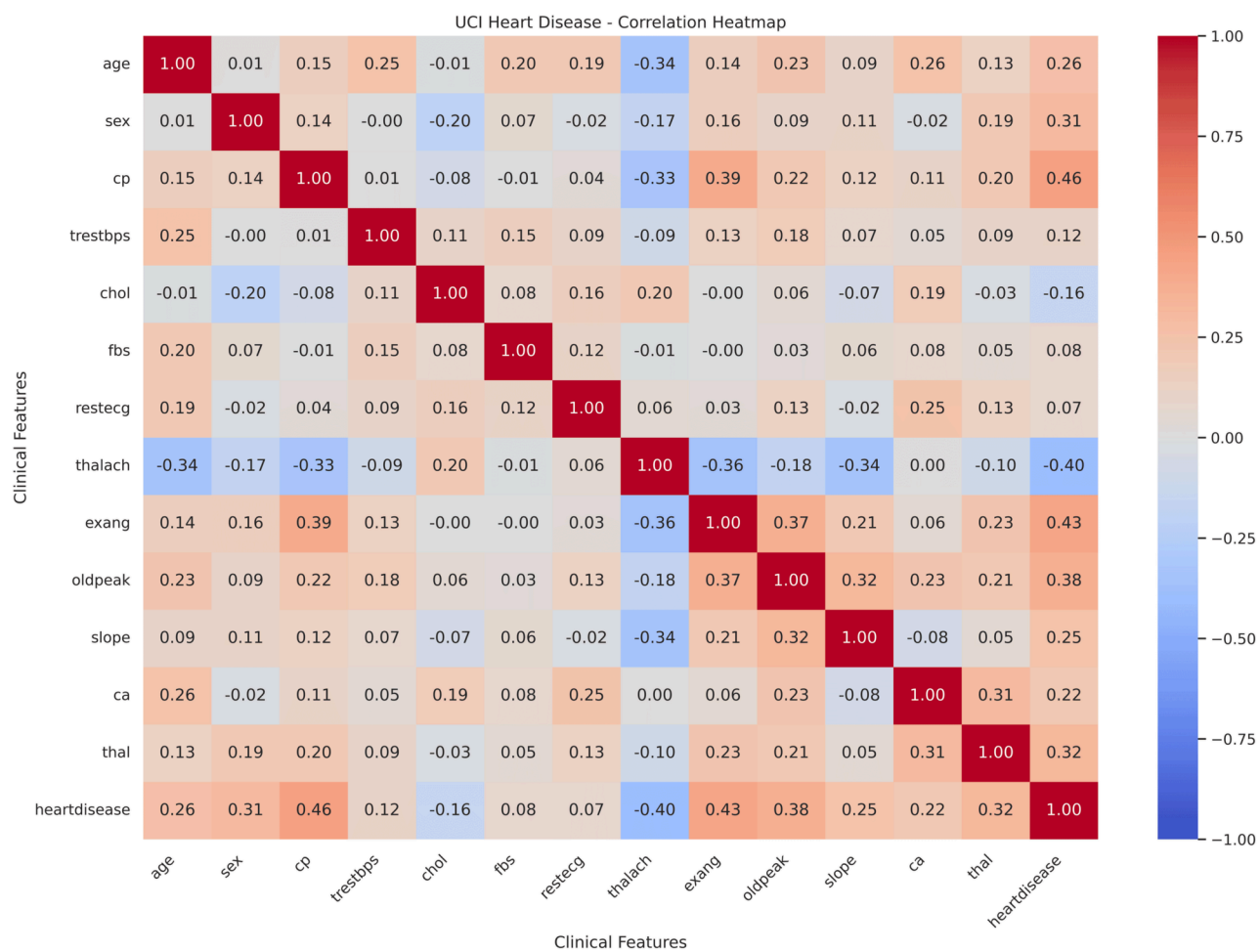


FIGURE 2: Correlation matrix for preprocessed UCI-Statlog heart disease dataset.

How to cite this article:

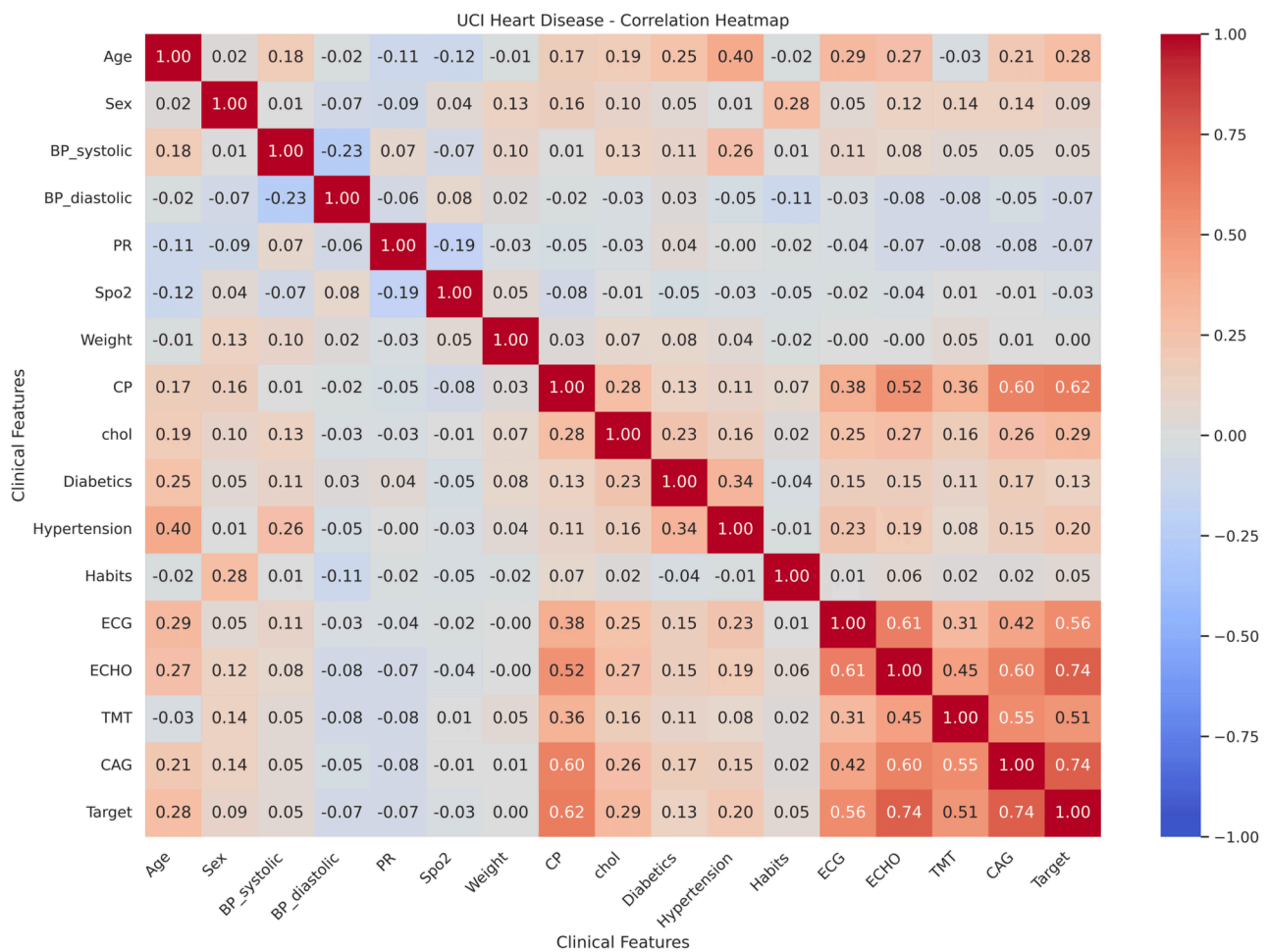


FIGURE 3: Correlation matrix for preprocessed AMSHRC heart disease dataset.

AMSHRC, Ayush Multi-Specialty Hospital and Research Centre; ECG, Electrocardiography; ECHO, Echocardiography; TMT, Tread Mill Test; CAG, Coronary Angiography; BP, Blood Pressure; CP, Chest Pain; PR, Pulse Rate

Features selection

Feature selection was carried out to identify the most informative variables and reduce dataset dimensions prior to model training. Two commonly applied statistical feature selection approaches, namely the Chi-Square test and the ANOVA F-Test, were applied in this work, both of which have demonstrated effectiveness in previous studies on heart disease prediction [10,12,17]. Both methods allow assessing the importance of individual features and selecting those that contribute to better classification results.

Chi-Square Test

The Chi-Square (χ^2) test was applied to evaluate the relationship between categorical variables and the target variable, with the Chi-Square statistic calculated according to Equation (4).

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (4)$$

How to cite this article:

Budihal S, Kawale S R, Agarwal N (July 03, 2026) A Machine Learning Framework for Heart Disease Prediction With Feature Selection and Hyperparameter Optimization Across Benchmark and Real-World Datasets. Cureus J Comput Sci 3 : es44389-026-00150-7. DOI <https://doi.org/10.7759/s44389-026-00150-7>

where O_i and E_i refer to the observed and expected frequencies, respectively, with larger Chi-Square values indicating a stronger dependence of the variable on the target class. In addition, appropriate encoding was used for categorical variables, and all feature values were transformed into a non-negative format to satisfy the requirements of the Chi-Square test.

ANOVA F-Test

The ANOVA F-Test was used to assess the discriminatory power of continuous variables concerning the target heart disease class. This approach allows evaluating the differences between means of the feature in question in case of different target classes. The F-statistic was calculated using Equation (5).

$$F = \frac{MS_{between}}{MS_{within}} \quad (5)$$

where $MS_{between} = SS_{between} / (k - 1)$ denotes the mean square between groups, $SS_{between}$ denotes the sum of squares between groups, k is the number of classes (in this study, $k = 2$: healthy and diseased classes), $MS_{within} = SS_{within} / (N - k)$ denotes the mean square within groups, SS_{within} denotes the sum of squares within groups, and N is the total number of samples.

The larger F-value and the smaller P -value imply the better discriminatory power of the feature in question. According to the rankings generated by the two feature selection approaches, the top eight features were selected for further use in the modeling process. The combination of the Chi-Square test and the ANOVA F-Test allowed identifying informative features and avoiding their redundancy.

ML models

For heart disease prediction, five supervised learning models were analyzed in this work: DT, RF, LR, XGBoost, and AB. These classifiers were selected because of their strong performance in binary classification tasks and their ability to handle heterogeneous clinical data comprising both numerical and categorical features. For all experimental studies, the 80:20 train/test split strategy was used for the analysis of the data.

Decision Tree

DT is a classification technique that builds a tree by splitting data into smaller parts according to certain feature values in a recursive manner. Each split maximizes inter-class variance and increases the homogeneity of nodes. In this research, node impurity is calculated using the Gini index, which is defined by Equation (6).

$$Gini = 1 - \sum_{i=1}^C p_i^2 \quad (6)$$

where p_i is the probability of class i , and C is the total number of classes. A smaller Gini Index value indicates higher node purity.

Random Forest

RF is an ensemble learning method that produces the final prediction by means of majority vote among several decision trees. Every decision tree is trained on a distinct bootstrap sample of the dataset. RF enhances model resilience and mitigates overfitting risk by consolidating the predictions of several trees, in contrast to a single DT. The prediction methodology of RF is defined in Equation (7).

$$\hat{y} = \text{majority}_{vote}(T_1(x), T_2(x) \dots T_n(x)) \quad (7)$$

where $T_i(x)$ represents the prediction of the i -th DT, and $\hat{y}$ denotes the final class prediction.

Logistic Regression

How to cite this article:

LR is a probabilistic classifier often used for binary classification tasks. This algorithm predicts the probability that an instance belongs to the target class using the logistic sigmoid function. The probability function of LR is formulated in Equation (8).

$$P(y = 1 | x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n)}} \quad (8)$$

where β_0 is the intercept term, β_i is the corresponding feature coefficient, and x_i is the feature value. Predicted probabilities were transformed to class labels by a 0.5 threshold.

XGBoost

XGBoost is an advanced gradient boosting technique used for building an ensemble of DTs sequentially. Each new DT attempts to correct the prediction errors of the previous trees and controls model complexity through regularization at the same time. The objective function of XGBoost is given in Equation (9).

$$Obj = \sum_{i=1}^n l(y_i, \hat{y}_i) + \sum_{k=1}^K \Omega(f_k) \quad (9)$$

where l represents the loss function, y_i represents the actual label, $\hat{y}_i$ represents the predicted label, and $\Omega(f_k)$ represents the regularization term that regulates model complexity.

AdaBoost

AdaBoost is a boosting method that combines several weak learners to build a strong predictive model. During training, greater emphasis is placed on observations that were previously misclassified, allowing subsequent learners to focus on more difficult cases. Predictions are made based on the weighted sum of all weak learners, as shown in Equation (10).

$$F(x) = \sum_{t=1}^T \alpha_t h_t(x) \quad (10)$$

where $h_t(x)$ denotes the weak learner built during the t -th iteration, while t, α_t is the weight of this weak learner, dependent on its classification ability.

Hyperparameter grid search cross-validation

Hyperparameter tuning was performed using HGSCV along with five-fold cross-validation. This allows for the evaluation of various hyperparameter settings to select the best among them while controlling the computation costs. During the tuning process, all possible parameter combinations were examined, and the one yielding the best performance in cross-validation was chosen. The proposed study employed five-fold cross-validation, in which the training data is split into five equal folds. In each step, four folds are used for training and one for validation. This process continues until all five folds have been used as the validation set. Finally, the overall cross-validation score is calculated as an average of scores obtained from all folds, according to Equation (11).

$$CV - Score = \frac{1}{k} \sum_{i=1}^k Score_i \quad (11)$$

where k is the number of folds, and $Score_i$ is the score obtained in the i -th fold. The average score serves as the model performance indicator.

The HGSCV technique was used in this study for the optimization of important parameters for DT, LR, RF, XGBoost, and AB models, namely the number of trees, max depth, learning rate, and regularization strength.

How to cite this article:

Performance evaluation

Standard evaluation metrics were used to assess model effectiveness: accuracy, precision, recall, and F1-score. They are constructed on the values of the confusion matrix: True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN). The accuracy, precision, recall, and F1-score are based on Equations (12-15).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (12)$$

$$Precision = \frac{TP}{TP + FP} \quad (13)$$

$$Recall = \frac{TP}{TP + FN} \quad (14)$$

$$F - Score = 2 \bullet \frac{Precision \bullet Recall}{Precision + Recall} \quad (15)$$

The relevance of the proposed methodology is evaluated through a series of experiments, with the corresponding findings and performance analysis presented in the following section.

Results And Discussion

Experimental setup

These experiments were conducted with both benchmark and real-world datasets. The benchmark datasets, namely the UCI Heart Disease dataset and Statlog Heart Disease dataset, were combined to create a dataset of 1,190 samples and 14 attributes. Also, a real-world clinical dataset from AMSHRC, India, consisting of 2,000 patient records with 17 attributes, was used. The target variable indicates the presence or absence of heart disease.

All the experiments were carried out in Python with the help of the Scikit-learn and XGBoost libraries in the Google Colab environment. In all experiments, a uniform data preprocessing pipeline was used, which included KNN imputation of missing values, IQR-based outlier removal, and label encoding of categorical variables.

An 80:20 train-test split was used to train five ML models, including DT, RF, LR, XGBoost, and AdaBoost. Grid search was used to tune the hyperparameters with five-fold cross-validation to improve model reliability. Model performance was assessed using standard measures such as accuracy, precision, recall, F1-score, and AUC-ROC.

Performance Comparison on UCI-Statlog and AMSHRC Datasets

This section presents and discusses the results obtained from ML models evaluated on the UCI-Statlog and AMSHRC datasets. The performance of the ML models that were fine-tuned using the HGSCV technique was analyzed using the following three experimental scenarios: no feature selection, feature selection based on the Chi-Square technique, and feature selection based on the ANOVA F-Test technique.

UCI-Statlog Dataset: The ML models were applied to the UCI-Statlog dataset under three experimental settings: (i) without feature selection, (ii) Chi-Square feature selection, and (iii) ANOVA F-Test feature selection. The equations of the second-order polynomial trends used in the graphical analysis are shown in Equations (12-15) and Figures 4-6, respectively. Figure 4 indicates that RF achieved an accuracy of 88% and an AUC of 0.94 with HGSCV when all features were retained. The results obtained using Chi-Square feature selection and HGSCV are shown in Figure 5, where the RF model still performed better, achieving 86% accuracy. The Chi-Square method selected the top features: exang, ca, thal, cp, slope, sex, oldpeak, and thalach. Similarly, Figure 6 shows that RF was still the best-performing model when ANOVA F-Test feature selection was used, with an accuracy of 85%. The features considered were cp, exang, thal, ca, oldpeak, thalach, slope, and sex. Table 2 demonstrates the accuracy of all models under different experimental settings.

How to cite this article:

The small decrease in accuracy after applying the Chi-Square and ANOVA F-Test feature selection techniques indicates that the original UCI-Statlog dataset is a concise and informative clinical feature set. The removal of features may have eliminated complementary information that influences model performance, resulting in only a minimal loss in predictive accuracy. The relatively small performance differences between the feature selection settings imply that the models achieved satisfactory predictive performance irrespective of the feature set used. All reported results were obtained using an 80:20 train-test split and are reported on the independent test set.

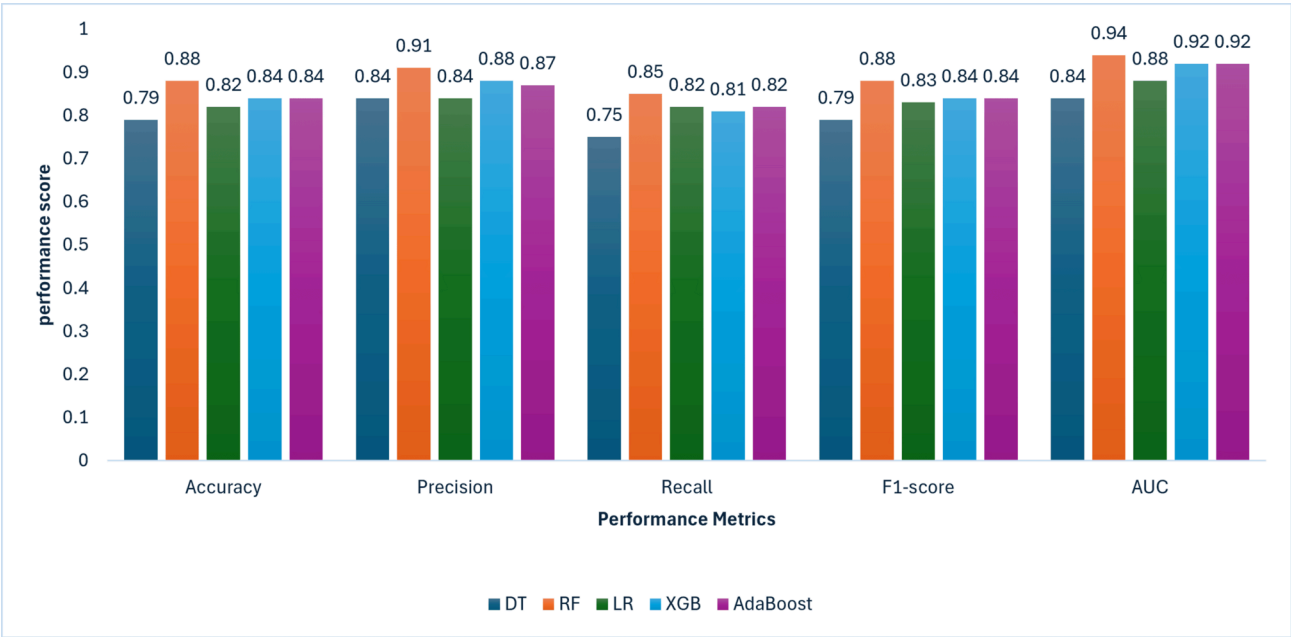


FIGURE 4: ML models' performance using HGSCV on the UCI-Statlog dataset without features selection.

AUC, Area Under the Curve; ML, Machine Learning; HGSCV, Hyperparameter Grid Search Cross-Validation; DT, Decision Tree; RF, Random Forest; LR, Logistic Regression; XGBoost, Extreme Gradient Boosting

How to cite this article:

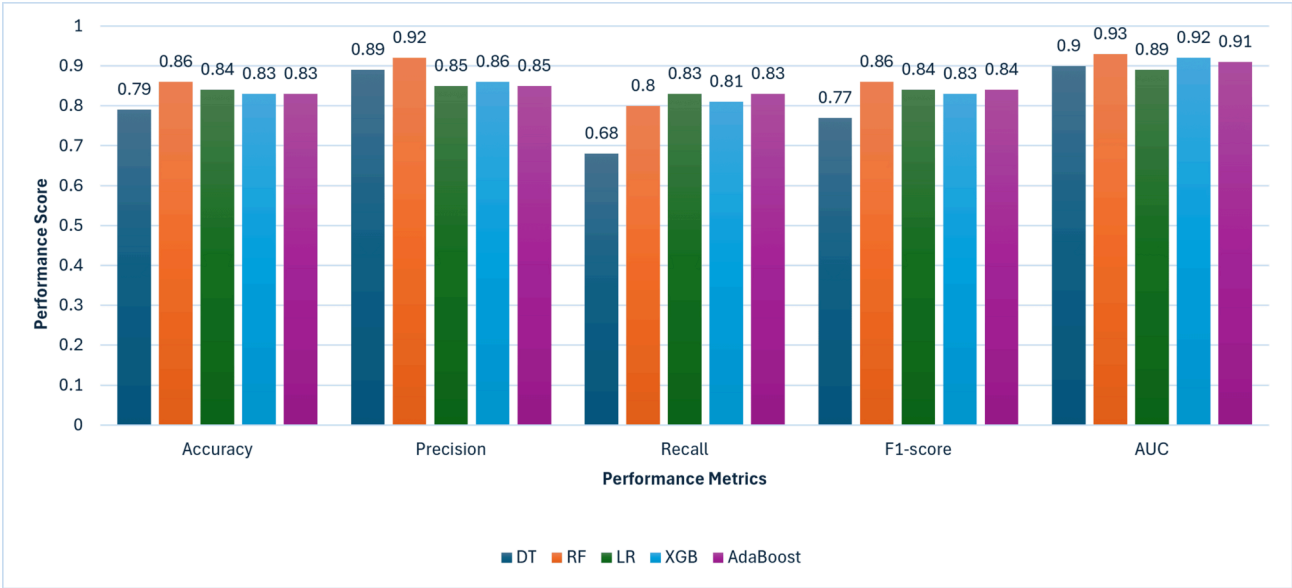


FIGURE 5: ML model's performance using HGSCV on the UCI-Statlog dataset with chi-square feature selection.

AUC, Area Under the Curve; ML, Machine Learning; HGSCV, Hyperparameter Grid Search Cross-Validation; DT, Decision Tree; RF, Random Forest; LR, Logistic Regression; XGBoost, Extreme Gradient Boosting

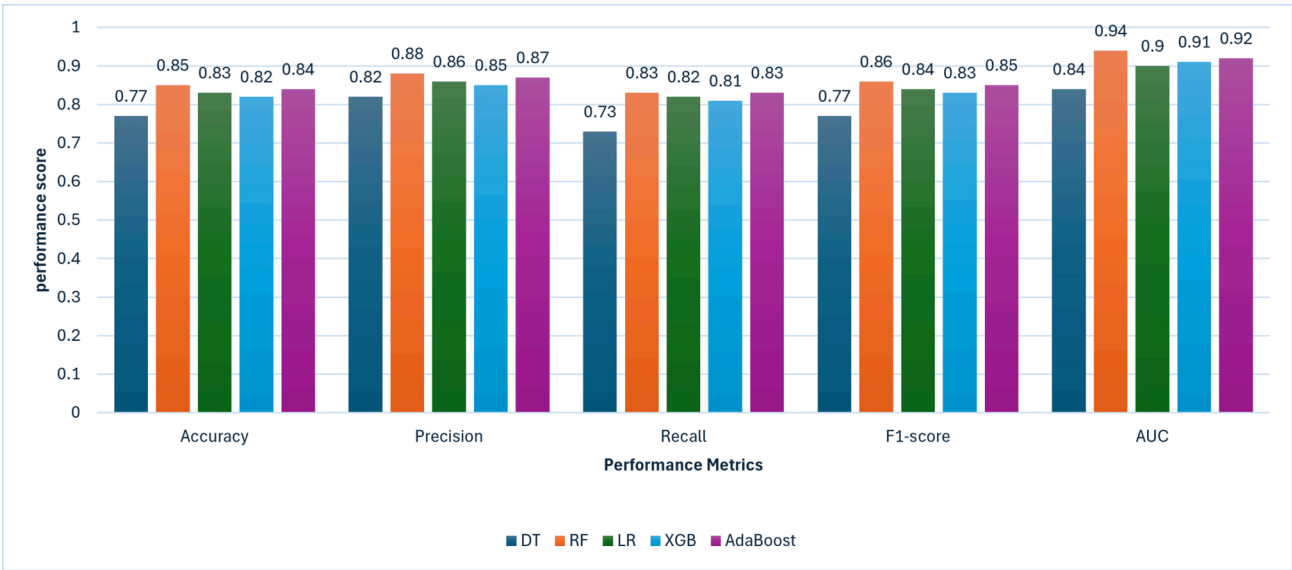


FIGURE 6: ML model's performance using HGSCV on the UCI-Statlog dataset with ANOVA F-Test feature selection.

AUC, Area Under the Curve; ML, Machine Learning; HGSCV, Hyperparameter Grid Search Cross-Validation; DT, Decision Tree; RF, Random Forest; LR, Logistic Regression; XGBoost, Extreme Gradient Boosting; ANOVA, Analysis of Variance

How to cite this article:

Budihal S, Kawale S R, Agarwal N (July 03, 2026) A Machine Learning Framework for Heart Disease Prediction With Feature Selection and Hyperparameter Optimization Across Benchmark and Real-World Datasets. Cureus J Comput Sci 3 : es44389-026-00150-7. DOI <https://doi.org/10.7759/s44389-026-00150-7>

Feature selection approaches	DT	RF	LR	XGB	AB
Without Feature Selection	79	88	82	84	84
Chi-Square	79	86	84	83	83
ANOVA F-Test	77	85	83	82	84

TABLE 2: ML models with HGSCV accuracy performance on the UCI-Statlog dataset.

ML, Machine Learning; HGSCV, Hyperparameter Grid Search Cross-Validation; DT, Decision Tree; RF, Random Forest; LR, Logistic Regression; AB, Adaptive Boosting; XGBoost, Extreme Gradient Boosting; ANOVA, Analysis of Variance

AMSHRC Dataset: The ML models were tested on the AMSHRC dataset under three different conditions, which are shown sequentially in Figures 7-9. The conditions were (i) without feature selection, (ii) with feature selection using the Chi-Square, and (iii) with feature selection using the ANOVA F-Test. In Figure 7, the XGBoost model showed a prediction accuracy of 96%, an F1-score of 0.95, and an AUC of 0.99. The Random Forest algorithm also showed commendable performance with 96% accuracy. DT and LR performed slightly worse but were still competitive. However, as shown in Figures 8 and 9, XGBoost still outperformed the other models when using the results of feature selection techniques, specifically the Chi-Square and ANOVA F-Test methods. The model achieved an accuracy of 96% and an AUC of 0.99. The ANOVA F-Test and Chi-Square methods showed that ECHO, CAG, CP, ECG, TMT, Age, Cholesterol, and Hypertension were the most important features. The performance of all studied models under different experimental conditions is presented in Table 3. The AMSHRC dataset has strong predictive ability, but diagnostic variables such as ECG, ECHO, TMT, and CAG are very important for the final diagnosis. Consequently, these variables may influence the expected performance, as noted above. The proposed framework should be considered a clinical decision support tool embedded in routine clinical practice rather than a pre-diagnostic screening model.

How to cite this article:

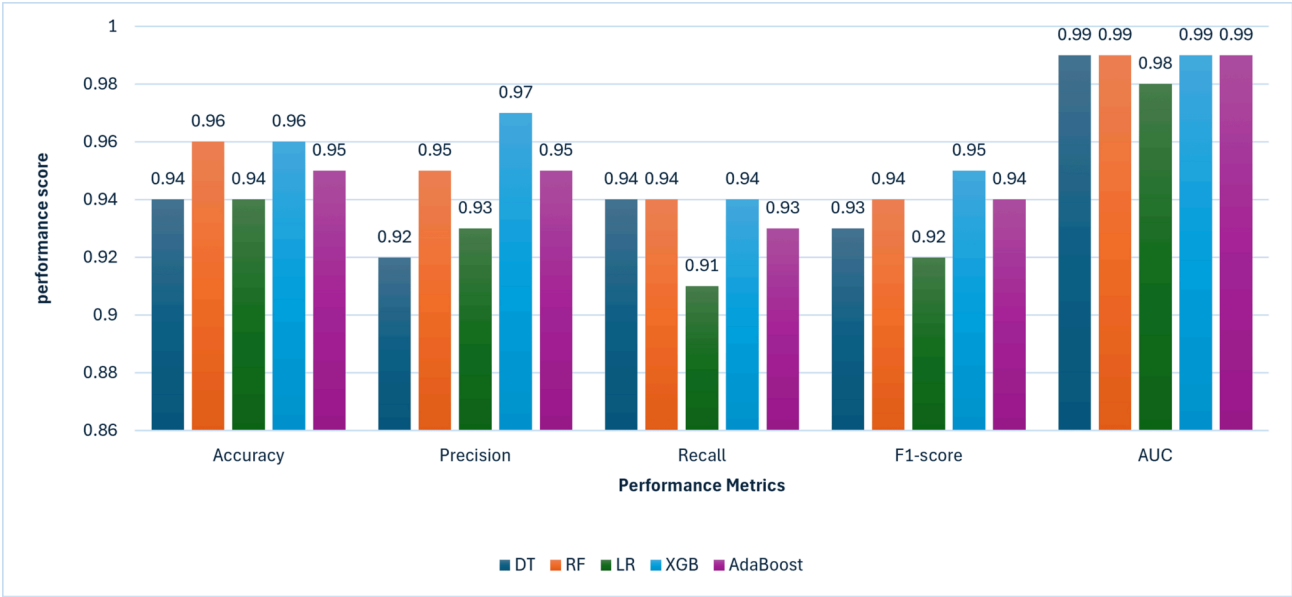


FIGURE 7: ML model's performance using HGSCV on the AMSHRC dataset without feature selection.

AUC, Area Under the Curve; ML, Machine Learning; HGSCV, Hyperparameter Grid Search Cross-Validation; AMSHRC, Ayush Multi-Specialty Hospital and Research Centre; DT, Decision Tree; RF, Random Forest; LR, Logistic Regression; XGBoost, Extreme Gradient Boosting

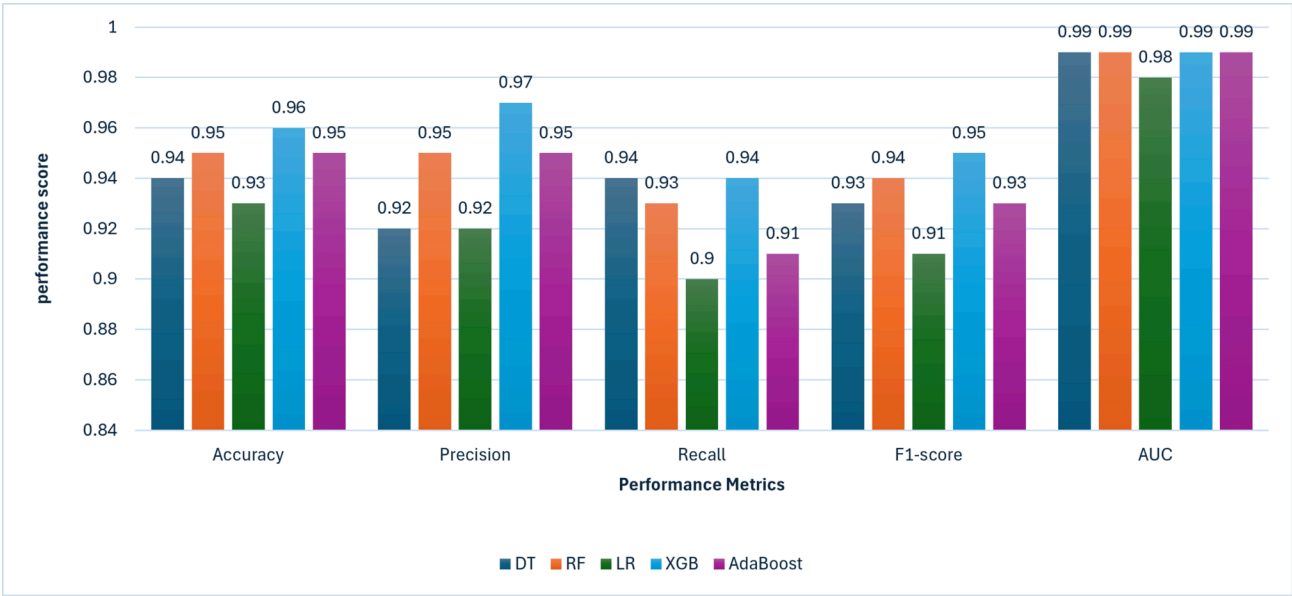


FIGURE 8: ML model's performance using HGSCV on the AMSHRC dataset with chi-square feature selection.

AUC, Area Under the Curve; ML, Machine Learning; HGSCV, Hyperparameter Grid Search Cross-Validation; AMSHRC, Ayush Multi-Specialty Hospital and Research Centre; DT, Decision Tree; RF, Random Forest; LR, Logistic Regression; XGBoost, Extreme Gradient Boosting

How to cite this article:

Budihal S, Kawale S R, Agarwal N (July 03, 2026) A Machine Learning Framework for Heart Disease Prediction With Feature Selection and Hyperparameter Optimization Across Benchmark and Real-World Datasets. Cureus J Comput Sci 3 : es44389-026-00150-7. DOI <https://doi.org/10.7759/s44389-026-00150-7>

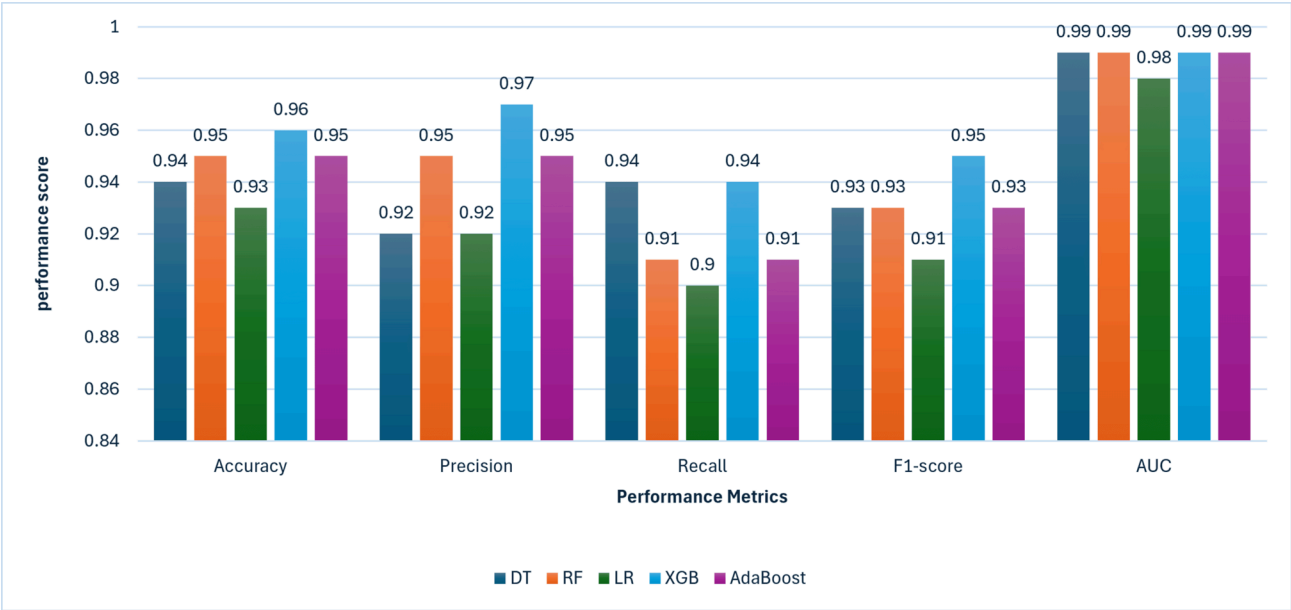


FIGURE 9: ML model's performance using HGSCV on the AMSHRC dataset with ANOVA F-Test feature selection.

AUC, Area Under the Curve; ML, Machine Learning; HGSCV, Hyperparameter Grid Search Cross-Validation; AMSHRC, Ayush Multi-Specialty Hospital and Research Centre; DT, Decision Tree; RF, Random Forest; LR, Logistic Regression; XGBoost, Extreme Gradient Boosting; ANOVA, Analysis of Variance

Feature selection approach	DT (%)	RF (%)	LR (%)	XGB (%)	AdaBoost (%)
Without Feature Selection	94	96	94	96	95
Chi-Square	94	95	93	96	95
ANOVA F-Test	94	95	93	96	95

TABLE 3: ML models with HGSCV accuracy performance on the AMSHRC dataset.

ML, Machine Learning; HGSCV, Hyperparameter Grid Search Cross-Validation; AMSHRC, Ayush Multi-Specialty Hospital and Research Centre; DT, Decision Tree; RF, Random Forest; LR, Logistic Regression; XGBoost, Extreme Gradient Boosting; AdaBoost, Adaptive Boosting; ANOVA, Analysis of Variance

Table 3 shows a comparison of model accuracy in three cases: no feature selection, Chi-Square, and ANOVA F-Test. Overall, the choice of features does not affect accuracy significantly; model performance remains largely consistent across all settings. XGBoost consistently achieves the highest accuracy (96%), indicating high robustness both with and without feature reduction. RF and AdaBoost also remain almost constant, while LR shows only a minimal decline in

accuracy with feature selection. These findings suggest that the input feature set is already highly informative. Overall, the results show similar patterns across all evaluation measures, and the ensemble models demonstrate strong baseline performance.

Model	CV accuracy (%)	CV AUC	Test accuracy (%)	Test AUC
DT	94.5 ± 1.0	0.99 ± 0.00	94.3	0.99
RF	95.2 ± 1.2	0.99 ± 0.00	95.8	0.99
LR	94.5 ± 1.3	0.99 ± 0.01	93.8	0.98
XGBoost	95.6 ± 1.1	0.99 ± 0.00	96.5	1.00
AdaBoost	95.6 ± 1.0	0.99 ± 0.00	95.5	0.99

TABLE 4: CV and test results on ML models of the AMSHRC dataset.

AUC, Area Under the Curve; CV, Cross-Validation; ML, Machine Learning; AMSHRC, Ayush Multi-Specialty Hospital and Research Centre; DT, Decision Tree; RF, Random Forest; LR, Logistic Regression; XGBoost, Extreme Gradient Boosting; AdaBoost, Adaptive Boosting

Table 4 indicates that the ensemble models performed better than the single classifiers. Cross-validation accuracy and AUC were highest for the RF and XGBoost models, indicating that both models have a high degree of generalization ability. The close match between cross-validation and independent test performance shows little overfitting and strong model stability. For example, XGBoost achieved a cross-validation accuracy of $95.6 \pm 1.1\%$ and a test accuracy of 96.5% , while RF achieved a cross-validation accuracy of $95.2 \pm 1.2\%$ and a test accuracy of 95.8% . The consistency between the cross-validation and test results indicates that the models have consistent predictive performance on unidentified data and adequate generalization capabilities across the datasets tested.

Comparative Analysis of the Proposed Framework with Existing Studies

Table 5 shows a comparative analysis of ML models for predicting heart disease on the UCI-Statlog dataset (1,190 samples). Previous studies have reported high classification accuracy, such as 96.1% with a Chi-Square feature selection QBE model [10]. Nevertheless, this performance was obtained on a somewhat limited dataset, and such results may limit its generalizability to heterogeneous real-world clinical data. Other studies have reported 93% accuracy with XGBoost and manual feature selection [22] and 98% accuracy with ensemble techniques such as stacking and voting [23]; however, none of these methods employed systematic feature selection techniques.

In contrast, the proposed framework compares several learning models using structured feature selection methods. Specifically, the proposed model assesses various ML algorithms based on systematic feature selection techniques, such as Chi-Square and ANOVA F-Test. Both the UCI-Statlog and AMSHRC datasets were subjected to the same set of ML algorithms, namely DT, RF, LR, XGBoost, and AdaBoost, to provide a fair and consistent comparison.

On the UCI-Statlog dataset, RF was the most accurate model, achieving 88% accuracy without feature selection, and it remained the effectively performing model under different feature selection settings. The minimal accuracy drop after the feature selection indicates that the initial UCI-Statlog dataset already contained highly informative clinical features and that the assessed models were resistant to the fluctuations in the selected feature set.

In the AMSHRC dataset, XGBoost achieved an accuracy of 96%, indicating strong predictive performance on real-world clinical data. Furthermore, the proposed framework was validated on both the benchmark UCI-Statlog dataset and the real-world AMSHRC dataset using consistent and similar preprocessing, feature selection, and evaluation techniques. This approach allows a more realistic assessment of model robustness and clinical applicability than studies based solely on benchmark datasets. Future work involving hybrid and stacked ML frameworks, external validation using multicenter datasets, and explainable artificial intelligence methodologies could further improve the predictive performance, transparency, and generalizability of the ML models.

Reference	Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	Feature selection
[10]	QBE	96.1	97.1	95.5	96.3	Chi-Square
[22]	XGBoost	93	92	92	92	Manual (7 features)
[23]	Stacking	98	98.8	98.1	98.4	None
[23]	Voting	98	97.5	98.7	98	None
Proposed	Random Forest (UCI)	88	91	85	88	Without feature selection
Proposed	XGBoost (AMSHRC)	96	97	94	95	Chi-Square and ANOVA

TABLE 5: Comparison and contrast of UCI-Statlog and AMSHRC dataset.

AMSHRC, Ayush Multi-Specialty Hospital and Research Centre; QBE, Quantum Bagging Ensemble; XGBoost, Extreme Gradient Boosting; ANOVA, Analysis of Variance

The stability of the presented framework is also increased by the fact that the performance was consistent across models. While the obtained results demonstrate the performance of the proposed models across different experimental settings, the following section provides a deeper interpretation of these findings in terms of clinical relevance, feature influence, and model generalization.

Discussion

The findings of this study are compared with the existing methods that employ the analogous ML algorithms, which include DT, RF, LR, XGBoost, and AdaBoost, as well as feature selection techniques such as Chi-Square and ANOVA F-Test. Some of the earlier studies have achieved good results with the use of a combination of ensemble algorithms and feature selection. However, most early research has been conducted on benchmark datasets, thus limiting the applicability to actual clinical settings.

However, in the case of the UCI-Statlog dataset, the maximum accuracy achieved by RF was 88%, whereas slightly lower accuracies were observed using Chi-Square and ANOVA F-Test feature selection techniques. These findings suggest that feature selection is not always beneficial, particularly when the dataset already contains highly informative features.

How to cite this article:

Budihal S, Kawale S R, Agarwal N (July 03, 2026) A Machine Learning Framework for Heart Disease Prediction With Feature Selection and Hyperparameter Optimization Across Benchmark and Real-World Datasets. Cureus J Comput Sci 3 : es44389-026-00150-7. DOI <https://doi.org/10.7759/s44389-026-00150-7>

XGBoost attained an accuracy of 96% and an AUC of 0.99, making it highly predictive across all experimental setups on the real-world AMSHRC dataset.

In addition, the close agreement between cross-validation and independent test results implies minimal overfitting and good generalization ability. For instance, XGBoost achieved cross-validation and test accuracies of $95.6 \pm 1.1\%$ and 96.5% , respectively, while RF achieved cross-validation and test accuracies of $95.2 \pm 1.2\%$ and 95.8% , respectively. From these results, it can be noted that the proposed models were able to maintain consistent performance when applied to unseen data.

Diagnostic variables also improved predictive power, but adding these variables could have brought in more information that was close to the final diagnosis. It is therefore regarded as a tool to support clinical decision-making in clinical routines and not as the sole predictive test for pre-diagnosis. Models using only baseline demographic, physiological, and lifestyle characteristics would be valuable for further evaluating early-stage heart disease risk.

A key contribution of this study lies in the comparative analysis of multiple ML models that are systematically assessed over benchmark and AMSHRC datasets through a standard pipeline that includes preprocessing, feature selection, a hyperparameter optimizing approach, and test methodology. There are still quite limited comparative assessments available in the literature. Furthermore, few studies have focused on the role of heart disease prediction without advanced diagnostic variables, as such variables are often necessary for clinical screening and risk assessment.

This study addresses existing gaps by conducting a systematic comparison of multiple ML algorithms with and without feature selection across benchmark and real-world clinical datasets. Unified preprocessing, hyperparameter optimization, and evaluation methods were used to ensure reliable performance evaluation. Furthermore, the addition of real-world clinical data improves the practical relevance of the proposed work.

Nevertheless, this study has several limitations. Findings from the AMSHRC dataset may not be generalizable to all patients or healthcare environments because the dataset was obtained from a single healthcare institution. In addition, a “not performed” category was retained for diagnostic tests such as ECG, ECHO, TMT, and CAG because these tests were not clinically required for certain patients, were contraindicated, or were deferred based on the patients' clinical condition and physician assessment. Therefore, this category may also reflect clinical decision-making and could influence model performance. Future work should assess the impact of this category on the predictive performance and generalizability of the models using independent multicenter datasets.

Future studies should focus on developing hybrid ML frameworks, performing external validation using multicenter datasets, and integrating explainable artificial intelligence methods such as SHapley Additive exPlanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) to enhance the transparency and explainability of the models.

Conclusions

In this paper, we present a comprehensive ML framework for heart disease prediction using both the benchmark UCI-Statlog and the real-world AMSHRC datasets. A systematic evaluation of five ML models (DT, RF, LR, XGB, and AB) was conducted on both datasets. For the UCI-Statlog dataset, RF showed the highest accuracy (88%) without any feature selection, with only marginal drops when using Chi-Square and ANOVA F-Test feature selection methods, indicating that the original features are quite relevant. XGBoost achieved the best performance on the AMSHRC dataset, with 96% accuracy, a 95% F1-score, and an AUC of 0.99 across all experimental conditions. The results indicate reliable predictive performance on both benchmark and real-world clinical datasets. Furthermore, the high correlation between cross-validation and independent test performance indicates low overfitting and supports the generalizability of the proposed methodology. However, the availability of diagnostic variables such as ECG, ECHO, TMT, and CAG in the AMSHRC dataset makes the framework more suitable as a clinical decision support tool rather than a pre-diagnostic screening model. In addition, the AMSHRC dataset was collected from a single healthcare facility, which may restrict the generalizability of

How to cite this article:

Budihal S, Kawale S R, Agarwal N (July 03, 2026) A Machine Learning Framework for Heart Disease Prediction With Feature Selection and Hyperparameter Optimization Across Benchmark and Real-World Datasets. *Cureus J Comput Sci* 3 : es44389-026-00150-7. DOI <https://doi.org/10.7759/s44389-026-00150-7>

the results. Future work will focus on hybrid ML frameworks, multicenter external validation, and explainable artificial intelligence approaches, such as SHAP and LIME, to improve model transparency, interpretability, and clinical decision support.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Sunanda Budihal, Sheetalrani R. Kawale

Acquisition, analysis, or interpretation of data: Sunanda Budihal, Sheetalrani R. Kawale , Nitin Agarwal

Drafting of the manuscript: Sunanda Budihal, Sheetalrani R. Kawale , Nitin Agarwal

Critical review of the manuscript for important intellectual content: Sunanda Budihal, Sheetalrani R. Kawale , Nitin Agarwal

Supervision: Sunanda Budihal

Disclosures

Human subjects: All authors have confirmed that this study did not involve human participants or tissue. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Data Availability Statements

The datasets (and/or code) supporting this study are available from the corresponding author upon reasonable request.

Acknowledgements

The authors gratefully acknowledge the academic support provided by the staff of Ayush Multi-Speciality Hospital and Research Centre (AMSHRC), Vijayapura, Karnataka, India, for their clinical support. The authors further acknowledge Dr. Laxmi Kattimani, College Librarian, Government First Grade College, Bagalkot, for assistance for plagiarism checking.

References

1. Di Cesare M, Perel P, Taylor S, et al.: [The heart of the world](#). Global Heart. 2024, 19:1-13. [10.5334/gh.1288](#)
2. Antza C, Gallo A, Boutari C, et al.: [Prevention of cardiovascular disease in young adults: focus on gender differences. A collaborative review from the EAS Young Fellows](#). Atherosclerosis. 2023, 384:1-14. [10.1016/j.atherosclerosis.2023.117272](#)
3. Ghodeswar GK, Dube A, Khobragade D: [Impact of lifestyle modifications on cardiovascular health: a narrative review](#). Cureus. 2023, 15:e42616. [10.7759/cureus.42616](#)
4. Młynarska E, Czarnik W, Fularski P, et al.: [From atherosclerotic plaque to myocardial infarction—the leading cause of coronary artery occlusion](#). International Journal of Molecular Sciences. 2024, 25:1-27. [10.3390/ijms25137295](#)
5. Costantino MF, Cortese F, D'Addeo G, Nicoletti A, Mancino S, Stolfi L: [Diagnostic modalities for ischemic heart disease: evaluating the role of stress echocardiography, cardiac CT, and myocardial perfusion scintigraphy in guiding coronary angiography](#). Exploration of Cardiology. 2025, 3:1-19. [10.37349/ec.2025.101243](#)

How to cite this article:

Budihal S, Kawale S R, Agarwal N (July 03, 2026) A Machine Learning Framework for Heart Disease Prediction With Feature Selection and Hyperparameter Optimization Across Benchmark and Real-World Datasets. Cureus J Comput Sci 3 : es44389-026-00150-7. DOI <https://doi.org/10.7759/s44389-026-00150-7>

6. Varnosfaderani SM, Forouzanfar M: [The role of AI in hospitals and clinics: transforming healthcare in the 21st century](#). Bioengineering. 2024, 4:1-38. [10.3390/bioengineering11040337](#)
7. Bekbolatova M, Mayer J, Ong CW, Toma M: [Transformative potential of AI in healthcare: definitions, applications, and navigating the ethical landscape and public perspectives](#). Healthcare. 2024, 12:1-31. [10.3390/healthcare12020125](#)
8. Taye MM: [Understanding of machine learning with deep learning: architectures, workflow, applications and future directions](#). Computers. 2023, 12:1-26. [10.3390/computers12050091](#)
9. Razzaq K, Shah M: [Machine learning and deep learning paradigms: from techniques to practical applications and research frontiers](#). Computers. 2025, 14:1-27. [10.3390/computers14030093](#)
10. Babu VS, Ramya P, Gracewell J: [Revolutionizing heart disease prediction with quantum-enhanced machine learning](#). Scientific Reports. 2024, 14:1-16. [10.1038/s41598-024-55991-w](#)
11. [Heart failure prediction dataset](#). Kaggle. Accessed: April 15, 2026: <https://www.kaggle.com/datasets/fedesoriano/heart-failure-prediction>.
12. Omkari DY, Shaik K: [An integrated two-layered voting \(TLV\) framework for coronary artery disease prediction using machine learning classifiers](#). IEEE Access. 2024, 12:56275-56290. [10.1109/access.2024.3389707](#)
13. [Cardiovascular disease dataset](#). Kaggle. Accessed: May 20, 2025: <https://www.kaggle.com/datasets/sulianova/cardiovascular-disease-dataset>.
14. [Heart disease prediction](#). Kaggle. Accessed: May 21, 2025: <https://www.kaggle.com/datasets/moazeldsokyx/heart-disease>.
15. Vijayaraj AR, Pasupathi S: [Nature inspired optimization in context-aware-based coronary artery disease prediction: a novel hybrid Harris Hawks approach](#). IEEE Access. 2024, 12:92635-92651. [10.1109/access.2024.3414662](#)
16. [Predicting heart disease using clinical variables](#). Kaggle. Accessed: May 21, 2025: <https://www.kaggle.com/datasets/thedevastator/predicting-heart-disease-risk-using-clinical-var>.
17. Mahajan A, Kaushik B, Rahmani MKI, Banga AS: [A hybrid feature selection and ensemble stacked learning models on multi-variant CVD datasets for effective classification](#). IEEE Access. 2024, 12:87023-87038. [10.1109/access.2024.3412077](#)
18. Janosi A, Steinbrunn W, Pfisterer M, Detrano R. Heart Disease [Dataset]. UCI Machine Learning Repository. (1989). Accessed: May 21, 2025: <https://archive.ics.uci.edu/dataset/45/heart+disease>.
19. Ram Kumar RP, Polepaka S, Manasa V: [Investigations on cardiovascular diseases and predicting using machine learning algorithms](#). Cogent Engineering. 2024, 11:1-13. [10.1080/23311916.2024.2386381](#)
20. Cenitta D, Vijaya Arjunan R, Paramasivam G, Arul N, Palkar A, Chadaga K: [Ischemic heart disease prognosis: a hybrid residual attention-enhanced LSTM model](#). IEEE Access. 2025, 13:4281-4289. [10.1109/access.2024.3524604](#)
21. Addisu EG, Yirga TG, Yirga HG, Yehuala AD: [Transfer learning-based hybrid VGG16-machine learning approach for heart disease detection with explainable artificial intelligence](#). Frontiers in Artificial Intelligence. 2025, 8:1-18. [10.3389/frai.2025.1504281](#)
22. Teja MD, Rayalu GM: [Optimizing heart disease diagnosis with advanced machine learning models: a comparison of predictive performance](#). BMC Cardiovascular Disorders. 2025, 25:1-12. [10.1186/s12872-025-04627-6](#)
23. Ganie SM, Pramanik PKD, Zhao Z: [Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets](#). Scientific Reports. 2025, 15:1-34. [10.1038/s41598-025-97547-6](#)
24. [Statlog \(Heart\) \[Dataset\]](#). UCI Machine Learning Repository. Accessed: May 21, 2025: <https://archive.ics.uci.edu/dataset/145/statlog+heart>.
25. Doppala BP, Bhattacharyya D: [Cardiovascular_Disease_Dataset](#). Mendeley Data, V1. (2021). Accessed: May 21, 2025: <https://data.mendeley.com/datasets/dzz48mvjht/1>.
26. Budhal S, Kawale S, Agarwal N: [Heart disease prediction using machine learning on a real-time clinical dataset with multi-diagnostic features](#). International Journal of Applied Mathematics. 2025, 38:2719-2742. [10.12732/ijam.v38i12s.1584](#)
27. Budihal S, Kawale S, Hallalli B, Agarwal N: [Systematic benchmarking of ANN, CNN, and LSTM models for heart disease prediction using multimodal clinical features](#). International Journal of Health Sciences and Research. 2026, 16:176-190. [10.52403/ijhsr.20260521](#)
28. Budihal S, Kawale SR, Agarwal N, Hallalli B: [Feature-group-aware cardiovascular disease prediction using gated neural fusion and CatBoost-based clinical feature analysis](#). Cureus Journal of Computer Science. 2026, 3:[10.7759/s44389-026-00068-0](#)

How to cite this article:

Budihal S, Kawale S R, Agarwal N (July 03, 2026) A Machine Learning Framework for Heart Disease Prediction With Feature Selection and Hyperparameter Optimization Across Benchmark and Real-World Datasets. Cureus J Comput Sci 3 : es44389-026-00150-7. DOI <https://doi.org/10.7759/s44389-026-00150-7>